*Economics and Philosophy*, **29** (2013) 235–261 © Cambridge University Press
doi:10.1017/S0266267113000217

# FOLK RETRIBUTIVISM AND THE COMMUNICATION CONFOUND

**Thomas Nadelhoffer**
*College of Charleston, USA*

**Saeideh Heshmati**
*University of Arizona, USA*

**Deanna Kaplan**
*University of Arizona, USA*

**Shaun Nichols**
*University of Arizona, USA*
*shaunbnichols@gmail.com*

Retributivist accounts of punishment maintain that it is right to punish wrongdoers, even if the punishment has no future benefits. Research in experimental economics indicates that people are willing to pay to punish defectors. A complementary line of work in social psychology suggests that people think that it is right to punish wrongdoers. This work suggests that people are *retributivists* about punishment. However, all of the extant work contains an important potential confound. The target of the punishment is expected to be aware of the punitive act. Thus, it's possible that people punish because they want to communicate something to the wrongdoer, e.g. disapproval, the presence of a norm, etc. In three studies, we examine whether people will punish even when the punishee will be ignorant. We find that people are no less likely to punish when the punishee will be

We thank Jason Dana for allowing us to adapt his ingenious new work (Dana *et al*. forthcoming) for our economic game. In addition, for discussion and comments on the paper, we thank Eyal Aharoni, Jerry Gaus, James Konow, Eddy Nahmias, Eric Schwitzgebel and two anonymous reviewers.



https://doi.org/10.1017/S0266267113000217 Published online by Cambridge University Press

ignorant. This finding emerges both in a survey study and in a monetized behavioural decision study.

## 1. INTRODUCTION

> Philosophies of punishment must . . . be assessed not merely as they appear on the page, but as they are (or could practicably be) realized in specific practices. Pragmatic penology must be subjected to close moral and political scrutiny as well as to rigorous empirical testing. (R.A. Duff and D. Garland 1994)

One of the recurrent themes in the growing empirical literature on the social psychology of punishment is that ordinary people are consequentialists in theory but not in practice – i.e. while people often claim to be interested in preventing and deterring crime, their actual judgements about punishment appear to be driven by retributive intuitions about just deserts (e.g. McFatter 1978, 1982; Kahneman *et al.* 1998; Darley *et al.* 2000; Carlsmith *et al.* 2002; Carlsmith 2006, 2008; Carlsmith and Darley 2008). However, one notable shortcoming of the extant literature is that researchers have failed to operationalize retributivism in a way that adequately distinguishes it from consequentialist decision procedures which take backward-looking factors into consideration. It is difficult to isolate purely retributive motives in concrete decisions about punishment because punishing wrongdoers often serves a communicative role – e.g. punishment both enables the public to express its disapproval of the offence and it can have a communicative effect on the wrongdoer in the form of moral education. Because this dual communicative role could have been partly driving people's punishment decisions in earlier studies, we can't draw any firm conclusions about the role that purely retributive motives play when it comes to the social psychology of punishment.

In order to control for this potential communication confound, we ran a series of new studies which were designed to more cleanly cleave apart retributive and consequentalist motives for punishment. After setting the stage with a brief discussion of the traditional philosophical debate between consequentialists and retributivists concerning the proper justification of punishment, we will discuss some of the shortcomings of the otherwise illuminating and important recent work on the social psychology of punishment. Then, we present the results of three new studies which control for the communication confound and provide new evidence (a) that people overwhelmingly believe that it is right to punish wrongdoers even if the punishment confers no (additional) benefits, and (b) that people's concrete decisions to punish are generated by a retributive motive.

https://doi.org/10.1017/S0266267113000217 Published online by Cambridge University Press

## 2. THE PHILOSOPHY AND PSYCHOLOGY OF PUNISHMENT

Historically, theories of punishment have been divided into two general groups – consequentialism and retributivism. In their attempts to justify both the institution and the implementation of punishment, consequentialists appeal to a wide range of potentially positive benefits produced by punishing wrongdoers including:

1. General deterrence – i.e. punishing in order to deter other would-be offenders from committing similar offences.
2. Incapacitation – i.e. punishing in order to prevent the offender from committing similar crimes while he is being detained and/or treated.
3. Rehabilitation and moral education – i.e. punishing in order to rehabilitate or re-educate the offender.
4. Communication – i.e. punishing in order to communicate or express disapproval of an action.
5. Catharsis – i.e. punishing in order to give victims and society more generally a healthy emotional release.
6. Norm reinforcement – i.e. punishing in order to highlight and reassert the importance of social values and norms.
7. Quelling revenge – i.e. punishing in order to keep the original or third parties from starting a blood feud.

In each of these cases, the *prima facie* moral badness of punishment is offset by the benefits that it produces. However, in the event that punishment fails to produce these positive consequences either in general or in a particular case, it loses its moral legitimacy. On these views, the moral propriety of punishment is entirely a function of the consequences of punishment.

Retributivist theories of punishment reject this instrumentalist approach. According to retributivism, punishment is not justified in terms of some future benefit.[1] Rather, retributivists believe that 'punishment institutions in general are justified by the giving of just deserts' and that 'the punishment of each offender is justified by the fact that he or she deserves it' (Moore 1997: 155). On this non-consequentialist view, if an agent is morally responsible for wrongdoing (of sufficient gravity), then that person deserves to be punished – i.e. harmed – even if the harm produces no (other) benefits. Of course, if punishing the morally deserving happens to yield positive side effects such as deterring crime or rehabilitating the offender, these consequences may be welcomed by the retributivist as a 'happy surplus' (Moore 1997: 89), but they do

[1] For a thorough taxonomy of the varieties of retributivism, see Cottingham (1979).

https://doi.org/10.1017/S0266267113000217 Published online by Cambridge University Press

not constitute any part of the justification of punishment for the pure retributivist. According to retributivists, giving wrongdoers their just deserts is intrinsically and not merely instrumentally valuable (Moore 1997: 157).

However, while the philosophical debate about punishment has focused primarily on the normative issue of why we *ought* to punish, psychologists have focused on the descriptive question of why we *actually* punish. For instance, Kevin Carlsmith and colleagues have recently built upon the earlier empirical work on the apparent discrepancy between people's stated theories of punishment and their decisions concerning how much to punish (Hogarth 1971; Vidmar 1974; McFatter 1978, 1982). In short, Carlsmith and colleagues maintain that (i) people profess to care a great deal about both giving offenders their just deserts and deterring crime and incapacitating offenders (Carlsmith *et al.* 2002: 295); however, (ii) when given examples of prototypical wrongdoing, people's actual judgements concerning how much to punish are driven nearly exclusively by retributivist considerations (e.g. Darley *et al.* 2000; Carlsmith *et al.* 2002; Carlsmith 2006, 2008; Carlsmith and Darley 2008).[2]

In order to identify participants' intuitions with one of the three main sentencing philosophies – namely, retributivism, deterrence and incapacitation – Carlsmith and colleagues associated each sentencing philosophy with the following factors (Carlsmith 2006: 440):

- Retributivism – magnitude of harm, perpetrator culpability, extenuating circumstances.
- Deterrence – general frequency of crime, detection rate, publicity.
- Incapacitation – likelihood of reoffending, prior record, self-control.

Having operationalized retributivism, deterrence, and incapacitation in this way, Carlsmith and colleagues found that participants' verbal reports suggest that they largely agree with all three sentencing philosophies. However, they also found, across a series of vignette-based experiments, that when making sentencing decisions, people neglect considerations important to consequentialism (e.g. deterrence and incapacitation) and focus instead on the factors that are important to retributivism (e.g. severity and culpability).

