Evolution and Human Behavior xxx (2017) xxx-xxx

Contents lists available at ScienceDirect

Evolution and Human Behavior

journal homepage: www.elsevier.com

The psychology of deterrence explains why group membership matters for third-party punishment☆

Andrew W. Delton[a, b, c, *, 1], Max M. Krasnow[d, 1]

[a] Department of Political Science, Stony Brook University, Stony Brook, NY 11794-4392, United States
[b] College of Business, Stony Brook University, Stony Brook, NY 11794-4392, United States
[c] Center for Behavioral Political Economy, Stony Brook University, Stony Brook, NY 11794-4392, United States
[d] Department of Psychology, Harvard University, Cambridge, MA 02451, United States

ARTICLE INFO

Article history:
Received 8 December 2016
Received in revised form 21 July 2017
Accepted 24 July 2017
Available online xxx

Keywords:
Deterrence
Third-party punishment
Intergroup relations
Evolutionary psychology

ABSTRACT

Humans regularly intervene in others' conflicts as third-parties. This has been studied using the third-party punishment game: A third-party can pay a cost to punish another player (the "dictator") who treated someone else poorly. Because the game is anonymous and one-shot, punishers are thought to have no strategic reasons to intervene. Nonetheless, punishers often punish dictators who treat others poorly. This result is central to a controversy over human social evolution: Did third-party punishment evolve to maintain group norms or to deter others from acting against one's interests? This paper provides a critical test. We manipulate the ingroup/outgroup composition of the players while simultaneously measuring the inferences punishers make about how the dictator would treat them personally. The group norm predictions were falsified, as outgroup defectors were punished most harshly, not ingroup defectors (as predicted by ingroup fairness norms) and not outgroup members generally (as predicted by norms of parochialism). The deterrence predictions were validated: Punishers punished the most when they inferred that they would be treated the worst by dictators, especially when better treatment would be expected given ingroup/outgroup composition.

1. Introduction

We are often opinionated about others' conflicts and occasionally even intervene. From Twitter wars raging around a celebrity's infidelity, to boycotts of businesses, states, or entire countries for their treatment of sexual minorities, to the good Samaritan detaining a mugger trying to make off with a stolen purse, third-parties are often provoked by the bad actions of others.

In humans, researchers have usually studied one particular type of such third-party intervention: *third-party punishment*. Third-party punishment involves third parties punishing someone for treating another person poorly (Fehr & Fischbacher, 2004). Third-party punishment has been seen in industrialized societies, in small-scale societies, in both laboratory experiments and field experiments, and among children as young as 6 (Fehr, Fischbacher, & Gächter, 2002; Henrich et al., 2010; Kurzban, Descioli, & O'Brien, 2007; McAuliffe, Jordan, & Warneken, 2015).

Third-party punishment is also a group-based phenomenon (McAuliffe & Dunham, 2016). People often punish more when the victimizer is an outgroup member or when the victim is an ingroup member (Bernhard, Fischbacher, & Fehr, 2006; Lieberman & Linke, 2007). Group-based third-party punishment occurs both for real-world groups and for artificial laboratory groups (Goette, Huffman, & Meier, 2006; Jordan, McAuliffe, & Warneken, 2014; Schiller, Baumgartner, & Knoch, 2014). But, why does ingroup/outgroup status matter for third-party punishment?

Different theories of third-party punishment make different predictions about why group membership should matter. One theory, *group norm maintenance theory*, suggests that people engage in third-party punishment to enforce ingroup norms. Group norm researchers have primarily studied two such norms. The norm of fairness requires ingroup members to split resources fairly with other ingroup members. The norm of parochialism requires that ingroup members treat outgroup members poorly when possible. Another theory, *deterrence theory*, suggests that people engage in third-party punishment as the output of a cue-driven, evolved psychology designed to deter poor treatment of oneself and one's allies. Deterrence theory suggests that one driver of punishment is the inferences punishers draw: Punishers should punish more when they infer that poor treatment of third parties reflects a disposition by the actor to treat the self or valued others poorly.

Despite the differences between the theories, testing between them has proved difficult and only a few studies have attempted to do so (Bone, Silva, & Raihani, 2014; Jordan, Hoffman, Bloom, & Rand, 2016; Jordan et al., 2014; Krasnow, Cosmides, Pedersen, & Tooby, 2012; Krasnow, Delton, Cosmides, & Tooby, 2016). The goal of the

? The raw data for this paper can be found on the Open Science Framework at: https://osf.io/6gyyd/.

* Corresponding author at: Department of Political Science, Stony Brook University, United States.
Email addresses: andrew.delton@stonybrook.edu (A.W. Delton); krasnow@fas.harvard.edu (M.M. Krasnow)

[1] Both authors contributed equally to this manuscript.

http://dx.doi.org/10.1016/j.evolhumbehav.2017.07.003
1090-5138/

present study is to investigate how differential group membership affects third party punishment by observing the inferences punishers draw from dictator behavior. If the deterrence view is correct, group membership should matter because of how it changes the inference punishers draw about how the dictator would treat them or those they value personally. For example, based on seeing an outgroup dictator treat an ingroup member poorly a punisher should infer that the dictator will also treat her poorly; such cases license the inference that the poor treatment was due to the victim's group membership, a property which the punisher shares, causing the inference to generalize. In contrast, this inference should be much weaker when seeing an ingroup dictator treat an outgroup member poorly. If the group norm view regarding the fairness norm is correct, group membership should matter because the fairness norm most properly applies to behavior within the group. If the parochialism norm is operative, we should expect general poor treatment of outgroup members. Notably, neither norm specifies how punishment should relate to inferred personal treatment, particularly in contrast to inferred treatment of others. We elaborate on these theories below.

1.1. Punishment as norm maintenance

One class of theories explains third-party punishment as flowing from a human ability to create and maintain group norms. On this group norm maintenance view, humans have an evolved psychology designed to acquire social norms from the local social environment, act on them, and enforce them in others (Chudek & Henrich, 2011; Fehr & Fischbacher, 2004; Henrich et al., 2006, 2010; Richerson & Boyd, 2005). A social norm is a learned rule that specifies both an action to be taken (or not) and simultaneously specifies punishment of people who do not obey the norm.

Norms are shared within groups, but might differ between groups—they are rules applied by a community on people within the community. This is important for making concrete predictions from group norm maintenance theory. As Chudek and Henrich (2011, p. 218) write, "By norms, we mean learned behavioral standards shared and enforced by a community." Again illustrating that norms are an ingroup phenomenon, Richerson and Boyd (2005, p. 219) write that humans "are inclined to punish fellow group members who violate social norms, even when such punishment is costly." A given norm, whatever it is, regulates behavior within a community. By punishing people who violate a norm, punishment has at least two effects: changing the norm violator so they follow the norm in the future and cueing other members of the group that norm violations will be punished.

Group norm maintenance theory also holds that people enforce norms regardless of personal benefits—punishing a norm breaker need not be in service of any anticipated direct benefits from punishing. This feature is often called "strong reciprocity" (Gintis, 2000). As Fehr and Henrich (2003, p. 57 emphasis original) write, "The essential feature of strong reciprocity is a willingness to sacrifice resources in … punishing unfair behavior, *even if this is costly and provides neither present nor future economic rewards for the reciprocator*."

There are many variations on group norm maintenance theory and many potential norms. A single paper cannot possibly investigate them all. Instead, we focus on the most prominent version of the theory—cultural group selection—and the most commonly studied norms—fairness and parochialism. On theories of culture group selection, virtually any norm is possible. This is because, on this theory, norm psychology uses *moralistic punishment*: not only are people who break the norm punished, but people who do not punish norm breakers are also punished (and, in principle, people who do not punish those who do not punish are punished, ad infinitum). Moralistic punishment can sustain any norm, even ones deleterious for the group or individual (Boyd & Richerson, 1992). So, if a group norm specifies burning down group members' homes, people who do not commit arson should be punished. Moreover, people who commit arson but do not punish non-arsonists should *also* be punished (and up through higher levels).

