

Learning moral values: another's desire to punish enhances one's own punitive behavior

Oriel FeldmanHall^a, A Ross Otto^b, Elizabeth A. Phelps^{c,d,e}

^aDepartment of Cognitive, Linguistics & Psychological Sciences, Brown University, RI. 02912

^bDepartment of Psychology, McGill University, Montreal QC H31 1G1

^cDepartment of Psychology, New York University, New York, NY. 10003

^dCenter for Neural Science, New York University, New York, NY. 10003

^eNathan Kline Institute, Orangeburg, NY. 10962

Correspondence to:

Oriel FeldmanHall

Brown University

190 Thayer St

Providence, RI 02906

Email: oriel.feldmanhall@brown.edu

FULL ABSTRACT

There is little consensus about how moral values are learned. Using a novel social learning task, we examine whether vicarious learning impacts moral values—specifically fairness preferences—during decisions to restore justice. In both laboratory and internet-based experimental settings, we employ a dyadic Justice Game where participants receive unfair splits of money from another Player and respond resoundingly to the fairness violations by exhibiting robust non-punitive, compensatory behavior (baseline behavior). In a subsequent learning phase, participants are tasked with responding to fairness violations on behalf of another participant (a Receiver) and are given explicit trial-by-trial feedback about the Receiver's fairness preferences (e.g., whether they prefer punishment). This allows participants to update their decisions in accordance with the Receiver's feedback (learning behavior). In a final test phase, participants again directly experience fairness violations. After learning about a Receiver who prefers highly punitive measures, participants significantly enhance their own endorsement of punishment during the test phase compared to baseline. Computational learning models illustrate the acquisition of these moral values is governed by a reinforcement mechanism, revealing it takes as little as being exposed to the preferences of a single individual to shift one's own fairness preferences when responding to moral violations. Together this suggests that even in the absence of explicit social pressure, fairness preferences are highly labile.

Key words: Moral preferences; learning; punishment; conformity, social norms

INTRODUCTION

Decades of research have established that moral behavior is dictated by a set of customs and values that are endorsed at the society level in order to guide conduct (Moll et al. 2005, Bowles and Gintis 2011, Chudek and Henrich 2011, Kouchaki 2011). These culturally normative rules are critically shaped by fairness considerations (Spitzer et al. 2007, Haidt 2010, Rai and Fiske 2011, Peysakhovich and Rand 2016), which is considered a core foundation of morality (Turiel 1983, de Waal 2006). Indeed, there is much evidence that humans are highly motivated to re-balance the scales of justice after experiencing or observing norm violations (Fehr and Gächter 2002, Fehr and Fischbacher 2004, Fehr et al. 2008, Mathew and Boyd 2011). The degree to which fairness is valued even extends to non-human primates, where animals routinely seek out ways to re-establish equality amongst group members when posed with a fairness infraction (de Waal 1996, Brosnan and De Waal 2003, de Waal 2006). Although sensitivity to fairness is well established across species, there is little consensus about how these fairness preferences are learned, and whether they are unwavering in their application or vulnerable to social influence.

On one hand, individual fairness preferences are often characterized as stable and fixed (Peysakhovich et al. 2014). This dovetails with a long history of research illustrating that moral attitudes (Turiel 1983, Haidt 2001)—such as beliefs about justice and honor—are resolutely and sacredly held (Sturgeon 1985, Skitka et al. 2005, Bauman and Skitka 2009), resistant to change (Hornsey et al. 2003, Tetlock 2003, Andrew Luttrell 2016), and firmly rooted even in the face of social opposition (Aramovich et al. 2012). And yet research in other domains illustrates how social pressure and influence can swiftly yield conformist behavior (Deutsch 1955, Cialdini 1998, Bardsley and Sausgruber 2005, Zaki et al. 2011), indicating that individuals desire the approval of others (Baumeister and Leary 1995, Wood 2000). There is an equally long history of evidence delineating how social attitudes are exquisitely attuned to perceived societal milieus (Jones 1994, Nook et al. 2016). For instance, perceptions of group consensus can induce compliant behavior, such as altering how readily one estimates or values an object (Asch 1956, Campbell-Meiklejohn et al. 2010), or exhibits cooperative behavior in a group setting (Fowler and Christakis 2010). Early work revealed the force by which social influence could modify decision-making (Cialdini and Goldstein 2004, Izuma and Adolphs 2013), demonstrating just how far individuals are willing to go when confronted with overt and forceful instruction (Milgram 1963, Haney 1973).

Much less is known, however, about the extent to which moral preferences are malleable in the absence of overt social pressure or perception of social consensus. In other words, does learning about another's moral values reveal the existence of a new social norm that bears on how one comes to value certain moral preferences? To examine whether such decisions to restore justice—referred to here as fairness preferences—are indeed susceptible to subtle social influences, we compare how participants respond to fairness violations before learning about another's fairness preferences, to those made afterward, in a series of Internet and laboratory-based experiments. To test and measure these putative behavioral changes, we adapted a dyadic economic task, the Justice Game (FeldmanHall et al. 2014), which measures different motivations for restoring justice.