As a package, the findings by Carlsmith and colleagues suggest an unnerving disconnect between theory and practice. As such, these researchers conclude that 'people do not have a good sense of their

[2] When Carlsmith and colleagues present their participants with characterizations of retributivism, they use the labels 'just deserts' and 'deservingness philosophy' in place of 'retributivism'. This is partly because they wanted to avoid possible negative connotations associated with the word 'retribution' and its cognates (Carlsmith 2008: fn 2).

https://doi.org/10.1017/S0266267113000217 Published online by Cambridge University Press

own motives for punishment' (2008: 207) since there is apparently 'a fundamental conflict within many individuals: they want and expect punishment to serve consequentialist goals, but when faced with a particular case they consistently choose punishments that serve retributive goals' (Carlsmith and Darley 2008: 207). However, while the studies by Carlsmith and colleagues might indeed show a kind of disconnect between people's theoretical claims about deterrence and their intuitions about cases, these studies do not permit any strong inferences about folk retributivism because the experimental materials confound the venerable distinction between a *criterion of rightness* and a *decision procedure*.

As has long been recognized by consequentialists, even if the consequences of an act are what make an act right, it does not follow that only forward-looking considerations are relevant in the optimal decision procedure for maximizing good consequences (see, e.g. Mill 1979 [1861]; Sidgwick 1874). Carlsmith and colleagues are right that severity and culpability loom large for retributivists. However, even if one is a consequentialist about punishment, it does not follow that one's decision procedure will neglect things like the severity of the crime or the culpability of the offender. What makes punishment right, according to the consequentialist, is a function of the consequences of punishment. But to achieve that moral goal might involve decision procedures that consult a wide range of factors. Indeed, both the severity of the crime as well as the offender's culpability are factors that could play an important role in the decision procedures of consequentialists. More severe offences produce more social outrage, and hence severity is relevant both to the need for deterrence and the need to provide victims and society with a means for expressing their moral disapprobation. And the culpability of an offender will typically be *predictively* valuable and so inform decisions on the most effective punitive strategies for rehabilitation and incapacitation.

Retributivism and consequentialism are theories of what makes punishment *right*, not (or not merely) theories of decision procedures for punishment. As a result, the claim that the folk are retributivists (or that the folk make judgements according to retributivist motives) is not just a claim about decision procedures. Rather, it is a claim about what the folk take to ultimately justify punishment – i.e. what they take the criterion of rightness or the *right-making* features to be. But once we allow the distinction between decision procedures and right-making accounts, we can see that the experiments by Carlsmith and colleagues fail to provide clear evidence that the folk reject consequentialism as the right-making account for punishment.

For instance, in one typical characterization, Carlsmith and colleagues present participants with the following as a general statement of retributivism: 'the major determinant of the sentence for a crime should

https://doi.org/10.1017/S0266267113000217 Published online by Cambridge University Press

be the sentence that the person deserves to receive given the seriousness of the crime he committed' (Carlsmith *et al.* 2002: 298). Although this statement is consistent with retributivism, it is naturally interpreted as characterizing a decision procedure for allotting punishment – namely, when we decide how much to punish we should apportion the punishment to the seriousness of the crime. And as noted above, consequentialists typically adopt decision procedures that attend carefully to the seriousness of the crime. Similar problems attend the vignette studies. The results show that people use a decision procedure that attends to the seriousness of the crime, but that doesn't entail that people are being driven by a truly retributivist principle.

The same problem arises for the other factors Carlsmith and colleagues associate with retributivism – the culpability of the offender, the extenuating circumstances surrounding the commission of the offence, and the moral outrage caused by the offence. In each case, consequentialists can (and arguably should) take these factors into consideration as part of their decision procedure for determining the most appropriate punishment. This shows that while consequentialists are interested in some factors that are admittedly irrelevant from the standpoint of retributivism – e.g. the general frequency of a crime or the likelihood that an offender will recidivate – the factors that Carlsmith and colleagues assume are *uniquely* retributivist may be relevant when it comes to the consequentialist's decision procedure even if these factors are not relevant to the consequentialist's criterion of rightness.

Indeed, one might worry that it is very difficult to isolate retributivist motives in concrete decisions about punishment. For instance, some prominent accounts of punishment hold that the benefit of punishment is given by its *communicative role* (Feinberg 1965; Lacey 1988; Braithwaite and Pettit 1990). And while communicative theories of punishment are often broadly retributive – focusing as they often do on the central importance of moral censure as the appropriate response to moral desert (von Hirsch 1993; Duff 2001) – consequentialists can quite consistently take the communicative effects of punishment into consideration. On this view, rather than focusing on the intrinsic value of conveying moral disapprobation to offenders, we ought to focus instead on the instrumental value of the communicative role played by punishment.

On the one hand, consequentialists could focus on the positive communicative effect that punishment has on *the public*. After all, by enabling the community to collectively vent their outrage in solidarity, punishment not only helps repair the damage done to society by giving the members of the community a common voice, but it also helps reinforce the underlying social norms, restore people's faith in the law, facilitate catharsis, and the like. Moreover, the public nature of punishment serves to convey the message to would-be offenders that society does not tolerate

https://doi.org/10.1017/S0266267113000217 Published online by Cambridge University Press

transgressors. Indeed, if punishments weren't public at least in some sense, they simply couldn't serve the communicative role required for general deterrence. On the other hand, rather than focusing on the positive effect that punishment has on society, consequentialists could focus instead on the communicative effect punishment has on the *wrongdoer*. On this view, punishment plays an essential role when it comes to moral education both by providing the offender with moral censure and by providing him with a prudential reason to refrain from crime in the future (regardless of whether he feels the sting of the moral censure itself). So, whether a consequentialist values the communicative benefits that punishment has on the public or on the offender (or both), the potentially instrumentally valuable communicative role played by punishment exacerbates the difficulty in showing that sentencing decisions are truly driven by retributive motives.

The extant psychological work on retributive punishment largely fails to address this *communication confound*, since it doesn't exclude the possibility that punishment is apportioned for its communicative role. As such, the otherwise important and illuminating research does not yet settle whether the folk are retributivists. In our attempt to shed some empirical light on this issue, we borrowed some ideas from a quite different research tradition – experimental economics. Social psychology focuses on judgements about state-sanctioned criminal punishment, but of course punishment itself is a much broader practice. Reducing the welfare of an individual as a result of perceived wrongdoing will count as punishment even if the welfare reduction is not conducted by the state or directed at criminals. In experimental economics, punishment is explored in the context of economic games. Research in this tradition shows that players will often pay to deduct money from other players who have defected or failed to reciprocate in the game. People will pay to deduct in these cases even when they know that they will never interact with the players again (Fehr and Gächter 2002). Indeed, even third parties will pay to deduct from players who violate norms (e.g. Fehr and Fischbacher 2004).

This work is commonly, and in our view rightly, presented as investigating punishment behaviour. Some philosophers maintain that the behaviour in these games is in fact driven by retributive impulses (see, e.g. Nichols 2007: 419; Sommers 2012: 37). However, the evidence doesn't yet establish that the punishment is retributive.[3] For the extant work in experimental economics, like the work in social psychology, bears the communication confound. In these experiments, the punisher expects the

[3] The experimental economists are not focused on retributivism. Rather, they tend to focus on the fact that the punisher can expect no personal monetary gain from punishing, suggesting that the punishment is 'altruistic' (e.g. Fehr and Gachter 2002).

https://doi.org/10.1017/S0266267113000217 Published online by Cambridge University Press

punishee to *know* when his income has been diminished. This raises the possibility that punishment is taken to have some communicative role, as many anti-retributivists have held. Although extant economic games do not address the communication confound, the economic game approach allows for a natural way to control for the confound because this approach allows us to design those games such that the punishee either knows or doesn't know about a reduction in welfare. This is exactly what we did for our first study.