Although any norm, useful or harmful, is possible, cultural group selection theory holds that the distribution of norms will not be random. Instead, group-beneficial norms should tend to predominate. In part, this is because a process of cultural selection happens between groups. Groups with norms favoring ingroup prosociality will tend to replace groups without such norms. This could happen because groups with more effective norms grow and reproduce faster or survive longer than other groups (Boyd, Gintis, Bowles, & Richerson, 2003). Or such norms could allow an ingroup to directly compete with outgroups, such as in war, and thereby replace those outgroups (Choi & Bowles, 2007; Gintis, 2000). This does not necessarily require that individual group members be killed; merely that members of dissipated groups join more effective groups or adopt their norms (Chudek & Henrich, 2011; Richerson & Boyd, 2005).

By far the most commonly studied potential norm is the fairness norm (Fehr & Fischbacher, 2004; Fehr et al., 2002; Henrich et al., 2006, 2010). This norm specifies that ingroup members should treat each other "fairly," typically meaning that a windfall gain should be split (more or less) evenly. For instance, if an experimenter randomly hands one subject of a pair $10, then the fairness norm specifies that this subject should give $5 to the other subject. Fairness norms have been suggested to underpin the amazing economic success of Western cultures (Henrich et al., 2010).

The other most consistently studied norm is parochialism (Choi & Bowles, 2007). Parochialism is often conceptualized as having two components: ingroup altruism or fairness (essentially the fairness norm discussed above) and outgroup aggression, spite, or derogation (Rusch, 2014). Parochialism's norm of outgroup derogation requires that ingroup members hurt, injure, or otherwise inflict costs on outgroup members when possible. (From this point on, whenever we refer to "parochialism," we will be referring to the outgroup derogation side.) Because norms are about ingroup members regulating other ingroup members' behavior, parochialism is *not* a norm that specifies how *out*group members should behave. Instead, the parochialism norm specifies how *in*group members should be behave towards outgroup members.

Proponents of the view that human third-party punishment flows from group norm maintenance point to a number of sources of evidence (Richerson et al., 2016). A chief source of evidence is that third-party punishment occurs when punishers cannot seemingly expect any material returns. For instance, third-parties will punish in anonymous, one-shot laboratory experiments. Typically, these experiments involve the third-party punishment game. One player, the dictator, is given (e.g.) a $10 stake. The dictator can divide the stake any way she sees fit between herself and another player, the recipient. The recipient has no say over this allocation. Finally, a third player, the punisher, knows how much the dictator allocated to the recipient. The punisher has a separate stake of (e.g.) $5 and can spend it to reduce the dictator's earnings. Dictators are aware in advance that punishers exist and can punish.

Because the experiment is one-shot and anonymous, punishers have no material incentive to punish: They do not know the recipient's or dictator's identity, nor will punishers knowingly interact with either again. Thus, punishers have no strategic reasons to spend on

punishment. Dictators, realizing this, have no material incentive to give anything to recipients for similar reasons. Nonetheless, people regularly punish in these experiments (Fehr & Fischbacher, 2004; Fehr et al., 2002; Goette et al., 2006; Henrich et al., 2010; Jordan, McAuliffe, & Rand, 2015; Jordan et al., 2014; Jordan et al., 2016; Krasnow et al., 2016; McAuliffe et al., 2015; Schiller et al., 2014).

On group norm maintenance theory, this reveals that third-party punishment has been organized to maintain group norms. Consistent with this account, third parties punish more when an ingroup member treats another ingroup member poorly than when an outgroup member treats another outgroup member poorly (Bernhard et al., 2006); this follows because norms regulate ingroup members' behavior, not outgroup members' behavior.

1.2. Punishment as deterring poor treatment of the self and valued others

A different perspective holds that third-party punishment flows from an evolved deterrence psychology designed to deter poor treatment of oneself or valued others (Krasnow et al., 2012, 2016; Lieberman & Linke, 2007; McCullough, Kurzban, & Tabak, 2013; Sell, Tooby, & Cosmides, 2009). Many resources are rivalrous. This leads to conflicts over who will consume such resources—conflicts of interests. Animals often defend their interests with force, or threats of force, and anticipate that others will do the same. But there will usually be uncertainty about what another animal's interests are and how strongly that animal is able to defend them. Thus, miscoordinations will be common, for example, between how much someone respects your interests and the level of respect you feel entitled to.

The existence of others with a disposition to act against your interests is an adaptive problem; should they continue to act as they have they will continue to impose fitness costs on you. Getting them to improve their behavior requires recalibrating representations in their mind, for example representations about your ability to use force to defend your interests. Many animals use punishment or the threat of punishment to change the behavior of others (Clutton-Brock & Parker, 1995; Raihani, Thornton, & Bshary, 2012). Punishment among animals straightforwardly increases inclusive fitness by causing the punisher to be treated better, or causing close kin to be treated better. For instance, a mother might use threats or aggression to drive predators away from her offspring.

Unlike many animals, however, humans create enduring friendships, alliances, and coalitions. Can deterrence straightforwardly extend to these cases? We believe it can. Friends are often irreplaceable and effective coalitions are often difficult to recreate. These features make them intrinsically valuable, much in the same way genetic relatives are intrinsically valuable (Tooby & Cosmides, 1996; Tooby, Cosmides, & Price, 2006), making it valuable to punish on their behalf (Lieberman & Linke, 2007; Roos, Gelfand, Nau, & Carr, 2014). Moreover, friends and allies may reciprocate deterrence: I help deter poor treatment of you now, you help me in the future. Just as it is easy to understand why pastoralists would deter poachers from stealing their animals, it is easy to understand why people would deter poor treatment of their valuable relationship partners. This predicts that humans should, at least in some instances, engage in group-based third-party punishment: If I benefit by the continued existence of a strong coalition, I am directly incentivized to defend its interests.

Punishment can also defend or secure the punisher's own reputation to bystanders. Such punishment could signal to observers who are not part of the dispute that the punisher is willing to enforce her interests (Krasnow et al., 2016) or has other valuable traits (Kurzban et al., 2007), such as trustworthiness (Jordan et al., 2016).

How does deterrence psychology work? At the deep level of evolutionary game theory, the potential cost of being punished must outweigh the benefits of treating others poorly; otherwise, continued poor treatment would still be best. Proximally, however, any given act of punishment or sanctioning is not usually deterring through its effect on immediate payoffs (Ostrom, 1998). Instead, it functions as a signal to the malefactor that such a behavior must stop, or else later sanctions will be more severe. Moreover, this signaling logic is not unique to theories of deterrence in humans. Bluffs and ritualized strength displays are common across animals.

Punishment as signaling is consistent with recent evolutionary models of anger (McCullough et al., 2013; Sell, 2011; Sell et al., 2009). Typically, the expression of anger is not costly aggression, but talking or arguing with the person who caused the anger, to get them to change their behavior (Averill, 1983). Similar results obtain in political science and economic research. Based on substantial fieldwork, Ostrom (1998, p. 8) writes that punishment involves "graduated sanctions for enforcing compliance" and that "by paying a modest fine, they [malefactors] rejoin the community in good standing and learn that rule infractions are observed and sanctioned." This is consistent with experimental economic research showing that purely nominal "disapproval points," which were cost-free to give and receive, were nearly as effective as costly punishment in maintaining cooperation in a public goods game (Masclet, Noussair, Tucker, & Villeval, 2003). To accomplish its deterrence function, punishment or sanctioning must change the malefactor's behavior; it is not strictly required that the malefactor be damaged—the threat of future aggression or withdrawal of benefits can be sufficient to change behavior.

Of course, talk is cheap and bad actors should not find threats universally credible. When bad actors appear to need a bigger signal to change their behavior, people switch to cost-imposing deterrence: "Repeated rule breakers are severely sanctioned and eventually excluded from the group" (Ostrom, 1998, p. 8). Alternatively, if bad behavior signals a severe enough disposition towards future bad behavior, punishment may immediately escalate to higher levels of severity (Kurzban & DeScioli, 2013; McCullough et al., 2013; Sell et al., 2009). Although in many cases punishment or sanctioning might be a pure signal—that is, merely a threat of future harm or withdrawal of benefits—in more extreme cases it can involve immediate costs. Thus, a key prediction of deterrence theory is that the greater disparity between how much you infer someone values you (given how they acted) and what you feel entitled to from them, the greater the punishment that is predicted (Kurzban & DeScioli, 2013; McCullough et al., 2013; Sell et al., 2009).

Generally, in the third-party punishment game, punishers do not have access to how they would be personally treated by the dictator. But they can infer this disposition based on dictators' treatment of the *recipient*. In support of this a previous study showed that punishers use dictators' treatment of *recipients* to infer how dictators would treat the punishers themselves (Krasnow et al., 2016). This inference was ecologically valid: Dictators' treatment of recipients predicted their treatment of punishers.