In the modified Justice Game, Player A is endowed with a sum of money [1] and can make fair or unfair offers by distributing the money as she sees fit between herself [1-x] and Player B [x] (Falk et al. 2008). Player B then has the opportunity to restore justice by redistributing the money between the players. Participants partake in three phases of the game, a Baseline (1st), Learning (2nd), and Transfer (3rd) phase (Fig 1), in which they play both as a second-party (Player B), directly experiencing fairness violations (Baseline and Transfer phases)—and as a third-party (Player C), making decisions on behalf of another victim, where they can also observe the victim's preferences after each choice (Learning phase). This allows us to explore whether observing and then implementing another's preferences during the Learning phase influences subsequent responses made for the self during the Transfer phase. If responses in the Transfer phase significantly differ from responses in the Baseline phase (a within-subjects design), it would suggest that preferences for handling fairness violations are indeed malleable, and the degree of behavioral change between these phases would enable us to measure the extent to which moral preferences transmit from one person to another.

Examining behavior in the Transfer compared to Baseline conditions, further permits us to test two key questions about the malleability of these preferences. First, while a rich economic literature suggests that punishing the perpetrator is the preferred method for restoring justice (Fehr and Gächter 2002, Fehr and Fischbacher 2004, Henrich et al. 2005, Gintis and Fehr 2012), recent work reveals that these punishment preferences are contextually bound (Dreber et al.

2008, Jordan et al. 2016), and that when other options are made available, individuals exhibit a resounding endorsement for less punitive and more restorative forms of justice (FeldmanHall et al. 2014). Accordingly, we were primarily interested in understanding whether punitive preferences can be acquired from learning about others who desire punishment. A robust expression of certain moral values—in this case, that punishment is a preferred form of justice restoration—could signal the existence of an alternative appropriate social norm, and being exposed to this social norm may alter how much people value punishment themselves. Therefore, assuming individuals initially exhibit non-punitive, compensatory preferences (FeldmanHall et al. 2014), will exposure to an individual who exhibits strong preferences to punish shift how readily one endorses punishment as a means of restoring justice for oneself?

Second, we can address the way in which this putative behavioral transfer occurs. That is, how might an alternative social norm become assimilated into an individual's moral calculus? For example, if the difference between how one individual values punishment compared to how another values punishment (i.e., a prediction error between what one prefers and what another prefers) biases one's own taste for punishment, it would suggest that a reinforcement mechanism governs the acquisition of moral preferences (Sutton 1998). Moreover, we can also test what type of learning subserves this process. It is possible that simply observing another's fairness preferences is enough to influence one's own preferences (i.e. a contagion effect; Chartrand and Bargh 1999, Suzuki et al. 2016). Alternatively, transmission of moral preferences may demand successful implementation of the victim's preferences. Here we leverage a task paradigm traditionally used to study reward learning (i.e., Reinforcement Learning) to elucidate the cognitive mechanisms underlying how moral preferences come to be valued.

Experiment 1

Participants. In Experiment 1, 300 participants (147 females, mean age=34.37 SD±10.07) were recruited from the United States using the online labor market Amazon Mechanical Turk (AMT) (Paolacci et al. 2010, Buhrmester et al. 2011, Horton et al. 2011, Mason and Suri 2012) (one participant from the Reverse Condition, two participants from the Indifferent Condition and five participants from the Compensate Condition failed to complete the entire task and therefore their responses were not logged). Participants were paid an initial \$2.50 and an additional monetary bonus accrued during the task from one randomly realized trial from either the

Baseline or Transfer phases (ranging from \$.10 to \$.90). Informed consent was obtained from each participant in a manner approved by New York University's Committee on Activities Involving Human Subjects. All participants had to complete a comprehension quiz before beginning the task (Crump et al. 2013). We also ran initial pilot experiment on AMT to assess the necessary levels of reinforcement (150 participant, 71 females, mean age=33.68 SD±10.01). Results fully replicate but are not reported here for succinctness.