## 3. FOLK RETRIBUTIVISM – THREE NEW STUDIES[4]

### Study 1

Our first study has two main parts, one involving concrete punishment decisions and another concerning general punishment philosophy. We used the experimental economic approach for the concrete punishment decisions. For our economic game, we used the 'strategy method', in which subjects are asked to indicate their preferences across a range of possibilities, even though only some of those preferences will be the basis for the payout. Our particular game was based on forthcoming studies by Jason Dana and colleagues in which participants are given an opportunity to destroy income of another, at a cost to themselves, based on various possible traits of the other. Dana and colleagues find that subjects do indeed often prefer to destroy the welfare of another at a cost to themselves. We adapted the idea of welfare destruction for our study on retributive judgement. In Dana's work, the income destruction was real, since the subjects really would get paid according to their preferences; for our first two studies, income destruction was presented as hypothetical.

In the first part of this study, some subjects made a hypothetical decision about allocation in a context in which the potential punishee would be aware of the reduction in welfare; other subjects made the allocation decision in a context in which the potential punishee would never be aware of any reduction in welfare. Our central prediction was that people would be no less likely to reduce welfare when the target would be ignorant (cf. Gollwitzer and Denzler 2009). That would show that the welfare reduction behaviour is not merely dependent on the communicative function of punishment, and it would thereby lend support to the idea that the welfare reduction is generated by a truly retributivist motive.

For the second part of our study, we had participants indicate agreement with an explicit characterization of a retributivist theory of punishment. We adopted a characterization that rectified the shortcomings of the renditions in Carlsmith and colleagues (2002). In

[4] The complete experimental materials are available as supplementary material online.

https://doi.org/10.1017/S0266267113000217 Published online by Cambridge University Press

particular, we stipulated as part of the retributive theory that punishing wrongdoers is appropriate even if it has no future benefits.

*Sample*. Participants were recruited from MTurk, an Amazon internet platform which provides researchers with access to a population of potential participants. Data obtained through MTurk is of comparable quality to data obtained through standard psychology subject pools, so long as the sample is restricted to US participants (Buhrmester *et al*. 2011). Participation was restricted to the USA, and participants were compensated $0.40 for their time. Eight-six participants began the study; 74 completed the task. Of those who completed the task, 44 were female. Participants ranged in age from 19 to 68, with a mean age of 34. As is typical with MTurk samples, the participants were diverse in ethnicity, educational background and employment.

*Materials*. Participants completed an online questionnaire divided into three sections.

*Section 1: Economic Game.* An economic game was described to participants, and they were asked to indicate their preferences as if they were actually taking part in the game. They received the following instructions for the game:

> You will be grouped randomly and anonymously with two other people in this room, also at computer stations. You will never be informed of the identity of these other people, either during or after the experiment. Similarly, the other participants will never be informed about your identity or about each other's identities. One participant will be designated as G and will be given $2. Another participant will be designated D, and that participant will be allowed to deduct $0, $1 or $2 from G (the money does not go to D or anyone else). You have been designated as the third participant, P. You have the opportunity to get some money and also to decide whether D gets a bonus payment. You need to decide among various options for what you would do, depending on what D does.

Participants were randomly assigned to one of two conditions, 'knowledge' or 'ignorance'. Participants in the knowledge condition saw the following sentence at the end of the scenario:

> D knows that he might get a bonus of up to $4, and he will know if you have deducted from his bonus.

Participants in the ignorance condition were told instead that because D *does not know* that he might get the bonus, he would *not know* whether anything had been deducted from his bonus. Participants were then asked to make decisions about how much money D would get under three

https://doi.org/10.1017/S0266267113000217 Published online by Cambridge University Press

different possible scenarios (D deducts either $2, $1 or $0 from G). For each scenario, participants were given the following choices:

> $4 for you, $4 for D
> $3 for you, $1 for D
> $2 for you, $0 for D

If participants selected an option on which D (and the participant) would get less than the optimal $4, we call this a 'welfare reduction' response.

*Section 2: Explicit Retributivism Question.* The next part of the experiment investigated explicit beliefs about the propriety of retributive punishment. Our primary interest was to measure beliefs in retributive punishment. But we also wanted participants to consider a wide range of potential justifications for punishment. So before they were given the key retributivism question, they were asked to rank the importance of the following different justifications for punishment (the order of the justifications was randomized):

1. People who commit crimes should be punished because by punishing them we give them what they deserve and giving offenders their just deserts is a good thing. [Desert]
2. People who commit crimes should be punished because by punishing them, we make it less likely they will harm others in the future. [Specific deterrence]
3. People who commit crimes should be punished because by punishing them, we make an example out of them, which makes it less likely others will engage in similarly harmful behaviour. [General deterrence]
4. People who commit crimes should be punished because by punishing them, we make it more likely they will be rehabilitated and safely reintegrated into society. [Rehabilitation]
5. People who commit crimes should be punished because by punishing them, we give the victims' family and friends the revenge they deserve. [Revenge]

The first item above was intended to approximate a retributivist consideration. Following the ranking, participants were asked to indicate their agreement with a more precise statement of an *explicit retributivist principle*:

> People who commit crimes deserve to be punished even if punishing them won't produce any positive benefits to either the offender or society – e.g. rehabilitation, deterring other would-be offenders, etc.

https://doi.org/10.1017/S0266267113000217 Published online by Cambridge University Press

Agreement with this statement was coded as indicating explicit agreement with retributivism.[5]

*Section 3: Personal Beliefs and Demographics.* Following the retributivist portions of the experiment, participants completed (i) an empathy measure (the subscales for *empathic concern* and *personal distress* from the Interpersonal Reactivity Index [Davis 1980; 1983]), (ii) Part 1 of the Free Will Inventory (Nadelhoffer *et al*. prep)[6] and (iii) a demographic questionnaire.

*Results*. There were 86 participants who began the study, though 12 dropped out over the course of the experiment. Of these 86, 39 (45.3%) were randomly assigned to the Ignorance condition, and 47 (54.7%) were randomly assigned to the Knowledge condition. In the Knowledge condition, 57% of participants gave a welfare-reduction response on the economic game. In the Ignorance condition, 69% gave a welfare-reduction response to the question. The difference between these proportions is not significant on a two-proportion $z$-test ($z = 1.12$, $p = 0.27$, NS); 95% CI ($-0.32$, 0.08). As expected, participants were more likely to reduce an individual's welfare if that individual had destroyed another's income in both the Ignorance condition (McNemar's test, $n = 39$, $p < 0.001$) and the Knowledge condition (McNemar's test, $n = 47$, $p < 0.001$).

To our surprise, several subjects indicated that they would reduce the welfare of player D even when D did not deduct at all. It is impossible to tell whether these subjects are retributivists in their responses, so we wanted to consider the data excluding these participants. If we exclude these subjects, and compare subjects who did not reduce welfare with subjects who reduced welfare when and only when D deducted from G, the results are largely the same: in the Knowledge condition, 49% of participants chose a welfare reduction when D had deducted from the other play; in the Ignorance condition 63% chose welfare reduction. Once again, there was no significant difference between conditions ($\chi^2$ (1, $n = 86$) = 1.35, $p = 0.34$). As a result, responses in the Knowledge and Ignorance conditions were collapsed for subsequent tests.[7] (See Figure 1 for frequency distributions in the economic game.)

[5] The retributivist consideration we used in the ranking task is less precise because it did not include the explicit stipulation that punishment is justified in the absence of positive benefits.