Deterrence theory, therefore, offers two mechanisms whereby third-party punishment would be differentiated by group membership. First, because ingroup members are intrinsically valuable, people should be more willing to punish on their behalf compared to outgroup members. Second, the relative group membership should change the inference the punisher makes about how much the dictator values the punisher based on how the dictator treats the recipient. If an ingroup member has been mistreated, you can more reasonably infer that this mistreatment would extend to you if the culprit was an outgroup member. This inference in and of itself should license pun-

ishment. In contrast, if an ingroup member mistreats an outgroup member, that likely does not predict how you will be treated by that ingroup member.

But how does a deterrence account explain the apparent irrationality of one-shot, anonymous third-party punishment? After all, in an anonymous, one-shot game punishment can have no rational deterrent effect. Deterrence theory assumes that, in the small-scale social worlds of human ancestors, a person who treats someone else poorly now might later treat you, your kin, or your allies poorly in the future (Krasnow et al., 2016). Much like craving sugar-rich foods was adaptive in the past, but may be harmful in abundant modern environments, an evolved punishment psychology may treat the anomalous situations of anonymous, one-shot laboratory games as if they represented more typical conditions where relationships and reputations persist over time (Delton, Krasnow, Cosmides, & Tooby, 2011; Hagen & Hammerstein, 2006; Krasnow, Delton, Tooby, & Cosmides, 2013; West, Griffin, & Gardner, 2007). Deterrence theory does not argue that all third-party punishment will be beneficial, nor that it was always beneficial in the past. Rather, the argument is that because the long-run average returns from attempting to deter bad treatment were positive, our present psychology bears this design.

In our view, deterrence theory is more parsimonious than group norm maintenance theory (see the Discussion). Thus, even if both theories' predictions and explanatory power entirely overlapped, we believe that deterrence theory should be favored. However, we also believe the two theories can be empirically distinguished and doing so is more productive than mere argument. After describing our methods, we lay out specific predictions from the theories.

2. Method

2.1. Participants

We analyzed data from 275 punishers (39% women; 60% liberal) and 303 dictators (44% women; 60% liberal) recruited from Amazon Mechanical Turk (Mturk), an online labor force. Participants came from Mturk's US worker pool. Responses to economic games played on Mturk are similar to responses in laboratory settings (Horton, Rand, & Zeckhauser, 2011), including in the third-party punishment game (Krasnow et al., 2016). All participants received $0.50 merely for playing and earned bonuses based on gameplay. See SI for additional details and for results as a function of liberalism/conservatism.

Although using Western, educated samples is sometimes seen as a flaw in research, here we think it is the most appropriate sample. This is because it gives best advantage to group norm maintenance theory, especially the fairness norm. The fairness norm has been offered as an explanation for the West's amazing economic success (Henrich et al., 2010). Thus, if it exists, it should certainly be operating in our US sample.

2.2. Procedure

Players participated in triads composed of a dictator, recipient, and punisher. The game was played over the Internet and gameplay was anonymous and one-shot. All decisions and guesses by dictators and punishers were incentivized with real money and no deception was used.

We used players' identification as liberals or conservatives to manipulate the ingroup/outgroup composition of the triads. We experimentally manipulated group composition: all three players from the same ingroup (i.e., same ideological persuasion), the dictator outgroup to both recipient and punisher, the dictator and recipient outgroup to the punisher, and the recipient and punisher outgroup to the dictator. (Neutral terminology was used in the actual materials, see SI.)

Dictators first completed an incentivized *valuation task* that measured how much they valued the material welfare of recipients and, separately, punishers (Delton & Robertson, 2016; Kirkpatrick, Delton, Robertson, & de Wit, 2015; Krasnow et al., 2016; Tooby, Cosmides, Sell, Lieberman, & Sznycer, 2008). This task assesses how much money dictators would forgo in order to give a fixed $1.00 to the other player and has been validated to predict behavior in other contexts (Delton & Robertson, 2016; Krasnow et al., 2016). The more money they are willing to forgo, the more they value the other player. Scores on the valuation task could range 0 to 0.75, with greater scores representing greater valuation. A score of, e.g., 0.25 implies a dictator would pass up as much as $0.25 to give $1.00 to their recipient.

Next, dictators made *divisions*, by dividing a fixed stake of $1 between themselves and the recipients, with allowable divisions of $1/$0 (0% transferred to recipients), $0.75/$0.25 (25% to recipients), and $0.50/$0.50 (50% to recipients). These stakes are typical of Mturk.

Punishers began by completing two tasks in the absence of any knowledge of dictators' behavior. First, punishers guessed the actual division the dictator would make (correct answers earned $0.10). Punishers were next shown the same valuation task that the dictator completed and asked to complete it *as they believed their dictator would*, measuring punishers' inferences about how much dictators valued the other players (correct answers earned $0.02 each).

Punishers then considered, one by one, the three possible divisions dictators could make to recipients: 0%, 25%, or 50% to recipients. Punishers completed the valuation task as they believed the dictator would under each of these assumed divisions (correct answers earned $0.02 each, though paid only for those corresponding to the dictator's actual division). Under each of these assumed divisions, punishers also decided how much they would punish the dictator—how much they would spend of a $0.50 stake to reduce the dictator's earnings. Each $0.10 punishers spent reduced the dictator's stake by 20% (Krasnow et al., 2016). Using a percentage reduction allows punishers to reduce dictators' earnings to zero, regardless of the division, while eliminating the possibility of reducing them below zero. Punishers' punishment decisions were enacted based on dictators' actual divisions.

To be clear, punishers only responded to the possible divisions dictators could make. They did not learn dictators' actual behavior (neither divisions nor valuation decisions) until after the experiment was over and payouts made. This "strategy method" has been frequently used to assess third party punishment and does not alter these behaviors relative to other methods (Jordan et al., 2015).

2.3. Predictions

Given this design, what predictions do the two approaches make? Group norm maintenance theories describe how group norms regulate the behavior of ingroup members; almost by definition, a norm could not apply to an outgroup member. Thus, in the context of a third-party punishment game—when considering the fairness norm—the proper target of punishment is ingroup members failing to share with other ingroup members. Therefore, punishment should be strongest when all three players are from the same ingroup. Punishment should be weakest when an outgroup member treats another outgroup member poorly—no ingroup norms are at stake here. Finally, there may be moderate punishment when an ingroup member treats an outgroup member poorly or vice versa; because only one player is ingroup, this

situation partially implicates ingroup norms. Fig. 1 summarizes the predictions for the fairness norm.

What predictions would a group norm maintenance approach make about the parochialism norm, particularly the outgroup derogation component? One interpretation is that whenever people have the chance to damage outgroup members they should do so; this predicts strong punishment of outgroup dictators *regardless of their recipient's group membership and regardless of whether they shared* (Fig. 1). The least ambiguous place to test this prediction is when the dictator splits the money equally: parochialism necessarily predicts that outgroup members should be punished substantially more. This norm should also predict that ingroup members who treat outgroup members *well* should be punished. That is, ingroup members should be punished *more* as they share *more* with outgroup members.

Note also that group norm maintenance theory does not require that punishers make any particular inferences about how dictators value punishers nor does it require that punishment be related to these inferences of valuation. Such inferences about personal valuation by malefactors are irrelevant to maintaining those malefactors' compliance with ingroup norms. The disconnect between valuation and punishment follows from the insistence that people should punish unfairness "even if this is costly and provides neither present nor future economic rewards" for the punisher (Fehr & Henrich, 2003, p. 57 emphasis removed). Finding that punishment tracks perceptions about personal costs would undermine this assumption of group norm maintenance theory.

Deterrence theory makes different predictions about the effect of group membership on punishment. At the most general level, this theory hypothesizes that punishment is used to deter poor treatment of the self or valued others. Moreover, deterrence theory also incorporates whether poor treatment of third-parties is likely to predict low valuation of the self or valued others. Thus, this theory predicts the most punishment when an outgroup member treats an ingroup member poorly. Such intergroup behavior is likely to reflect the relationship between the groups, rather than anything specific to the dyad. Thus, the punisher can predict that she (or other ingroup members) will be treated poorly by an outgroup member who treats an ingroup member poorly. For similar reasons, an ingroup member who treats an outgroup member poorly should be punished the least; his actions are unlikely to predict how he will treat other ingroup members. Moderate punishment should be elicited when an ingroup member treats an ingroup member poorly or an outgroup member treats an outgroup member poorly. That these dictators are willing to treat members of their own group poorly suggests they might generally be willing to treat others like the punisher poorly, but also may reveal something else is effecting their behavior besides group membership (Krasnow et al., 2016). See Fig. 1.