Justice Game. The Justice Game (FeldmanHall et al. 2014) was adapted to comprise three phases. In phase one—the Baseline phase (Fig 1A)—Player A has the first move and can propose any division of a \$1 or \$10 pie (initial endowment for Player A depended on the experiment) with Player B, the participant (Player A: $1 - x$, Player B: x ; Fig 1A illustrates only unfair offers ranging from [\$.60, \$.40] to [\$.90, \$.10]). Player B can then reapportion the money by choosing from the following three options; 1) *Accept*: agreeing to the proposed split ($1 - x, x$; Guth et al. 1982); 2) *Compensate*: increasing Player B's own payout to match Player A's payout, thus enlarging the pie and maximizing both players' outcomes ($1 - x, 1 - x$; Lotz et al. 2011); or, 3) *Reverse*: reversing the proposed split so that Player A is maximally punished and Player B is maximally compensated ($x, 1 - x$; Straub and Murnighan 1995, Pillutla and Murnighan 1996), a highly retributive option that punishes the perpetrator proportionate to the wrong committed (Carlsmith et al. 2002). In other words, reversing the Players' payouts results in Player A receiving what was initially assigned for Player B, and vice versa—a direct implementation of the 'just deserts' principle. Importantly, the Compensate option is entirely non-punitive and highly prosocial, since Player A is not punished with a reduced payout for behaving unfairly. In contrast, the Reverse option provides the opportunity to punish for free, since Player B's payout remains the same in either the Compensate or Reverse options (see SI for a more in depth explanation of each option).

Past research illustrates that choosing to non-punitively Compensate is the most frequently endorsed option when responding to fairness violations (FeldmanHall et al. 2014). To ensure that Player B perceives an offer of [\$.90, \$.10] as unfair, participants were told that one trial would be randomly selected to be paid out, and that half of the time it would be paid out according to Player A's split (like a Dictator game), and half the time according to Player B's

decision. This incentive structure means that an unfair offer signals Player A's desire to maximize and enrich their payout at the expense of Player B's payout (e.g. a selfish decision).