[6] See Appendix for the full scale. We included FWI because of the recent interest in both social psychology and experimental philosophy in the relationship between free will and judgements concerning moral responsibility. See, e.g. Nahmias *et al*. (2006); Nichols and Knobe (2007); Vohs and Schooler (2008); Monroe and Malle (2010); Shariff *et al*. (in prep).

[7] There were no gender differences in responses to the game, whether we include all participants ($\chi^2$ (1, $n = 73$) = 0.002, $p = 0.968$) or exclude those who reduced welfare when D did not deduct ($\chi^2$ (1, $n = 64$) = 0.009, $p = 0.924$).

https://doi.org/10.1017/S0266267113000217 Published online by Cambridge University Press

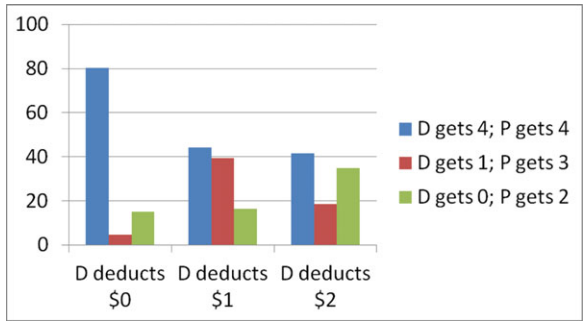



FIGURE 1. (Colour online) Frequency of responses in study 1.

For the ranking task, participants did *not* tend to rank the retributive-like consideration above all the consequentialist considerations (see Table 1). The highest ranking items were consequentialist: specific and general deterrence. More importantly, though, when asked simply to indicate agreement with the explicit retributivist statement, people tended to agree with the retributivist statement (M = 2.74) on a scale from 1 (strongly agree) to 6 (strongly disagree). This showed greater agreement than what would be expected by chance alone ($t$ (75) = −4.310, p < 0.001).[8] Women gave somewhat higher agreement (M = 2.52) than men (M = 3.14), but this difference was not significant (t (71) = 1.61, p = 0.099). We also conducted a pre-planned t-test comparing responses on the explicit retributivist statement by participants who reduced welfare in the game (N = 36, M = 2.58) and those who did not (N = 31, M = 2.97). (We excluded participants who reduced welfare even when D did not deduct.) We found no significant difference (t (65) = −1.01, p = 0.315).

Turning now to the correlational tests, for FWI, there was a positive correlation between the free will subscale – which measures strength of belief in free will – and agreement with the explicit retributivist question (r = 0.362, n = 74, p = 0.002), with strong beliefs in free will associated with strong agreement to the explicit retributivist question.[9] These results are unsurprising given the intimate relationship between free will and moral responsibility. After all, the more free will and responsibility one attributes to wrong doers, the more sense retributivist punishment makes. Sceptics

[8] Note that there is no contradiction in embracing retributivism and also ranking consequentialist considerations higher than retributive ones since it is perfectly sensible to hold that the justification for punishment is over-determined. One might consistently maintain that retributivism provides one reason to punish and that consequentialism provides other reasons to punish.

[9] These results are consistent with some recent findings which show that when people's beliefs in free will are challenged, they become less retributivistic (Shariff *et al*. in prep.).

https://doi.org/10.1017/S0266267113000217 Published online by Cambridge University Press

| | Desert | Specific deterrence | General deterrence | Rehabilitation | Revenge |
|---|---|---|---|---|---|
| % ranking item #1 | 13.2 | 43.4 | 14.5 | 22.4 | 6.6 |
| % ranking item #2 | 14.5 | 23.7 | 27.6 | 25 | 9.2 |
| % ranking item #3 | 14.5 | 18.4 | 35.5 | 17.1 | 14.5 |
| % ranking item #4 | 34.2 | 9.2 | 14.5 | 11.8 | 30.3 |
| % ranking item #5 | 23.7 | 5.3 | 7.9 | 23.7 | 39.5 |

TABLE 1. Rankings of justifications for punishment (study 1).

about free will and responsibility, on the other hand, categorically reject retributivism as a viable justification for punishment. So, the positive correlation we found between scores on the free will subscale of FWI and retributive punishment is precisely what we predicted. By contrast, there was no correlation between responses to the explicit retributivist statement and either the dualism subscale ($r = 0.043$, $n = 74$, $p = 0.713$) or the determinism subscale ($r = -0.122$, $n = 74$, $p = 0.3$).

For the empathy measures, there were no significant correlations between the explicit retributivism question and either Personal Distress ($r = -0.08$, $n = 74$, $p = 0.498$) or Empathic Concern ($r = -0.127$, $n = 74$, $p = 0.28$).

*Discussion*. As predicted, we found that participants' behaviour on the economic game was unaffected by whether the punishee would be ignorant of the welfare reduction. Indeed, a slightly higher proportion of participants in the ignorance condition preferred to impose welfare reduction. This suggests that the responses cannot be explained by supposing that participants are assigning punishment for its communicative role. This helps to secure the claim that welfare reduction in this experiment is generated by a retributive principle, not by a purely consequentialist criterion. In addition, many people explained their welfare reduction preference by citing the wrongness of D's behaviour. For instance:

- 'Because it is a form of stealing when D takes away $2 from G. Even worse, the money doesn't go to anyone. That is just evil, and D does not deserve to receive money because of this, even if I have to take a cut in my earnings.'

https://doi.org/10.1017/S0266267113000217 Published online by Cambridge University Press

- 'It didn't seem fair that D would take all of G's money but then get $4. If D didn't take any money from G, he should be rewarded for his good behaviour.'
- 'Someone who deducted money just for fun is obviously a jerk and shouldn't be rewarded for their behaviour.'[10]

In addition to the apparently retributivist behaviour in the economic game, participants tended to agree with the explicit statement of a retributive principle. This suggests that people in Western culture embrace the retributivist principle and rely on it in making decisions about welfare reduction. So, whereas Carlsmith and colleagues focused on deterrence and found a disconnect between explicit theory and practice, we focused on retributivism and found that, broadly speaking, there is a coincidence between theory and practice.

We've been arguing that the welfare reduction responses in our economic task reflect a retributive motive. However, an alternative explanation is that people are reducing welfare in an attempt to generate income equity among participants. Reducing the income of D (the deducter) when he has deducted from G brings D's income into greater parity with G's income. Evidence from experimental economics indicates that egalitarian motives do plausibly play a role in these kinds of games (see e.g. Fowler *et al.* 2005) even if they cannot explain all punitive behaviour in economic games (Fehr and Fischbacher 2004; Fehr and Gachter 2005). An equity-based explanation might explain some of the behaviour in the economic game in Study 1. Indeed, a few participants in Study 1 invoked equity considerations in explaining why they preferred welfare reduction – e.g. '[I] tried to be fair and give everybody the same' and 'I tried to be fair to the participants. So I tried to do was best for all of us'. Other participants gave explanations that were vague enough to be consistent with an equity motivation, e.g. 'I think it would be fair'. As a result, it was important to explore the extent to which egalitarian motives can explain preferences for welfare reduction. So, we ran a follow-up study.