Of course, group-based third-party punishment has already been studied and its empirical patterns are already known (Bernhard et al., 2006; Goette et al., 2006; Jordan et al., 2014; Schiller et al., 2014). Although the deterrence theory is somewhat supported by cogently retrodicting these patterns, strong evidence can come only from testing novel predictions. Thus, we turn to the valuation task.

The deterrence theory predicts that punishers' inferences of valuation should follow the same pattern described above for punishment. For example, punishers should infer that they are not valued much by an outgroup dictator if they observe that dictator treating an ingroup recipient poorly. Moreover, punishers should *not* infer especially low valuation if they observe an ingroup dictator treating an outgroup recipient poorly. Throughout, punishers' inferences of low valuation from dictators should predict punishment. Importantly, this parallel prediction for valuation reduces the degrees of freedom for the ordering of punishment based on the group membership conditions: Given the ordering of the conditions based on punishment, on the deterrence theory the order should apply to valuation inferences as well.

To be clear, we have discussed two separate routes for punishment under deterrence theory. One route is that punishers may punish on behalf of valued others, independently of how they personally expect to be treated. If your child is bullied by a classmate, you do not have to expect poor treatment yourself before you step in. The other route uses poor treatment of others to predict poor treatment of the self or valued allies, which further licenses punishment. The valuation task is specifically designed to assess this second route. These multiple routes mean that in some cases punishment is "overdetermined." For instance, when an outgroup member treats an ingroup member poorly, punishment should result both because punishers are defending a valuable ingroup member *and* because they infer that they (or other valuable ingroup members) could later be targets of bad treatment by the outgroup member.

3. Results

3.1. Basics of division and punishment

As in past research on the third-party punishment game, dictators sometimes gave money to recipients in the division task: 61% transferred at least \$0.25 and 37% transferred half of the stake (Fig. 2). Dictators valued both recipients and punishers at about 0.27, meaning dictators would forgo up to \$0.27 to give \$1.00 to punishers and recipients.

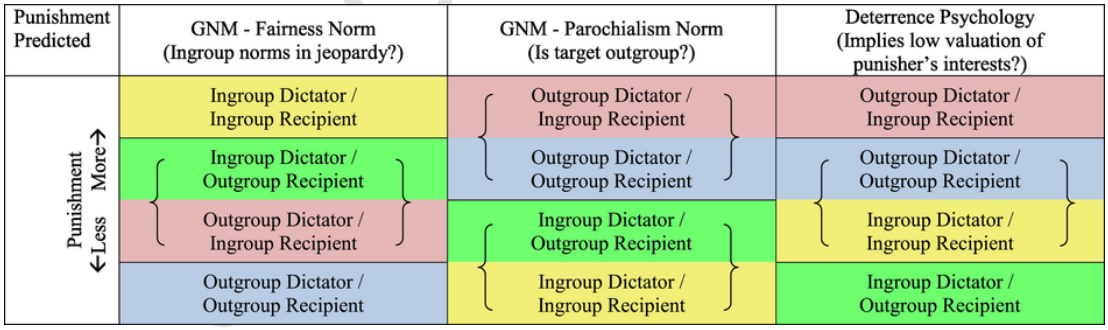

Fig. 1. Predictions from group norm maintenance theory and deterrence psychology theory. Colors correspond to the same experimental conditions as in Figs. 4 and 5. Conditions connected by brackets are not predicted to be different; ordering within the brackets is arbitrary. For the fairness norm and deterrence theory, the predictions are most cleanly tested when dictators share nothing; for the parochialism norm, when dictators split the stake evenly. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

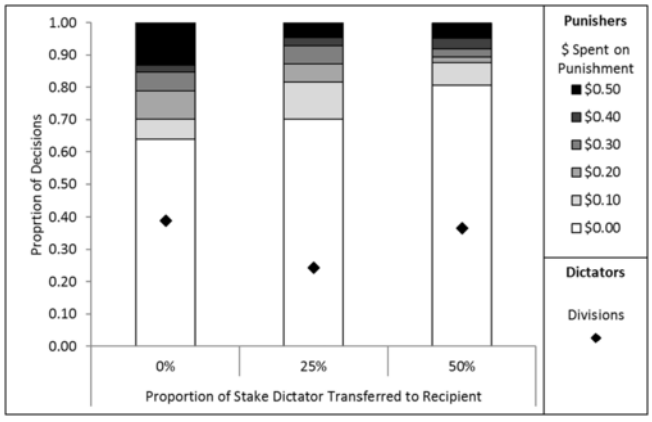

Fig. 2. Bars show punishers' punishment decisions as a function of how much of the stake dictators transferred to recipients. Diamonds show the proportion of dictators making each transfer decision. For punishers, $N = 275$; for dictators, $N = 303$.

Punishers sometimes punished: For example, when dictators transferred nothing, 36% spent some money on punishment and 13% spent their complete $0.50 stake (Fig. 2). Replicating past work, punishers punished more when dictators transferred less to recipients (Fig. 3; $F(2, 273) = 17.88$, $p < 0.001$, partial $\eta^2 = 0.12$).

3.2. The link between punishment and inferences of valuation

The present experiment replicated past research on the relationship between punishment and inferences regarding valuation (Krasnow et al., 2016); see SI. To start with, dictators' divisions, their valuations of recipients, and their valuations of punishers were all tightly correlated (*r*s ranged 0.47 to 0.82). This means that punishers can use dictators' treatment of *recipients* to predict dictators' treatment of *punishers*.

Given this connection between treatment of others and the self, and as predicted by deterrence theory, when punishers believed dictators did not value the punisher and the recipient, they punished more (Fig. 3). A punisher who expects zero valuation punishes about $0.09 more (17% of the stake) than a punisher who expects 0.75 valuation (zero to 0.75 spans the possible range of the valuation measure; see Tables S1–3).

But these results are correlational; they do not experimentally show that manipulating the predictive link between treatment of recipients and inferred valuation of punishers affects inferences and third-party punishment. To do this, we turn to our manipulation of group composition.

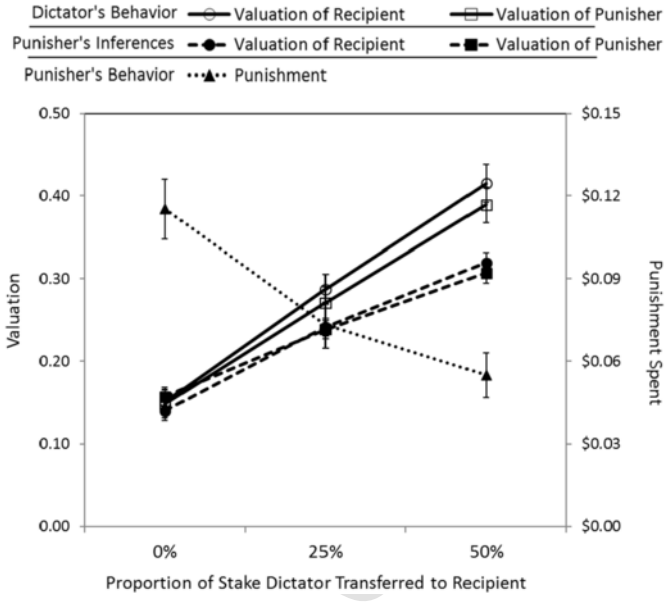

Fig. 3. Left vertical axis graphs dictators' mean valuation decisions and punishers' mean inferences about valuation, as a function of dictators' transfer to recipients (solid and dashed lines). Right vertical axis graphs punishers' mean spending on punishment (dotted line). Error bars ± 1 SEM. For punishers, $N = 275$; for dictators, $N = 303$.