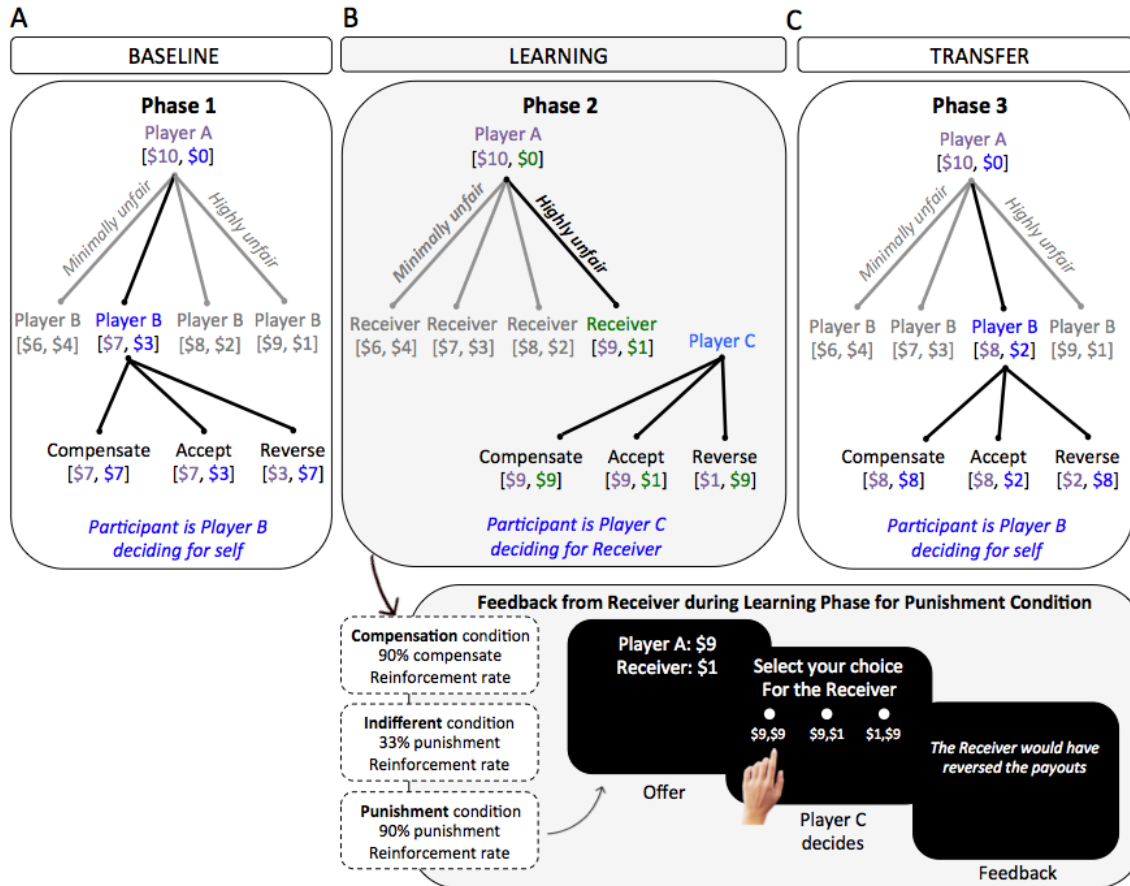


Fig 1 | Experimental Set-up of Justice Task. **A) Baseline phase.** Player A is the first mover and endowed with a sum of money, in this case \$10 (monetary amounts in the brackets denote the payouts for each player at every stage in the decision tree: A's payout is always on the left, B's payout is always on the right). After receiving an offer from Player A, in this case [\$7, \$3], participants—who take the role of Player B—are presented with three options to restore justice: they can 'Accept' the offer as is, 'Compensate' themselves and not punish player A by only increasing their own payout to match that of Player A, or punish Player A by reversing the payouts (simultaneous punishment of Player A and compensation of the self, known as the 'Reverse' option). Splits from Player A were made out of \$1 in the online experiments and \$10 in the laboratory experiment. **B) Learning phase.** In the next phase of the task, participants change roles and make decisions as a non-vested third party (Player C). On each trial, participants observe Player A making an offer, in this case [\$9, \$1] to another participant (the Receiver) and can respond on behalf of the Receiver. After making a decision about how to reappropriate the money, participants receive feedback about how the Receiver would have responded. In the example given, the participant, who is in the Punishment condition, choose the option to Compensate (denoted by \$9, \$9) and then received feedback that the Receiver would have preferred the Reverse option (the highly punitive response). Participants are randomly selected to be in one of three conditions, where they receive feedback from a highly punitive Receiver (Punishment condition), a non-punitive Receiver (Compensate condition), or a Receiver who exhibited no strong preferences (Indifferent condition). Participants observe many different

Player A making offers to the same Receiver. **C) Transfer phase.** In the third phase of the task, participants again take the role of *Player B*, directly receiving offers from various *Player A*s, in this case an [\$8, \$2] offer.

During this Baseline phase, participants were told that they were pre-selected to be *Player B* and completed four trials, one for each level of fairness transgression (i.e., offer type), ranging from a relatively fair [\$0.60, \$0.40] split to a highly unfair [\$0.90, \$0.10] split (NB: participants were told that any offer was possible, which included a highly fair offer of [\$0.50, \$0.50], although participants were never presented with this amount as we were only interested in fairness transgressions). The varying types of offers from *Player A* were randomly presented to each participant. The Baseline phase enabled us to gather behavior prior to any social manipulations. In phase two—the Learning phase—participants were told they would be playing the game as *Player C*, a non-vested third party (Fig 1B). Accordingly, participants were asked to make decisions on behalf of two other players, such that payoffs would be paid to *Players A* and another *Player*—denoted as the Receiver. As *Player C*, participants first observed *Player A* making an offer to the Receiver. *Player C* could then redistribute the money between *Player A* and the Receiver with the same three options described above (Accept, Compensate, Reverse). Once a decision was made, *Player C* was informed of what the Receiver would have preferred to do (e.g., feedback). This feedback was given on every trial so that *Player C* could learn the justice preferences of the Receiver they were deciding for.

In one condition participants received feedback from a highly retributive Receiver, who reported that she would have wanted the initial split from *Player A* to be reversed (Punishment Condition: across 80 trials there was 90% reinforcement rate of Reverse, 5% reinforcement rate of Accept, and 5% reinforcement rate of Compensate). Effectively, by providing this trial-by-trial feedback, the Receiver signals that she not only values increasing her own payout, but she also prefers that *Player A* is punished for making unfair splits. In a second condition, participants received feedback from highly compensatory (non-punitive) Receiver (Compensate Condition: across 80 trials there was 90% reinforcement rate of Compensate, 5% reinforcement rate of Accept, and 5% reinforcement rate of Reverse). In this case, participants learned that the Receiver did not value punishment, and instead preferred that only her own payout be increased to match that of *Player A*. In a third condition, participants received feedback from a Receiver who was indifferent in their preferences, preferring to Compensate themselves,

Reverse the outcomes, or Accept the offer at equal reinforcement rates (Indifferent Condition: 33% for each type of feedback). These reinforcement feedback rates in each of the three conditions were predetermined by the experimenter to signal robust preferences that were highly punitive, not punitive at all, or completely random with an aim to ensure that participants could learn another player's judicial preferences, even during varying levels of fairness violations. Participants were randomly assigned to be in one of the three conditions such that each participant received feedback from only one Receiver (between-subjects design). As in the Baseline phase, offers from Player A ranged from minimally unfair to highly unfair.

In the third phase of the Experiment—the Transfer phase—participants again played the same Justice Game but this time as the victim (Player B) again. The Transfer phase was the same as the Baseline phase except there were 20 trials, where various unfair offer types were randomly presented. Following the experiment, participants were asked to report their strategies given these offer types (see supplement for participants' reported strategies).

Learning. To examine candidate learning processes involved in implementing another's preferences, we fit a family of computational reinforcement learning models that differently account for how feedback is incorporated, the magnitude of the fairness violation, and whether one's own initial fairness preferences influence learning rates (see SI). These models—which allowed us to examine various different psychological notions for how learning might unfold—were then fit to participants' behavior and tested via model comparison.

