Evolution and Human Behavior 38 (2017) 734–743

Contents lists available at ScienceDirect

Evolution and Human Behavior

journal homepage: www.ehbonline.org

Original Article

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:
Initial receipt 8 December 2016
Final revision received 24 July 2017

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.

☆ 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.
E-mail 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/

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 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).