## Study 2

To investigate the egalitarian interpretation, for our second study, we revised the economic game by making the welfare reduction *increase* inequity. Thus, if participants prefer welfare reduction, this cannot be explained in terms of an egalitarian motive. Another important difference between this economic game and the earlier one is that we designed the game so that the participant can reduce the welfare of the other player

[10] All of these participants were in the Ignorance condition.

https://doi.org/10.1017/S0266267113000217 Published online by Cambridge University Press

without having to sacrifice any of his or her own payment. We did this for two reasons. First, it is much simpler to generate clear egalitarian vs. non-egalitarian options if we factor out self interest. Second, the participants in Study 1 who did *not* prefer welfare reduction typically invoked self-interest as the reason for their choice. So we wanted to know whether a greater proportion of people would prefer welfare reduction if it came for free. Note that the fact that the welfare reduction is free does not mean it isn't retributive punishment. Experimental economists tend to focus on costly punishment because they want to explore the extent to which people are willing to endure costs in order to exact punishment. However, in the present context, our interest is in whether people operate with a retributive principle, and punishment is no less retributive just because it is free.

*Sample.* As with Study 1, participants were recruited from MTurk. Participants were compensated $0.40 for their time, and participation was restricted to the USA. Forty-five participants began the study, 41 of which completed the task. Of those who completed the task, 21 were male, and 20 were female. Participants ranged in age from 18 to 70, with a mean age of 37.

*Materials.* The three-section questionnaire used in Study 2 included the new economic game, the ranking task and the explicit retributivism question from Study 1, the free will scale, and a scale to measure belief in egalitarianism.

*Section 1: The New Economic Game.* Since there was no difference between the ignorance and knowledge conditions in Study 1, the knowledge condition was eliminated for Study 2. All participants were presented with a game in which the target of deduction would never know about any reduction in welfare. Instructions for the new economic game were the same as before, except this time participants were (a) told that G would be paid $4, (b) told that they had been designated as the third participant P, and (c) instructed as follows:

> You will receive $2. And you will decide whether D gets a bonus payment. You need to decide among various options for what you would do, depending on what D does.

Participants were then asked to make decisions about how much money D would get under three different possible scenarios (D deducts $2, $1, or $0 from G). For each scenario, participants indicated whether they thought D should receive $2, $1, or $0.

*Section 2: Explicit Retributivism Question.* The questions used for this section were identical to the question used in Study 1.

*Section 3: Egalitarianism, Free will beliefs, and Demographics.* Following the retributivist portions of the experiment, participants completed (i) a

https://doi.org/10.1017/S0266267113000217 Published online by Cambridge University Press

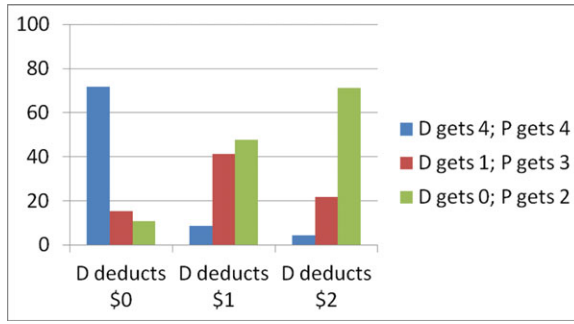



FIGURE 2. (Colour online) Frequency of responses in study 2.

free will questionnaire identical to that used in Study 1, (ii), a demographic questionnaire, and (iii) part of the Social Dominance Orientation scale (Pratto *et al.* 1994), used with reverse coding to measure egalitarianism.[11]

*Results*. There were 45 participants who began the study, three of whom dropped out during the course of participation. Of those that began participation, 95.6% (43 out of 45) preferred a welfare-reduction response. Once again, we found that several participants chose to reduce D's welfare even when D had not deducted. If we exclude these subjects, we are left with 33 participants, and the results are the same; 97% of the remaining participants opted to reduce, only 3% (a single subject) opted not to reduce. Once again, participants were more likely to reduce an individual's welfare if that individual had destroyed another's income (McNemar's test, $n = 45$, $p < 0.0001$). Since virtually all participants preferred welfare reduction, the pre-planned t-test was not conducted for this component of the study.[12] (See Figure 2 for frequency distributions in the economic game.)

On the ranking task, once again the highest ranking items were consequentialist (see Table 2). But again, more importantly, for the explicit retributivist question, people tended to agree with the retributivist statement ($M = 2.68$ on a scale from 1 (strongly agree) to 6 (strongly disagree)). As in study 1, this showed greater agreement than what would be expected by chance alone ($t(40) = -2.985$, $p = 0.005$). Once again, women gave somewhat higher agreement ($M = 2.25$) than men ($M = 3.1$), but this difference was not significant ($t(39) = 1.57$, $p = 0.124$). No

[11] We created the egalitarianism subscale from the social dominance orientation scale by omitting items 3, 7, 11, 12 and 14 from the scale. We are grateful to Kathleen Vohs and Adam Waytz for suggesting this as a measure of egalitarianism.

[12] Again, there were no gender differences in responses to the game, whether we include all participants ($\chi^2$ (1, n = 41) = 0.976, p = 0.323) or exclude those who reduced welfare when D did not deduct ($\chi^2$ (1, n = 29) = 0.967, p = 0.326).

https://doi.org/10.1017/S0266267113000217 Published online by Cambridge University Press

| | Desert | Specific deterrence | General deterrence | Rehabilitation | Revenge |
|---|---|---|---|---|---|
| % ranking item #1 | 19.5 | 36.6 | 22.0 | 9.8 | 12.2 |
| % ranking item #2 | 17.1 | 29.3 | 24.4 | 22.0 | 7.3 |
| % ranking item #3 | 31.7 | 17.1 | 14.6 | 19.5 | 17.1 |
| % ranking item #4 | 14.6 | 9.8 | 26.8 | 19.5 | 29.3 |
| % ranking item #5 | 17.1 | 7.3 | 12.2 | 29.3 | 34.1 |

TABLE 2. Rankings of justifications for punishment (study 2).

significant correlations were observed between the single retributivism question and the egalitarian subscale created from the Social Dominance Orientation scale ($r = -0.193$, $n = 41$, $p = 0.226$).

For FWI, as in study 1, there was a significant correlation between the free will subscale and agreement with the explicit retributivist question ($r = 0.314$, $n = 41$, $p = 0.045$), with strong beliefs in free will associated with strong agreement to the explicit retributivist question. Once again, there was no correlation between the explicit retributivist question and either the dualism subscale ($r = 0.243$, $n = 41$, $p = 0.126$) or the determinism subscale ($r = 0.283$, $n = 41$, $p = 0.073$).

*Discussion*. Study 1 left open the possibility that welfare reduction in the economic game was a product of an egalitarian motive rather than a retributive one. Study 2 puts that explanation to rest. Nearly all participants (95%) preferred welfare reduction even though this would increase inequity. This provides further reason to think that people in these games are motivated to reduce welfare for retributivist reasons. Moreover, the fact that virtually everyone chose welfare reduction in this condition suggests that the smaller proportion of welfare reducers in Study 1 is not to be explained by lack of a retributive motive, but to the costliness of the punishment. When punishment is free, almost everyone responds as a retributivist. While we believe the results from Study 1 and Study 2 shed some new light on the role that retributive motives play when it comes to the psychology of punishment, the fact that our economic games were merely hypothetical rather than monetized was admittedly an important limitation. So, while asking participants to play hypothetical economic games was an important first step, we also wanted to ensure that our findings held up in situations within which

https://doi.org/10.1017/S0266267113000217 Published online by Cambridge University Press

there are real economic consequences. In order to address this concern, we ran a monetized follow-up study to see whether our earlier findings would generalize to the behavioural realm.