Results

Baseline Behavior. Replicating earlier findings (FeldmanHall et al. 2014), behavior in the Baseline phase revealed that participants overwhelmingly prefer to Compensate themselves and not punish Player A, even if punishment is free and costs nothing to implement (i.e. the Reverse option). Because up until this point all conditions (e.g., Compensate, Indifferent, Punishment) had identical experiences, we reasoned that the rate at which Compensate (or Reverse) was endorsed should not vary across conditions. Indeed, decisions to Compensate were endorsed on average 70% in all conditions, and extremely punitive decisions, which Reverse the players' payouts, were only endorsed on average 21%—regardless of condition (Fig 2A). These robust baseline non-punitive preferences did not vary across conditions ($Reverse Rate_{i,t} =$

$\beta_1 \text{Condition} + \varepsilon$ Table 1; same non-significant results are found for $\text{Compensate Rate}_{i,t} = \beta_1 \text{Condition} + \varepsilon$). Moreover, paralleling past work (FeldmanHall et al. 2014), these highly compensatory, non-punitive decisions were observed even after very unfair offers are made (67% mean endorsement of Compensation for highly unfair offers across conditions).

TABLE 1 | Mixed Effects logistic regression predicting endorsement of Reverse option.

$\text{Reverse Rate}_{i,t} = \beta_1 \text{Condition} + \varepsilon$

DV	Coefficient (β)	Estimate (SE)	z-value	P value
Reverse	Intercept	-2.58 (.39)	-6.69	<0.001***
	Indifferent Condition	0.18 (.46)	0.41	0.68
	Punish Condition	-0.07 (.46)	-0.17	0.86

*Note: Where Reverse (1=Reverse, otherwise 0) is indexed by subject and trial and condition is an indicator variable, such that the compensate condition serves as the intercept. *** $p < .001$*

Transfer Effects. To test whether exposure to another's preferences to restore justice results in the transmission of fairness preferences, we examined whether responses in the Transfer phase changed from those made during the Baseline phase. We computed the difference in the proportion of times an option (Compensate, Reverse, Accept) was endorsed between the Baseline versus Transfer phases for each participant (not accounting for the magnitude of fairness violation). Given that the Baseline phase revealed highly compensatory and non-punitive behavior, we speculated that the greatest transmission of fairness preferences should be observed for participants in the Punishment condition, as this is where the Receiver's preferences will differ most dramatically from participants' default, non-punitive preferences (Fig 2A). In contrast, participants in the Indifferent condition should exhibit small transfer effects, as the Receiver's feedback did not systemically endorse one response type, and those in the Compensate condition should show no behavioral changes as there was no conflict between judicial preferences the participant made for herself and what she observed the Receiver desired.

As predicted, the Indifferent condition—where participants received feedback from a Receiver who did not show a strong singular preference (albeit slightly stronger punitive preferences than a typical participant)—did not manifest a significant transmission of fairness preferences. Participants endorsed the Reverse option 23% in the Baseline phase, and 27% in the Transfer phase (a 4% increase, Fig. 2B) which statistically revealed no evidence of significant transfer

effect (rmANOVA DV=Endorsement of Reverse, IV=Phase: $F(1,97)=1.64$, $p=0.20$). In the Compensate condition, where participants learned about a Receiver who strongly preferred the non-punitive, compensatory option, participants increased their endorsement of Compensate by 1% in the Transfer phase (rmANOVA DV=Endorsement of Compensate. IV=Phase $F(1,94)=0.01$, $p=0.90$.), while decisions to Reverse were endorsed at similar same rates in both Baseline and Transfer phases. In other words, exposure to a Receiver who endorses compensatory measures more often than oneself results in effectively no transmission of preference, which is likely due to the fact that there was little conflict or prediction error between what the Receiver and participant valued when deciding to restore justice.

In contrast, for those in the Punishment condition, exposure to the Receiver's fairness preferences exerted marked effects upon participants' own decisions to restore justice (Fig 2B). Participants significantly bolstered their endorsement of the punitive Reverse option for themselves after learning about a Receiver who preferred highly retributive measures (16% increase: rmANOVA DV=Endorsement of Reverse, IV=Phase, $F(1,98)=31.29$, $p<0.001$, $\eta^2=.24$).