3.3. Does punishment depend on group composition?

Yes. As shown in Fig. 4, group composition changes how much punishment is delivered as a function of transfer (for the interaction, $F(6, 542) = 2.65$, $p = 0.015$, partial $\eta^2 = 0.03$; see Table S4). Differences in punishment based on group composition are especially strong when dictators transfer nothing ($F(3, 271) = 5.39$, $p = 0.001$; $\eta^2 = 0.06$). This is primarily driven by the condition where an outgroup dictator treats an ingroup recipient poorly (see Table S5). Punishment also differs by group composition when dictators transfer only a quarter of the stake ($F(3, 271) = 2.95$, $p = 0.033$, $\eta^2 = 0.03$) and this was again driven primarily by the condition where an outgroup dictator treated an ingroup recipient poorly. Punishment did not vary by group when dictators transferred half the stake ($F(3, 271) = 1.20$, $p = 0.31$, $\eta^2 = 0.01$).

Importantly, the specific pattern of punishment as a function of group composition follows from the deterrence theory (see Fig. 1). Punishers punish most when the dictator is outgroup to both the punisher and recipient. On deterrence theory, this strong punishment occurs because outgroup behavior towards one ingroup member generalizes to other ingroup members like the punisher. Punishers punish least when an ingroup member treats an outgroup member poorly. On deterrence theory, this low punishment occurs because such behavior will not generalize to other *in*group members like the punisher (though it might to other *out*group members). Punishers punish a moderate amount when the dictator and recipient are both from the ingroup or both from the outgroup. On deterrence theory, this moderate punishment occurs because, although there is no group-based element, people who treat others poorly now may treat you or valued others poorly in the future.

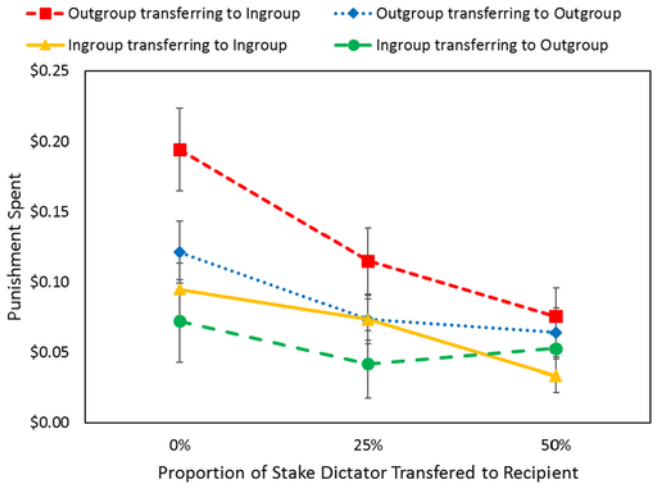

Fig. 4. Mean spending on punishment as a function of group composition and dictators' transfers. Error bars ± 1 SEM. $N = 275$. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

The pattern does not follow that predicted by the fairness norm (see Fig. 1). Using the data from the condition where dictators shared nothing, we can use focused contrasts to test the difference between the pattern predicted by deterrence theory and that predicted by the fairness norm (Rosenthal, Rosnow, & Rubin, 2000). First, a contrast for the pattern predicted by deterrence theory significantly accounted for variance in punishment ($t(271) = 3.84$, $p < 0.001$, $r^2 = 0.05$; see SI for contrast codes). Second, a contrast for the pattern predicted by the fairness norm did not significantly predict variance in punishment ($t(271) = -0.93$, $p = 0.35$, $r^2 = 0.00$). Finally, the deterrence contrast predicted more variance than the fairness norm contrast ($t(271) = 3.47$, $p = 0.001$, $r^2 = 0.04$).

Another way to test the fairness norm is to focus on two conditions that provide a particularly clear test. Because the fairness norm regulates ingroup behavior, punishment should be greater when an ingroup member treats an ingroup member poorly compared to when an outgroup member treats an outgroup member poorly. Our data provided no evidence for this straightforward prediction (Fig. 4 and Table S5). We think this failure is particularly revealing because, even if other norms (including ones we haven't considered) are at play, the fairness norm should still be visible in this focused comparison.

The parochialism norm also fared poorly. Outgroup members were not punished especially strongly irrespective of how they treated the recipient. This is clearest when dictators split the money equally; punishment did not significantly differ in this condition (see above analysis and Table S5). Moreover, punishers did not punish ingroup members *more* when they shared more with outgroup members (green line, Fig. 4); if anything, the trend was less punishment. That is, we observe no second-order punishment of parochialism norm violators. So, while participants use the context of group membership to inform their decision to punish the dictator, punishing some outgroup dictators more than some ingroup dictators, these data cannot be explained merely by parochialism per se.

Readers might notice that even in the condition where an ingroup member treats an outgroup member poorly—transferring nothing—punishers still punish, spending about $0.07 of their $0.50 stake. This might reflect a slight inference that this ingroup member will treat the punisher or valued others poorly in the future. But it also might be an experimental artifact. Research using a game without group information showed that once punishing is no longer the only option available to participants—they can also reward—participants no longer punish on average (Pedersen, Kurzban, & McCullough, 2013). Had reward been an option in the present design, it is possible that no punishment would have been observed in this condition. Importantly, however, the experimental artifact explanation, in and of itself, cannot explain why third-party punishment varies by group membership, as shown here and elsewhere (e.g., Bernhard et al., 2006; Lieberman & Linke, 2007).

Although the results from this section are as predicted by deterrence theory, they resemble previously known results. To show that deterrence theory also makes unique and novel predictions about group-based third-party punishment we next turn to punishers' inferences about dictators' valuation of punishers and recipients.

3.4. Do punishers' inferences about how much dictators' value recipients depend on group composition?

When punishers estimate the dictators' valuation of the recipient, they can do so with full knowledge of the dictators' transfer to the recipient. As such, we would not expect punishers' inferences about how much dictators' value recipients to vary by group composition. And they do not (at all three levels of transfer, $Fs(3271) \leq 1.65$, $ps \geq 0.18$.

3.5. Do punishers' inferences about how much dictators' value punishers depend on group composition?

Testing this prediction is the key test that our experiment was designed for. Deterrence theory predicts that punishers' inferences should depend on whether dictators' treatment of *recipients* will predict dictators' valuation of *punishers*. This should vary by group composition. And it does, for all three levels of transfer ($Fs(3271) \geq 4.24$, $ps < 0.01$, $\eta^2s \geq 0.04$; Table S5). Fig. 5 displays punishers' inferences about how much dictators value punishers when dictators transfer nothing to recipients.

As can be seen in Fig. 5, punishers infer that an outgroup dictator values them poorly when that outgroup dictator treats an ingroup member poorly; punishers assume poor treatment of an ingroup member will generalize to another ingroup member, namely themselves. Conversely, punishers infer that an ingroup dictator values them relatively highly when that ingroup dictator treats an outgroup recipient poorly; punishers do not assume that treatment of an *out*group member will generalize to another *in*group member, namely themselves.

In between these two cases, punishers infer that dictators value them a moderate amount when all three players are from the same group. This likely reflects several competing factors: For example, that the dictator treated a member of her own group poorly suggests she may treat others poorly in the future; but, in contrast to the outgroup dictator/ingroup recipient case, the poor treatment is less likely due to the mere fact of the recipient's group membership.

Across these three conditions the ordering of punishers' inferences about how much dictators value punishers is identical to the ordering of actual punishment (mirror-reversed, of course; see Fig. 5). In other words, when punishers expect they personally will not be valued by the dictator, they are more likely to punish bad behavior by the dictator—even when the bad behavior was directed at an otherwise anonymous third-party. Recall also that punishers did *not* infer any differences in dictators' valuation of *recipients*. The covariation between inferred valuation and punishment is specific to inferences of per-

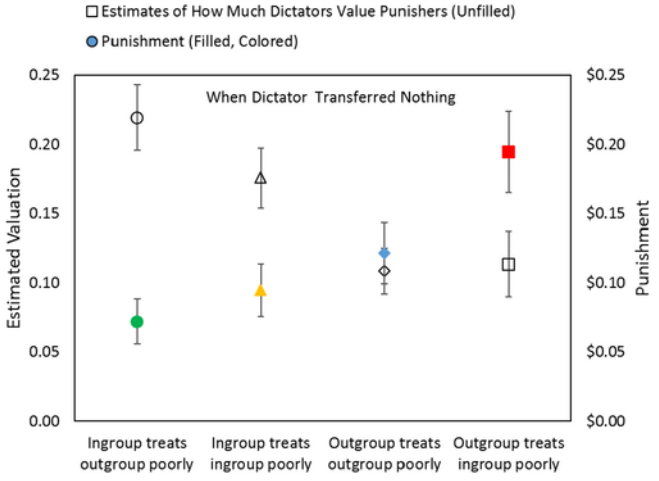

Fig. 5. Left vertical axis graphs punishers' mean estimates about how much dictators value them (the punishers), as a function of group composition (unfilled shapes). Right vertical axis graphs punishers' mean punishment of dictators, as a function of group composition (filled, colored shapes). All data come from cases where dictators transferred nothing. Error bars ± 1 SEM. $N = 275$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

sonal treatment—as predicted by deterrence theory. We emphasize that this ordering of punishment by conditions and the connections between inferred valuation of the self are clear, a priori predictions made only by deterrence theory; the norms cannot cogently explain the ordering nor do they directly predict the connections to inferred valuation.