## Study 3

For our third study, we once again utilized a strategy method whereby some of the participants were provided with the opportunity to reduce the welfare of others – except this time, the welfare reduction had real-world consequences and was not just part of an imagined scenario. Each participant was randomly assigned to play one of three roles within a group of three people. Two of the roles involved decisions concerning whether to reduce the welfare of another member of the group (see below for details). Mirroring Study 1, some participants in the third study had to make punishment decisions in contexts within which the punishee would know he had been punished and some participants had to make punishment decisions in contexts within which the punishee would not know. The goal was to once again try to control for the communication confound in an effort to see whether participants' punitive behaviours are driven (at least partly) by purely retributive motives. After playing the monetized economic game, participants were then asked to note their level of agreement with explicit retributive statements used in the first two studies and asked to provide some basic demographic information.

*Sample*. As with Studies 1 and 2, participants were recruited from MTurk. Participants were compensated $0.50 for participating, with the possibility of further payment depending on which condition they were assigned to and the decisions of other players. Participation was restricted to the USA. A total of 193 participants began the study; 13 failed comprehension checks and were excluded from the analysis, leaving 180 participants. Of those 180, 78 were female.

*Materials*. Participants were randomly assigned to one of five conditions and they completed an online questionnaire that was divided into three sections.

*Section 1: Economic Game.* For starters, each participant received the general instructions. Participants were then presented with the following description of the economic game:

> The game will be described in terms of 'tokens'. Each token is worth 12 cents. After the game, you will receive (through MTurk), a sum of money equivalent to the sum of tokens you have. Here again, the sum of money you receive for tokens will be in addition to the money you receive for the HIT.

https://doi.org/10.1017/S0266267113000217 Published online by Cambridge University Press

At this point, participants were randomly assigned to one of five conditions: Knowledge C (KC), Ignorance C (IC), Knowledge B (KB), Ignorance B (IB) and A.

*Conditions C*: Participants in both KC and IC received the following general information:

> As noted above, you have been grouped (randomly and anonymously) with two other participants. One participant, designated as A, is scheduled to receive 20 tokens. Another participant, designated as B, will be allowed to deduct 0, 10 or 20 tokens from A (the deducted tokens do not go to B or anyone else). The last participant, C, will receive 10 tokens.
>
> You have been randomly assigned to be participant C. Recall that B is in a position to deduct tokens from A. In addition, B is scheduled to receive a payment of up to 10 tokens. But you get to decide whether to deduct tokens from B, depending on what he decides about deducting from A's tokens. If you deduct tokens from B, they would not go to A or you or anyone else.
>
> Before you make your decision, we want to explain the situation fully.
>
> You will decide about deducting from B for three cases, corresponding to *all three possible decisions of B*. Recall that B is in a position to deduct 0, 5 or 10 tokens from A (again, the deducted tokens don't go to B or anyone else). After we finish explaining the situation, you will be asked to decide about each possibility. After the experiment, we will check B's decision about deducting from A, and we will then look at your corresponding decision about that possibility. We will then follow through on your decision about that case and implement your decision.

At this point, participants in KC received the following:

> Before you make your decision about these three possibilities, we want to clarify that B knows that he is scheduled to receive 10 tokens. He also knows that participant C is in a position to have tokens deducted from his account depending on his decision. So if he receives less than 10 tokens, he will know that participant C has deducted from his account (though, of course, your identity is completely anonymous).

Participants in IC, on the other hand, were reminded that B *does not know* he is scheduled to receive any tokens, so he will *not know* whether anyone has deducted from his account. Participants in both KC and IC were then asked to state what they would do if (a) B deducts 0 tokens from A (leaving A with 20 tokens), (b) B deducts 5 tokens from A (leaving A with 15 tokens), and (c) B deducts 10 tokens from A (leaving A with 10 tokens). In each of these three cases, participants were asked to specify whether they would:

- Deduct 10 tokens from B (leaving B with 0 tokens).

https://doi.org/10.1017/S0266267113000217 Published online by Cambridge University Press

- Deduct 5 tokens from B (leaving B with 5 tokens).
- Deduct 0 tokens from B (leaving B with 10 tokens).

Having made their decision, participants were then asked to explain their choices. They were also asked whether the fact that B would know (or would not know) whether they deducted from his funds affected their decision concerning whether to deduct from B's funds.

*Conditions B*: Participants in both KB and IB received the following information:

> As noted above, you have been grouped (randomly and anonymously) with two other participants. One participant, designated as A, is scheduled to receive 20 tokens. Another participant, designated as B, will be allowed to deduct 0, 10, or 20 tokens from A (the deducted tokens don't go to B or anyone else). The last participant, C, will receive 10 tokens.

Participants in KB then received the following information:

> You have been randomly assigned to be participant B. That means that you get to decide how much, if any, money, will be deducted from A. Before you make your decision, we want to explain the situation fully. Recall that A is scheduled to receive 20 tokens. You are scheduled to receive 10 tokens. However, the total number of tokens you get might be 10, 5 or 0, depending on what C decides. C will make decisions corresponding to each of your three possible decisions. In particular, C will decide whether (and how much) to deduct if you (i) deduct 10 tokens from A, (ii) deduct 5 tokens from A, (iii) deduct no tokens from A. After the experiment, we will check your decision and then look at what C decided about that option. We will then follow C's decision and deduct 10, 5 or 0 tokens from your scheduled allotment of 10 tokens.

Participants in IB, on the other hand, received the following information:

> You have been randomly assigned to be participant B. That means that you get to decide how much, if any, money will be deducted from A. Recall that A is scheduled to receive 20 tokens.

Participants in both KB and IB were then asked to select to do one of the following:

- Deduct 10 tokens from A (leaving A with 10 tokens).
- Deduct 5 tokens from A (leaving A with 15 tokens).
- Deduct 0 tokens from A (leaving A with 20 tokens).

Having made their decision, participants in both conditions were then asked to explain their choices. All participants in Condition C were compensated with an additional $1.20. Participants in condition B were compensated with an additional $0, $0.60 or $1.20, depending on the

https://doi.org/10.1017/S0266267113000217 Published online by Cambridge University Press

decisions of the corresponding player in the C condition. Participants in condition A were compensated with an addition $1.20, $1.80, or $2.40, depending on the decisions of the corresponding player in the B condition.[13]

*Section 2: Explicit Retributive Question.* Each participant was asked to state their level of agreement with the explicit retributive question we used in Studies 1 and 2 on a scale from 1 (strongly disagree) to 6 (strongly agree).

*Section 3: Demographics.* Upon participating in the economic game and noting their agreement with the explicit retributive statement, participants were asked to respond to a demographic questionnaire.

Finally, each participant was asked to (a) state their level of agreement with the explicit retributive question we used in Studies 1 and 2, and (b) respond to a demographic questionnaire.

*Results. Condition C.* A total of 67 participants started this condition, 7 failed comprehension checks and were excluded from analysis. Half of the remaining 60 participants were randomly assigned to the Ignorance condition and the other half to the Knowledge condition. In the Ignorance condition, 77% (23 out of 30) preferred a welfare-reduction response; in the Knowledge condition, 83% (25/30) preferred a welfare-reduction response. The difference between these proportions is not significant on a two-proportion z-test ($z = 0.646$, $p = 0.519$, NS); 95% CI ($-0.14$, 0.27). As before, participants were more likely to reduce welfare if the individual had destroyed another's income in both the Ignorance condition (McNemar's test, $N = 30$, $p < 0.01$) and the Knowledge condition (McNemar's test, $N = 30$, $p < 0.001$).