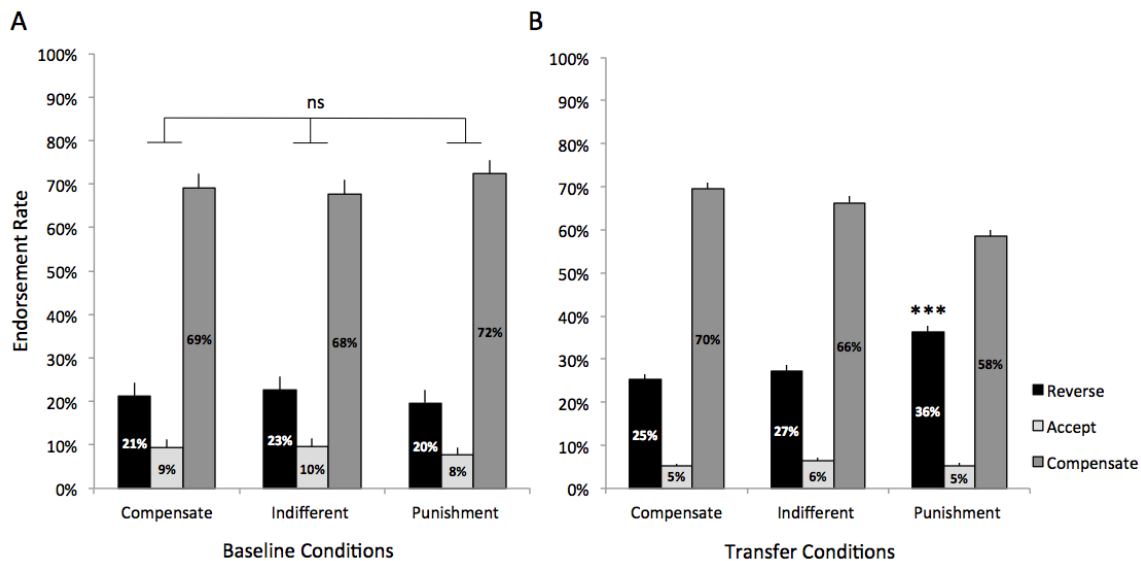


Fig 2 | Responses to fairness violations as the victim in Experiment 1. A) Behavior in the Baseline Phase for each condition reveals participants overwhelmingly endorsed the non-punitive option to compensate themselves and not punish Player A for making an unfair offer (70% endorsement of the Compensate option across three conditions). Participants equally endorsed decisions to Compensate, Reverse, and Accept at the same rate across conditions. **B) Behavior in the Transfer phase.** After engaging with a Receiver with highly punitive preferences (Punishment condition), participants significantly altered how they responded to fairness violations, increasing their endorsement of decisions to punish Player A (i.e. choosing Reverse) in the

Transfer phase. Significant enhancement of choosing to punish was only observed in the Punishment condition ($p < 0.001$).

To compare if these observed transfer effects in the Punishment condition are significantly different from the effects observed in the Compensate and Indifferent conditions, we computed Response Δ scores as the difference between endorsement rates for a particular response between the Transfer phase and the Baseline phase (averaging across all types of fairness violations). In the Punishment and Indifferent conditions, the Response Δ is computed with respect to the Reverse response, while for participants in the Compensate condition, the Response Δ is computed for the Compensate response (note in the Indifferent condition we compute Response Δ for Reverse given that the reinforcement rate is the same for each option). Thus, Response Δ s were calculated as a function of condition, such that every participant's unique score is contingent on the condition in which they participated. Statistically, participants in the Punishment condition exhibited significantly greater transfer effects (mean Response $\Delta = .16$ $SD \pm .27$) compared to those in the other two conditions (Fig 3: Compensate condition mean Response $\Delta = .003$ $SD \pm .30$, Indifferent condition mean Response $\Delta = .04$ $SD \pm .35$; ANOVA: $F(2, 289) = 6.14$, $p = .002$; post hoc LSD tests against Punishment condition: all P s < 0.01).

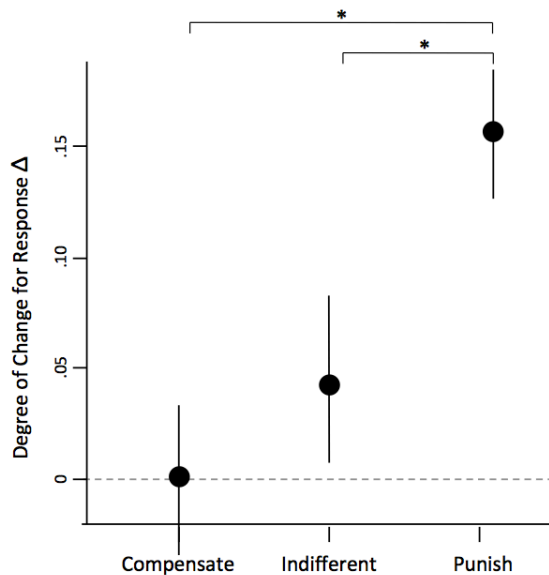


Fig 3 | Degree of transfer effect in Experiment 1. Amount of change in participants' responses from Baseline to Transfer phase in the three conditions. Results indicate the greatest transfer effects were unique to the Punishment condition. Error bars reflect 1 SEM.