One condition, however, did not exactly match this pattern. Punishers expected very low valuation from an outgroup dictator who treats another outgroup member poorly (see Fig. 5). Based on inferences of personal valuation alone, deterrence theory would predict that punishers should therefore punish strongly. In fact, they only punished a moderate amount (see Figs. 4 and 5). One possibility is that this low value resulted from noise. Fig. 5 presents eight sample means and it would be surprising if all these statistics exactly matched their true population parameters. If this low value represents reality, however, we speculate this mismatch is due to lower expectations of valuation from outgroup members in general; if punishers do not expect outgroup dictators to value them, then they may not need to update this expectation very much when they find out the dictator treats her own ingroup member poorly. The data support this post-hoc speculation: Even before learning of the dictator's decision punishers in the outgroup dictator/outgroup recipient condition expect the lowest personal valuation by the dictator (outgroup/outgroup mean valuation compared to average of other three conditions: M = 0.17, SD = 0.19 vs M = 0.27, SD = 0.21; $t(273) = 3.94$, $p < 0.001$, $r^2 = 0.05$), and update this expectation the least when considering the dictator gave 0% to the recipient (updating scores = drop in valuation scores from no info to known that dictator gave 0%; outgroup/outgroup compared to other three conditions: M = − 0.06, SD = 0.16 vs M = − 0.098, SD = 0.19; $t(273) = 1.71$, $p = 0.089$, $r^2 = 0.01$).

Regardless, as can be clearly seen from Figs. 3 and 5, punishers' inferences about how dictators value them closely track punishment of poor treatment of a third party—as predicted by deterrence theory.

We do not think the fairness norm makes direct predictions about the valuation task. Nonetheless, we tested contrasts comparing the pattern predicted by deterrence theory versus the pattern (assuming it follows that of punishment) for the fairness norm. We used the data from the condition where dictators shared nothing as the least ambiguous case for drawing predictions from these theories. First, a contrast for the pattern predicted by deterrence theory accounted for significant variance in valuation ($t(271) = -3.29$, $p = 0.001$, $r^2 = 0.04$; see SI for contrast codes). Second, a contrast for the pattern predicted by the fairness norm accounted for significant variance but in the wrong direction ($t(271) = 2.32$, $p = 0.021$, $r^2 = 0.02$). Finally, the deterrence contrast predicted more variance than the fairness norm contrast ($t(271) = -4.00$, $p < 0.001$, $r^2 = 0.06$).

3.6. Do punishers' inferences about dictators' valuation statistically mediate the effect of group composition on punishment?

This study cannot test a direct causal connection between valuation inferences and punishment; while group composition was manipulated, its effects on valuation inferences and punishment were only measured. But we think there are strong reasons to believe the connection is not spurious. Past research and theory has consistently connected inferences of valuation to anger, revenge, and punishment (Kurzban & DeScioli, 2013; McCullough et al., 2013; Sell et al., 2009; Tooby et al., 2008). For instance, people who feel entitled to higher valuation from others are more likely to be angered (Sell et al., 2009). A recent study showed experimentally that manipulating the information available to punishers changed their inferences about valuation in subtle and complex ways, which were mirrored in punishment decisions (Krasnow et al., 2016). Because valuation inferences are an internal variable, in an important sense they can never be "directly" manipulated. Instead, to understand their role, researchers must make reasonable inferences based on the totality of the evidence. We believe that this study, in conjunction with past research, helps to do so.

This issue of causality is a problem for deterrence theory, to be sure. But it is always a problem for *any* psychological theory, including norm theory. We can never directly manipulate internal psychological variables, including any potential internalized norms; all we can ever manipulate are external stimuli. So it is always an abductive argument to say, for example, that internalized norms cause a particular behavior. Indeed, whereas we measure some of deterrence theory's putative psychological causes, none of the high-profile group norm maintenance papers even attempt this. They simple measure behavior and by assumption treat that behavior as necessarily reflecting the norm.

For readers interested in statistical mediation, we report and discuss models in the supplemental information. This analysis shows that punishers' inferences about how much they are valued by dictators do statistically mediate the relationship between group composition and punishment, as predicted by deterrence theory (across the omnibus and contrast tests, two ps = 0.06 and two ps < 0.05). Although the mediation tests are only on the edge of significance or slightly above conventional values for two-tailed tests, we note that the size of the mediation effects were fairly large, accounting for about 20% percent of the total effect of group membership on punishment. This effect was unique to inferences about *personal* valuation. Given that inferences of *recipient* valuation did not differ by group composition (see Section 3.4), it is not surprising that they did not mediate the effect of group composition on punishment.

4. Discussion

Laboratory third-party punishment, and the real-world processes it is meant to model, often depend on group membership. In unframed laboratory games, subjects punish the mistreatment of others differently when the offender (or victim) is ingroup or outgroup. In the real world, our moral sentiments are not identically engaged when a community member is victimized (e.g., outrage following the San Bernardino Massacre) as when these victims are foreign citizens a world away (e.g., apathy to the ongoing Syrian civil war). It is easy to see these patterns of data and read into them inherently group-based processes like group norm maintenance. The results of the present study suggest this characterization, though appealing, is not correct. The mind instead appears designed to use the context of group membership as information regarding how generalizable the treatment of another is for the self. That is, what may look on the surface like the selfless defense of ingroup interests, may instead be the message "don't tread on me and mine".

We hasten to add that the deterrence theory is *not* a theory of self-interested egoism. For a given person, others can be valuable and their interests worth defending. This is uncontroversial when it comes to genetic kin. For instance, a mother putting herself in danger to defend her children has clear inclusive fitness interests. Whether the mother thinks or believes that protecting her children is in some sense to her own benefit is beside the point; designs that pay a small cost in mothers but reap a large enough reward in offspring will out-reproduce designs that do not. Similarly, friends and reciprocity partners can be intrinsically valuable if they cannot easily be replaced (Tooby & Cosmides, 1996). If a friend is threatened, coming to their defense

can, over the long-term, lead to benefits by sustaining the relationship, despite the short-term costs. People need not consciously consider this logic, so long as their mind embodies it.

Although it is less often appreciated, similar arguments apply to ingroup members. Ingroup members, especially those with whom one has a long history of coordination and cooperation, are also intrinsically valuable (Tooby et al., 2006), making their interests worth defending. Combining this with the fact that our evolved psychology is necessarily adapted for long-running features of ancestral environments—not anomalous situations like one-shot anonymous games or recently emerged market economies—explains why people would defend group interests even when there appear to be no rationally warranted benefits to doing so in a given situation.

4.1. Further comparing group norm maintenance and deterrence psychology

Besides the mere fact that third-parties will punish in one-shot, anonymous games, another source of evidence that group norm maintenance proponents offer is that societies with more third-party punishment are also fairer when making economic decisions (Henrich et al., 2006); they argue that this shows that groups with stronger punishment norms maintain better fairness norms. Similarly, the more market integrated a small-scale society is, the fairer its members are; proponents argue that greater involvement in anonymous markets requires fairness norms (Henrich et al., 2010).

However, these data points both follow from a deterrence psychology as well: First, if the local environment has more punishment of poor behavior, then all else equal, people should be more likely to behave well just by following the incentives; it's not clear why group norms are necessary to explain such a straightforward finding. Second, greater market integration predicts greater benefits to successful exchange, trade, and cooperation. Modeling work shows that, even in the absence of group norms, greater benefits to cooperation selects for psychologies willing to engage in anonymous cooperation (Delton et al., 2011). This creates a virtuous cycle (Ridley, 2010) whereby greater benefits create more willingness to cooperate and trust others, which in turn further enlarges the benefits (Delton, Krasnow, Cosmides, & Tooby, 2010).