As in the other studies, several participants chose to reduce a player's welfare even when that player hadn't deducted. Interestingly this only occurred in the Ignorance condition, though. In the Ignorance condition 13% of participants chose to reduce the welfare of player B when he had not deducted. In the Knowledge condition, not a single subject did this. One natural explanation is that people would be more cautious about such apparently malicious treatment when the player would know. In any case, if we exclude these subjects, the results are largely the same: in the Knowledge condition, 83% of participants chose a welfare reduction when but only when B had deducted from the other player; in the Ignorance condition 73% chose welfare reduction under those conditions. Once again, there was no significant difference between conditions ($\chi^2$ (1, n =

[13] Since the participants in the A condition were not asked to do anything, we have omitted the description of that condition. See online materials for complete details (supra note 5).

https://doi.org/10.1017/S0266267113000217 Published online by Cambridge University Press

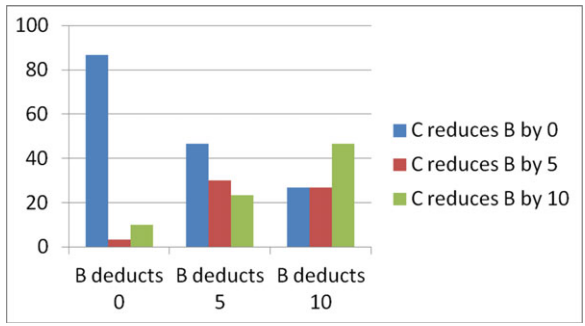



FIGURE 3. (Colour online) Frequency of responses in the Ignorance Condition (Study 3).

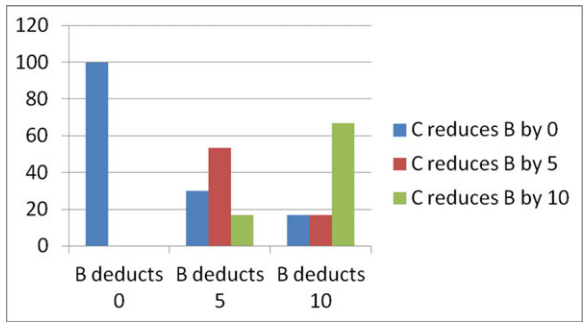



FIGURE 4. (Colour online) Frequency of responses in the Knowledge Condition (Study 3).

56) = 0.87, p = 0.35). (See Figures 3 and 4 for frequency distributions in these two conditions.) [14]

Since there was no significant difference in responses in the Knowledge and Ignorance conditions, responses in these conditions were collapsed for subsequent tests. For the explicit retributivist question, people tended to agree with the retributivist statement (M = 4.33 on a scale from 1 (strongly disagree) to 6 (strongly agree)). As in Studies 1 and 2, this showed greater agreement than what would be expected by chance alone (t(59) = 6.44, p < 0.001). We also conducted a pre-planned t-test comparing responses on the explicit retributivist statement by participants who reduced welfare in the game (N = 44) and those who did not (N = 12). (We excluded participants who reduced welfare even when B did

[14] There were no gender differences in responses in the Ignorance condition, whether we include all participants ($\chi^2$ (1, n = 30) = 0.096, p = 0.757) or exclude those who reduced welfare when D did not deduct ($\chi^2$ (1, n = 26) = 0.001, p = 0.973). In the Knowledge condition, there were too few female participants to run the test.

https://doi.org/10.1017/S0266267113000217 Published online by Cambridge University Press

not deduct.) We found a tendency for those who reduced welfare to give higher responses on the retributive question ($M = 4.432$) than those who did not reduce welfare ($M = 3.833$); however, this difference did not reach significance ($t(54) = 1.86$, $p = 0.068$).

The critical issue for us is whether the fact that the punishee would be ignorant would undermine the decision to reduce the person's welfare. We found that it did not significantly do so. In addition, we asked participants in the ignorance condition the following question: 'Did the fact that B would not know whether you deducted from his funds affect your decision about whether to deduct from his funds? Please answer Yes or No and explain your answer.' Here are some representative answers:

- 'No, since the survey is anonymous I did not factor whether or not B would know that tokens would be deducted into my decision. I felt that if B did or did not know it would not matter since I felt that is what he deserved for deducting for no good reason.'
- 'No, not really. I feel that B should get the same treatment he/she is giving to A, and it doesn't affect my decision if B knows it or not.'
- 'No. He gets what he deserves either way.'
- 'No. My decision is based on principle; not to teach B a lesson.'

Many subjects thus explicitly maintain that their punishment was the right thing to do despite the punishee's ignorance.

*Condition B.* Although our primary interest is in condition C, we also analysed responses from condition B. To our surprise, many participants opted to deduct money from participant A, even though this money would not go to any other player. In other contexts, this behaviour has been dubbed *money burning* (Zizzo and Oswald 2001). In the Knowledge condition (in which player B was told that C would know whether they deducted), 7 out of 30 players burned money; in the Ignorance condition, 12 of 30 burned money. The difference between those conditions is not significant ($\chi^2$ $(1, n = 60) = 1.926$, $p = 0.165$). But it is a striking fact that so many participants opted for money burning. Since we found no difference between conditions, we collapsed these conditions for the subsequent tests. And, again to our surprise, we found a significant gender effect on this task: 44% (16 out of 36) female participants opted to deduct from A; while only 12.5% (3/24) of the male participants did. This is a significant difference ($\chi^2$ $(1, n = 60) = 6.79$, $p = 0.009$). We did not predict this difference and have no good explanation for it. One possibility is that what is driving the effect is an egalitarian impulse. Previous research suggests that women are more egalitarian than men (Andreoni and Vesterlund 2001; Dufwenberg and Muren 2006). And by burning A's money, one brings A (who is scheduled to get 20 tokens) closer to parity with the other players. We see this kind of reasoning made explicit by some participants. For instance, one female participant in the Ignorance condition explained

https://doi.org/10.1017/S0266267113000217 Published online by Cambridge University Press

her decision as follows: 'I have no tokens, Player C has 10 tokens, I saw it fair to make it a little more equal in the game'. In any case, this is an issue that invites further research.

*Discussion*. The results on our monetized economic game largely paralleled the results in Studies 1 and 2. Of particular significance, a clear majority of people in both the knowledge and ignorance conditions still chose to reduce welfare. The proportion of such retributive punishment is higher in the hypothetical version of the task in Study 2 (97% of participants reduced welfare) than in the monetized version of the present study (73%), but there were other differences in the studies, so it is hard to know whether monetizing the task made the difference. In any case, the important point is that people still punish even when the punishee will not know a thing.

## 4. GENERAL DISCUSSION

Unlike previous work on retributivism and the psychology of punishment, our findings cannot easily be explained in terms of consequentialist decision procedures. Participants were presented with an explicit presentation of retributivism which made clear that punishing the wrongdoer would not yield any positive benefits. Yet participants overwhelmingly endorsed retributivism thusly defined – which indicates that people are retributivists in theory. The results from our economic games provide evidence that people are also retributivists in practice. In all three studies, a clear majority of participants opted to reduce a wrongdoer's welfare even when the wrongdoer wouldn't know about the reduction; this suggests that people favour punishment even in the absence of any communicative effects on the offender. Moreover, the fact that willingness to pay to punish correlates with responses to the explicit retributivism question further supports the interpretation that a retributivist motive is operative in the welfare reduction responses to our economic games. Finally, our findings cannot be explained by an egalitarian motive since in Study 2, participants were willing to reduce welfare even though doing so *increased* inequity. In short, our findings suggest that people are committed to retributivist punishment in both theory and practice.