Learning about another's moral preferences. These findings indicate that decisions to restore justice can be altered after learning about another's preferences, specifically when another

person holds markedly different moral values. To understand how individuals modify their choices in response to feedback in the Learning phase, we leveraged Reinforcement Learning (RL) models, which provide a useful computational framework for understanding how trial-by-trial learning unfolds. These RL models formalize how individuals adapt their behavior to recently observed outcomes. In particular, we modeled participant choice during the Learning phase as a function of feedback presented from the Receiver on the previous trial, allowing the influence of feedback to decay over time. A strength of this approach is that it allows us to test different psychological assumptions about how learning might occur. In the present study, we used these models to understand what aspects of the decision space (for example, unfairness levels) guide learning processes. Furthermore, our modeling approach makes no assumptions about the correct action—which ensures that we can use it across the three Learning conditions—and places no constraints on what the participant’s initial preferences are (which can often be problematic in the interpretation of delta-type measures).

The simplest model considered, termed the Basic model, assumes that people learn the values of the candidate actions (Accept, Compensate, or Reverse) based on direct experience with the actions and the feedback they experience (see Supplemental Materials for model details), but are agnostic to the fairness level of the offer at hand. In other words, the Basic model assumes that people are insensitive to the magnitude of the fairness infraction and will therefore respond in similar ways, despite the severity of violation. However, given that past research suggests that people deeply care about the magnitude of the fairness violations (Cameron 1999), we also explored whether learning is contingent on how unfairly the Receiver was treated. Thus, the next model considered—termed the Fairness Model—assumes that people are sensitivity to the fairness infraction (Cameron, 2003). If it is the case that social learning is sensitive to the extent of the fairness infraction, then such a model should do a better job of reflecting actual behavior than the Basic Model.

The last model we considered is termed the Extended Fairness Model. This model builds upon the Fairness model, but with the addition of a separate learning rate for each offer type, which are used to perform updates based on the fairness violation on the present trial. We conceived of this model based on the idea that depending on how sensitive people are to fairness violations, they may be differently reactive to feedback depending on the magnitude of the

transgression. Since higher learning rates allow estimates to reflect the consequences of more recent outcomes and lower learning rates reflect longer outcome histories—which leads to more stable estimates—having separate learning rates for each infraction level allows us to test whether the severity of the violation might hasten learning. In other words, people may be attuned to how unfairly one is being treated, and this might lead to faster learning depending on how egregious the infraction is. Together, these three models capture a range of how sensitive people are to various fairness violations, and afford an understanding of how an individual's preferences change from trial-to-trial in response to feedback from the Receiver.

This formal model based approach supported the behavioral findings that people acquire the punitive preferences of another. First, we found that the Fairness Model best fit participants' learning patterns (Table 2): it consistently had the lowest average Bayesian Information Criterion (BIC) scores—which quantifies model goodness of fit, penalizing for model complexity—indicating that participants were sensitive to the extent of the fairness infraction, but did not exhibit different learning rates (i.e., how efficiently they updated) across each offer type. Simply put, when learning about the moral preferences of others, people are in fact sensitive to how unfairly the victim is treated.

Second, this reinforcement-learning mechanism seems to accurately describe moral learning in this task. Comparing model goodness-of-fit across the Punishment, Indifferent, and Compensate conditions revealed that participants in the Compensate condition were best characterized by this learning account (insofar as having the best model fit), while participants in the Punishment condition had the worst model fit (Table 2). While at first blush this pattern of results may appear to conflict with the robust transfer effects observed in the Punishment condition, it should be noted that participants in the Compensate condition were receiving feedback that already aligned with their behavior. Thus, there was little to learn (i.e., small prediction errors) in the Compensate condition as the Receiver's feedback already reflected the participant's own preferences. In contrast, there was much to learn and implement in the Punishment condition, as the Receiver's preferences significantly differed from participants' initial preferences. In the Indifferent condition, Receivers were expressing slightly more punitive preferences (33%) compared to participants' baseline punitive preferences (23%), which amounts to a smaller learning gap and thus slightly better model fits compared to the Punishment condition.

Table 2 | Summary of Model Goodness-of-fit Metrics for Experiment 1

RL Models	Mean (SE) BIC scores by Condition			
	<i>Compensate</i>	<i>Indifferent</i>	<i>Punishment</i>	<i>Average</i>
<i>Basic Model</i>	137.61 (3.94)	180.81 (3.01)	154.75 (6.38)	157.92 (2.89)
<i>Fairness Model*</i>	87.68 (6.04)*	118.56 (4.90)*	132.20 (5.36)*	113.1 (3.31)*
<i>Extended Fairness Model</i>	98.71 (6.07)	129.57 (4.87)	143.29 (5.44)	124.18 (3.32)

Experiment 2

Given that the transfer effects were specific to the punishment condition, our goal in Experiment 2 was to probe the generality of these acquired punitive preferences within a laboratory setting. Accordingly, in the next experiment we only tested the Punishment condition.