The deterrence hypothesis helps to explain a data point collected by group norm maintenance proponents that they acknowledge is difficult for a norm of fairness to account for: In a group-based third-party punishment game, *outgroup* members are punished more harshly for treating an ingroup member poorly than *ingroup* members are (Bernhard et al., 2006; Goette et al., 2006; Schiller et al., 2014). This is odd on a group norm maintenance account because punishment is about preserving ingroup norms; outgroup members are comparatively irrelevant. But it is directly predicted by the cue-based deterrence view: If an outgroup member treats your ingroup member poorly, they likely did so because of group membership, strongly suggesting they would likely treat you or other group members poorly as well.

A group norm maintenance theorist might point out that their theory *is* a theory of deterrence, broadly construed. For instance, it holds that people punish against their personal interests to deter norm breakers from breaking the norm again (and perhaps to deter observers from breaking the norm). We do not think our analysis in this paper is affected by acknowledging this. We used the terms "group norm maintenance" and "deterrence theory" for rhetorical convenience and because they have been used in the past. Regardless of the names, what matters is the specific predictions each theory makes and how those predictions fare when tested empirically. The direct predictions of group norm maintenance theory that we tested in this study, the contexts under which people would be predicted to deter on behalf of their group, were not confirmed in our data.

4.2. Alternative hypotheses from group norm maintenance

Proponents of groups norm maintenance theory might respond that group norms are responsible for the pattern of punishment and the valuation inferences are merely epiphenomenal to punishment. On this alternative explanation, deterrence theory is not responsible for punishment, group norm maintenance is; inferences about valuation are produced by a different process and do not affect punishment despite tracking it. We acknowledge that this is possible—the statistical mediation certainly does not prove true causal mediation. But this alternative has serious difficulties as well. As noted above, the patterns of punishment across conditions are inconsistent with the two primary norms that group norm maintenance theorists have discussed, even if the norms are considered in combination. This would be a problem for group norm maintenance even if we had not measured inferences. Second, we remind readers that proponents of these theories have been quite clear that the punishment (and cooperation, fairness, and so on) produced by this mechanism is different because people do not expect any personal benefits from such behavior (e.g. quote from Fehr & Henrich, 2003, in Section 1.1). That players make inferences about personal treatment in these games appears, to us at least, seriously inconsistent with this claim.

Another possibility is that dictators are assessing whether the dictator poses a *general* threat to their group. If so, punishers would infer poor treatment of themselves—they are, after all, a member of their group. But they would also infer poor treatment of other members of their group who were not part of the game at all (which we did not measure). On this view, the self is just one substitutable member of their ingroup among many. This is logically possible and we cannot rule it out definitively. Nonetheless, it suffers from the same problem noted above that to preserve group norm maintenance theory, punishers are being allowed to assume that personal benefits will come from punishment. Also, other data (albeit not collected in a group context) have shown that the self is special (Krasnow et al., 2016). In this previous experiment, we presented punishers with information about how dictators shared with both the recipient and, in a separate decision, the punisher herself. Punishment was entirely driven by how the punisher was treated—it did not matter how the recipient was treated. But if this particular gloss on group norm maintenance was correct, treatment of the punisher and the recipient should have both played a role, perhaps equal roles, in determining punishment. At the least, treatment of the recipient should have had some relationship to punishment. But it did not.

A third possibility is that we are observing third-party punishment as predicted by group norm maintenance but that a general bias against outgroup members is acting as well. In other words, people *are* punishing group norm violators, but they also have a bias against outgroup members which increases the observed punishment in some cases. One reading of this alternative is that it is simply a combination of the fairness norm and the parochialism norm, which we ruled out above. But even if not, it still has problems. First, to account for our data, this alternative must predict that norm violation effects are swamped by outgroup bias—recall that the most punishment is directed at outgroup members treating ingroup members poorly, rather than the most serious norm violators, ingroup members treating ingroup members poorly. This alternative rescues the fairness norm only by severely weakening it. Second, if there was a generalized outgroup bias, it acts in a convoluted way. Rather than harming out-

group members whenever possible, or harming outgroup members whenever they treat others poorly, punishers harmed outgroup dictators more only when they did not share with the punisher's ingroup members. This pattern seems hard to capture in a simple notion of outgroup bias.

A fourth possibility is that, aside from fairness or parochialism norms, there could also be a norm for deterring poor treatment of ingroup members by outgroup members. This is possible in principle, but aside from one speculation invoking it to explain away findings difficult for group norm maintenance theory (Bernhard et al., 2006, p. 914), we know of no work addressing this issue theoretically or empirically. Until researchers articulate the structure of this norm in more detail, it is difficult to test how it operates. Invoking this norm also gives group norm maintenance theory yet another degree of freedom: With fairness norms leading to punishment of unfairness, parochialism norms leading to poor treatment of outgroups, and group-based deterrence norms leading to punishment of outgroup unfairness towards ingroups, in combination and with variable weights, these three norms could be consistent with almost any pattern of data. This makes the norms theory nearly unfalsifiable.

Moreover, we think that appealing to a more general deterrence psychology is more straightforward. As we described in the introduction, many animals have a deterrence psychology that defends their personal interests. No one is surprised when an animal uses threats to ward off others trying to aggress against it. Importantly, personal interests do not stop at the animal's own body. Again, no one is surprised if a mother deters poor treatment of her offspring or if genetic siblings defend each other. It should be no more surprising that in a species with friends, alliances, and coalitions, these relationships get incorporated into the category of personal interests. Once this step is made, group-based third-party punishment is a straightforward outcome of deterrence. Thus, it is unclear what additional aspects of group-based third-party punishment need explanation or what group norm maintenance theory uniquely adds to such an explanation.

4.3. Beyond deterrence

These data join past work in describing a psychology of deterrence that acts to defend one's interests and the interests of those one values. Yet, third-party punishment and third-party intervention more broadly are likely to be multiply determined by other psychologies as well. There could be motivations besides deterrence to punish as second parties, which generalize to the third-party case as we have shown for deterrence (Krasnow et al., 2016). By intervening on behalf of others, I can develop a reputation as a trustworthy cooperative partner, leveraging my way into more or better relationships (Jordan et al., 2016). I might be particularly vulnerable to the type of violation that was made, and therefore staked in preventing *this* behavior from becoming commonplace. By harming another, an actor could be expressing a bid for higher status, a move that I may be keen to prevent rather than see my own status shrink. The data we report show four basic effects that we attribute to deterrence theory: that punishers infer personal valuation by the dictator merely on the basis of behavior towards the recipient; that this inference is moderated by the relative group membership of the participants in ways consistent with how substitutable the punisher is for the recipient in the dictator's eyes; that the inference of low personal valuation by the dictator predicts punishment; and that the previously known group difference in punishment is partially mediated by the group difference in valuation inference. These effects do not account for all of the variance in third party punishment we observed, suggesting that while deterrence is part of the story, it is only part of the story.

5. Conclusion

Understanding deterrence psychology is only possible by taking the psychology of small-scale social life seriously. The last two decades have seen waves of research purporting to rule out the possibility that this kind of psychology—a psychology concerned with costs and benefits reliably present in the small-scale social world of the human past—could explain third-party punishment. The experimental protocols of one-shot, anonymous interactions were meant to rule out the possibility that this kind of psychology would influence behavior. Yet, a growing literature suggests that otherwise puzzling phenomena—participants cooperating more than they "should," punishing more than they "should," and for reasons that "should" not be relevant—can simply and parsimoniously be explained as the result of our psychology of small-scale social interaction doing what it should: foraging for potentially valuable social relationships, and defending the self and valuable others against potential exploitation and mistreatment.

Uncited references

Cheney, 1981
Krasnow and Delton, 2016a
Krasnow and Delton, 2016b
Wilson and Wrangham, 2003

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.evolhumbehav.2017.07.003.