Although our studies correct some shortcomings of previous work, there are several critical issues that our studies leave unresolved. First, although we took pains to exclude a communicative role for punishment, it is possible that when participants punish, they are doing so with an eye to communicating with the *experimenter*. We find it unlikely that this can explain much of the punishment in our studies, but it would be useful to see whether people punish less (or more) in conditions of heightened anonymity, e.g. when the participants believe that their

https://doi.org/10.1017/S0266267113000217 Published online by Cambridge University Press

responses will only be seen by experimenters after being aggregated (cf. Hoffman *et al*. 1994). Second, while we found that people's endorsement of retributivism was correlated with their belief in free will, it is an open question exactly what drives the correlation. Recent work has explored how people understand the relationship between free will, determinism and responsibility (e.g. Nahmias *et al.* 2006; Nichols and Knobe 2007; Monroe and Malle 2010), and we hope that our present findings will encourage research on the relationship between people's views about moral agency and their views on punishment. Finally, and perhaps most importantly, our population was drawn from the USA, and it is unclear whether our results will extend to other cultures as well. To be sure, cross-cultural experimental economics on punishment has been full of surprises (e.g. Hermann *et al.* 2008), and we would not want to proclaim any confidence that our results will generalize to other cultures. But for our culture, there appears to be a powerful strand of folk retributivism. Adequately tracing the contours of folk retributivism both here and abroad, however, is a task for another day – one which will require researchers to continue to reach across disciplinary boundaries in an effort to better understand why we punish.

## SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit http://dx.doi.org/doi:10.1017/S0266267113000217

### REFERENCES

Andreoni, J. and L. Vesterlund. 2001. Which is the fair sex? Gender differences in altruism. *Quarterly Journal of Economics* 116: 293–312.

Braithwaite, J. and P. Pettit. 1990. *Not Just Deserts*. Oxford: Oxford University Press.

Buhrmester, M., T. Kwang and S. Gosling. 2011. Amazon's Mechanical Turk: a new source of inexpensive, yet high-quality, data? *Psychological Science* 6: 3–5.

Carlsmith, K. M. 2006. The roles of retribution and utility in determining punishment. *Journal of Experimental Social Psychology* 42: 437–451.

Carlsmith, K. M. 2008. On justifying punishment: The discrepancy between words and actions. *Social Justice Research* 21: 119–137.

Carlsmith, K. M. and J. M. Darley. 2008. Psychological aspects of retributive justice. *Advances in Experimental Social Psychology* 40: 193–236. San Diego, CA: Elsevier.

Carlsmith, K. M., J. M. Darley and P. H. Robinson. 2002. Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology* 83: 284–299.

Cottingham, J. G. 1979. Varieties of retributivism. *Philosophical Quarterly* 29: 238–246.

Dana, J., C. Dave and E. Friedman. (forthcoming). Nosy preferences and welfare destruction in the laboratory. University of Pennsylvania.

Darley, J. M., K. M. Carlsmith and P. H. Robinson. 2000. Incapacitation and just deserts as motives for punishment. *Law and Human Behavior*, *24*, 659–683.

https://doi.org/10.1017/S0266267113000217 Published online by Cambridge University Press

Davis, M. H. 1980. A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology* 10: 85.
Davis, M. H. 1983. Measuring individual differences in empathy: evidence for a multidimensional approach. *Journal of Personality and Social Psychology* 44: 113–126.
Duff, R. A. 2001. *Punishment, Communication, and Community*. New York: Oxford University Press.
Duff, R. A. and D. Garland. 1994. Introduction: Thinking about punishment. In *A Reader on Punishment*, ed. R. A. Duff & D. Garlan, 1–43. Oxford: Oxford University Press.
Dufwenberg, M. and A. Muren. 2006. Gender composition in teams. *Journal of Economic Behavior and Organization* 61: 50–54.
Fehr, E. and S. Gächter. 2002. Altruistic punishment in humans. *Nature* 415: 137–140.
Fehr, E. and S. Gachter. 2005. Human behaviour: Egalitarian motive and altruistic punishment (reply). *Nature* 433: E1–E2.
Fehr, E. and Fischbacher, U. 2004. Third party punishment and social norms. *Evolution and Human Behavior* 25: 63–87.
Feinberg, J. 1965. The expressive function of punishment. *The Monist* 49: 397–423.
Fowler, J., T. Johnson and O. Smirnov. 2005. Egalitarian motive and altruistic punishment. *Nature* 433: E1.
Gollwitzer, M. and M. Denzler. 2009. What makes revenge so sweet: seeing the offender suffer or delivering a message? *Journal of Experimental Social Psychology* 45: 840–844.
Hermann, B., C. Thoni and S. Gachter. 2008. Antisocial punishment across societies. *Science* 319: 1362–1367.
Hoffman, E., K. McCabe, K. Shachat and V. L. Smith. 1994. Preferences, property rights and anonymity in bargaining games. *Games and Economic Behavior* 7: 346–380.
Hogarth, J. 1971. *Sentencing as a Human Process*. Toronto: University of Toronto Press.
Kahneman, D., D. Schkade and C. R. Sunstein. 1998. Shared outrage and erratic awards: the psychology of punitive damages. *Journal of Risk and Uncertainty* 16: 49–86.
Lacey, N. 1988. *State Punishment: Political Principles and Community Values*. London: Routledge.
McFatter, R. M. 1978. Sentencing strategies and justice: effects of punishment philosophy on sentencing decisions. *Journal of Personality and Social Psychology* 36: 1490–1500.
McFatter, R. M. 1982. Purposes of punishment: effects of utilities of criminal sanctions on perceived appropriateness. *Journal of Applied Psychology* 67: 255–267.
Mill, J. S. 1979 [1861]. *Utilitarianism*. Indianapolis, IN: Hackett Publishing.
Monroe, A. E. and B. F. Malle. 2010. From uncaused will to conscious choice: the need to study, not speculate about, people's folk concept of free will. *Review of Philosophy and Psychology* 1: 3–4.
Moore, M. 1997. *Placing Blame: A General Theory of the Criminal Law*. Oxford: Clarendon Press.
Nadelhoffer, T., E. Nahmias and C. Sripada. (in prep). The Free Will Inventory: Measuring beliefs about free will.
Nahmias, E., S. Morris, T. Nadelhoffer and J. Turner. 2006. Is incompatibilism intuitive? *Philosophy and Phenomenological Research* 73: 28–53.
Nichols, S. 2007. After incompatibilism: a naturalistic defense of the reactive attitudes. *Philosophical Perspectives* 21: 405–428.
Nichols, S. and J. Knobe. 2007. Moral responsibility and determinism: the cognitive science of folk intuitions. *Nous* 41: 663–685.
Pratto, F., J. Sidanius, L. M. Stallworth and B. F. Malle. 1994. Social dominance orientation: a personality variable predicting social and political attitudes. *Journal of Personality and Social Psychology* 67: 741–763.
Shariff, A. F., J. D. Greene and J. W. Schooler. (in prep). His brain made him do it: encouraging a mechanistic worldview reduces punishment.

https://doi.org/10.1017/S0266267113000217 Published online by Cambridge University Press

Sidgwick, H. 1874. *Methods of Ethics*. London: Macmillan.
Sommers, T. 2012. *Relative Justice: Cultural Diversity, Free Will, and Moral Responsibility*. Princeton, NJ: Princeton University Press.
Vidmar, N. 1974. Retributive and utilitarian motives and other correlates of Canadian attitudes toward the death penalty. *Canadian Psychologist* 15: 337–356.
Vohs, K. and J. Schooler. 2008. The value of believing in free will: encouraging a belief in determinism increases cheating. *Psychological Science* 19: 49–54.
Von Hirsch, A. 1993. *Censure and Sanctions*. Oxford: Oxford University Press.
Zizzo, D. and A. Oswald. 2001. Are people willing to pay to reduce others' incomes? *Annales D'économie et de statistique* 63–64: 39–65.

https://doi.org/10.1017/S0266267113000217 Published online by Cambridge University Press