References

Averill, J.R., 1983. Studies on anger and aggression: Implications for theories of emotion. American Psychologist 38, 1145–1160.
Bernhard, H., Fischbacher, U., Fehr, E., 2006. Parochial altruism in humans. Nature 442, 912–915.
Bone, J., Silva, A.S., Raihani, N.J., 2014. Defectors, not norm violators, are punished by third-parties. Biology Letters 10, 20140388.
Boyd, R., Gintis, H., Bowles, S., Richerson, P.J., 2003. The evolution of altruistic punishment. Proceedings of the National Academy of Sciences of the United States of America 100, 3531–3535.
Boyd, R., Richerson, P.J., 1992. Punishment allows the evolution of cooperation (or anything else) in sizable groups. Ethology and Sociobiology 13, 171–195.
, -
Choi, J.-K., Bowles, S., 2007. The coevolution of parochial altruism and war. Science 318, 636–640.
Chudek, M., Henrich, J., 2011. Culture–gene coevolution, norm-psychology and the emergence of human prosociality. Trends in Cognitive Sciences 15, 218–226.
Clutton-Brock, T.H., Parker, G.A., 1995. Punishment in animal societies. Nature 373, 209–216.
Delton, A.W., Krasnow, M.M., Cosmides, L., Tooby, J., 2010. Evolution of fairness: Rereading the data. Science 329, 389.
Delton, A.W., Krasnow, M.M., Cosmides, L., Tooby, J., 2011. The evolution of direct reciprocity under uncertainty can explain human generosity in one-shot encounters. Proceedings of the National Academy of Sciences of the United States of America 108, 13335–13340.
Delton, A.W., Robertson, T.E., 2016. How the mind makes welfare tradeoffs: Evolution, computation, and emotion. Current Opinion in Psychology 7, 12–16.
Fehr, E., Fischbacher, U., 2004. Third-party punishment and social norms. Evolution and Human Behavior 25, 63–87.
Fehr, E., Fischbacher, U., Gächter, S., 2002. Strong reciprocity, human cooperation, and the enforcement of social norms. Human Nature 13, 1–25.
Fehr, E., Henrich, J., 2003. Is strong reciprocity a maladaptation? On the evolutionary foundations of human altruism. In: Hammerstein, P. (Ed.), Genetic and cultural evolution of cooperation. MIT Press, Cambridge, MA, US, pp. 55–82.
Gintis, H., 2000. Strong reciprocity and human sociality. Journal of Theoretical Biology 206, 169–179.

Goette, L., Huffman, D., Meier, S., 2006. The impact of group membership on cooperation and norm enforcement: Evidence using random assignment to real social groups. The American Economic Review 96, 212–216.
Hagen, E.H., Hammerstein, P., 2006. Game theory and human evolution: A critique of some recent interpretations of experimental games. Theoretical Population Biology 69, 339–348.
Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., … Ziker, J., 2010. Markets, religion, community size, and the evolution of fairness and punishment. Science 327, 1480–1484.
Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, H.C., Bolyanatz, A., … Ziker, J., 2006. Costly punishment across human societies. Science 312, 1767–1770.
Horton, J.J., Rand, D.G., Zeckhauser, R.J., 2011. The online laboratory: Conducting experiments in a real labor market. Experimental Economics 14, 399–425.
Jordan, J.J., Hoffman, M., Bloom, P., Rand, D.G., 2016. Third-party punishment as a costly signal of trustworthiness. Nature 530, 473–476.
Jordan, J.J., McAuliffe, K., Rand, D., 2015. The effects of endowment size and strategy method on third party punishment. Experimental Economics 1–23.
Jordan, J.J., McAuliffe, K., Warneken, F., 2014. Development of in-group favoritism in children's third-party punishment of selfishness. Proceedings of the National Academy of Sciences 111, 12710–12715.
Kirkpatrick, M., Delton, A.W., Robertson, T.E., de Wit, H., 2015. Prosocial effects of MDMA: A measure of generosity. Journal of Psychopharmacology 29, 661–668.
Krasnow, M.M., Cosmides, L., Pedersen, E.J., Tooby, J., 2012. What are punishment and reputation for?. PloS One 7, e45662.
, ,
Krasnow, M.M., Delton, A.W., Cosmides, L., Tooby, J., 2016. Looking under the hood of third-party punishment reveals design for personal benefit. Psychological Science 27, 405–418.
Krasnow, M.M., Delton, A.W., Tooby, J., Cosmides, L., 2013. Meeting now suggests we will meet again: Implications for debates on the evolution of cooperation. Nature Scientific Reports 3, 1747.
Kurzban, R., DeScioli, P., 2013. Adaptationist punishment in humans. Journal of Bioeconomics 15, 269–279.
Kurzban, R., Descioli, P., O'Brien, E., 2007. Audience effects on moralistic punishment. Evolution and Human Behavior 28, 75–84.
Lieberman, D., Linke, L., 2007. The effect of social category on third party punishment. Evolutionary Psychology 5, 289–305.
Masclet, D., Noussair, C., Tucker, S., Villeval, M.C., 2003. Monetary and nonmonetary punishment in the voluntary contributions mechanism. American Economic Review 93, 366–380.
McAuliffe, K., Dunham, Y., 2016. Group bias in cooperative norm enforcement. Philosophical Transactions of the Royal Society B 371, 20150073.
McAuliffe, K., Jordan, J.J., Warneken, F., 2015. Costly third-party punishment in young children. Cognition 134, 1–10.
McCullough, M.E., Kurzban, R., Tabak, B.A., 2013. Cognitive systems for revenge and forgiveness. Behavioral and Brain Sciences 36, 1–58.
Ostrom, E., 1998. A behavioral approach to the rational choice theory of collective action. American Political Science Review 92, 1–22.
Pedersen, E.J., Kurzban, R., McCullough, M.E., 2013. Do humans really punish altruistically? A closer look. Proceedings of the Royal Society of London B: Biological Sciences 280, 20122723.
Raihani, N.J., Thornton, A., Bshary, R., 2012. Punishment and cooperation in nature. Trends in Ecology & Evolution 27, 288–295.
Richerson, P.J., Baldini, R., Bell, A., Demps, K., Frost, K., Hillis, V., … Zefferman, M., 2016. Cultural group selection plays an essential role in explaining human cooperation: A sketch of the evidence. Behavioral and Brain Sciences 39, 1–19.
Richerson, P.J., Boyd, R., 2005. Not by genes alone. University of Chicago Press, Chicago.
Ridley, M., 2010. The rational optimist. HarperCollins, New York City.
Roos, P., Gelfand, M., Nau, D., Carr, R., 2014. High strength-of-ties and low mobility enable the evolution of third-party punishment. Proceedings of the Royal Society B: Biological Sciences 281, 20132661.
Rosenthal, R., Rosnow, R.L., Rubin, D.B., 2000. Contrasts and effect sizes in behavioral research: A correlational approach. Cambridge University Press, Cambridge, U.K..
Rusch, H., 2014. The evolutionary interplay of intergroup conflict and altruism in humans: A review of parochial altruism theory and prospects for its extension. Proceedings of the Royal Society of London B: Biological Sciences 281, 20141539.
Schiller, B., Baumgartner, T., Knoch, D., 2014. Intergroup bias in third-party punishment stems from both ingroup favoritism and outgroup discrimination. Evolution and Human Behavior 35, 169–175.
Sell, A., 2011. Applying adaptationism to human anger: The recalibrational theory. In: Shaver, P.R., Mikulincer, M. (Eds.), Human aggression and violence: Causes, manifestations, and consequences. American Psychological Association, Washington, DC, US, pp. 53–70.
Sell, A., Tooby, J., Cosmides, L., 2009. Formidability and the logic of human anger. Proceedings of the National Academy of Sciences of the United States of America 106, 15073–15078.
Tooby, J., Cosmides, L., 1996. Friendship and the banker's paradox: Other pathways to the evolution of adaptations for altruism. In: Runciman, W.G., Maynard Smith, J., Dunbar, R.I.M. (Eds.), Evolution of social behaviour: Patterns in primates and man. 88, pp. 119–143.
Tooby, J., Cosmides, L., Price, M.E., 2006. Cognitive adaptations for n-person exchange: The evolutionary roots of organizational behavior. Managerial and Decision Economics 27, 103–129.
Tooby, J., Cosmides, L., Sell, A., Lieberman, D., Sznycer, D., 2008. Internal regulatory variables and the design of human motivation: A computational and evolutionary approach. In: Elliot, A.J. (Ed.), Handbook of approach and avoidance motivation. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 251–271.
West, S.A., Griffin, A.S., Gardner, A., 2007. Social semantics: Altruism, cooperation, mutualism, strong reciprocity and group selection. Journal of Evolutionary Biology 20, 415–432.