Method

Participants. For Experiment 2, we recruited 37 participants (23 females, mean age=23.4 SD±4.56) from New York University and the surrounding New York City community to serve as a within laboratory replication of the condition of interest—the Punishment condition. Our aim was to recruit 35 participants (sample size determined from previous work; (Murty et al. 2016) on the assumption that some participants would fail to believe the experimental structure and should be removed before analysis. While two participants reported high disbelief during funnel debriefing, because the results remain robust with or without their inclusion, we report behavior from all 37 participants. Participants were paid an initial \$15 and an additional monetary bonus accrued during the task from one randomly realized trial from either the Baseline or Transfer phases (between \$1-\$9). Informed consent was obtained from each participant in a manner approved by the University Committee on Activities Involving Human Subjects.

Task. Experiments 1 and 2 were the same in their structure and aim, with only two key differences. First, Experiment 2 was run within the laboratory. This demanded a slightly more comprehensive cover story of who the other players were (see SI), which reduced the perception of anonymity compared with the players in the online version of the task

(Experiment 1). Second, in Experiment 2, unfair offers made to participants were of much larger monetary value—all splits were made out of \$10—eliciting justice violations of greater magnitude.

Results

Baseline Behavior. Participants exhibited a remarkably similar pattern of baseline behavior as those in Experiment 1, forgoing the free punitive option and instead choosing to Compensate at high rates of 70% of the time (Fig 4A). As before, participants endorsed Accept 9% of the time, and Reverse 21% of the time during the Baseline phase.

Transfer effects. After learning about another who strongly prefers to punish the perpetrator, participants demonstrated a 15% increase in endorsing the punitive option when deciding for as the victim (rmANOVA DV: Endorsement of Reverse, IV: Phase, $F(1,36)=7.35$, $p<0.01$, $\eta^2=.17$, Fig 4B)—replicating the findings from Experiment 1. Finally, as we had done in Experiment 1, by comparing model fits, we again found evidence that the Fairness Model best characterizes overall learning (lowest BIC scores, Table S3).

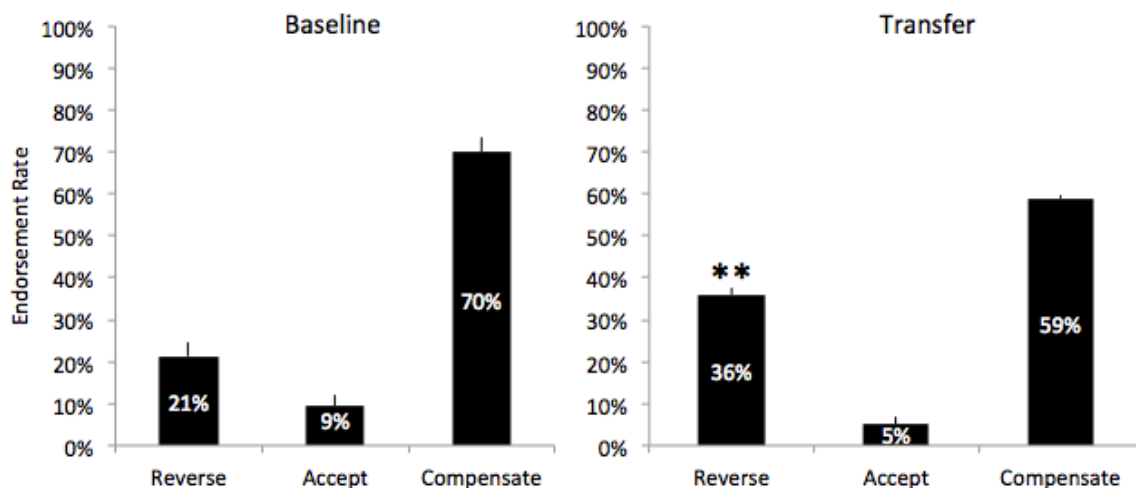


Fig 4 | Baseline and Transfer behavior in Experiment 2. Results reveal that participants significantly increased their endorsement of Reverse after learning about another person that values punishment.

Experiment 3

To further test the generalizability of our findings, we next Investigated these transfer effects in a different social context. In the Ultimatum Game—which leverages a classic game theoretic

setting—the only way to restore justice is to punish Player A. Thus, punishment rates are typically higher in the Ultimatum Game compared to the Justice Game. If under these circumstances individuals still learn to acquire the punitive preferences of another, it would provide robust converging evidence across contexts that a reinforcement mechanism governs how people learn to value moral preferences. In addition, given the task structure of the Justice Game, we wanted to be sure that our participants perceived a [\$9, \$1] or [\$.90, \$.10] split from Player A as a truly unfair offer. Because offers consisting of 10% of the pie are clearly—and unambiguously—intended to maximize Player A's payout in the Ultimatum Game, this additional testbed would ensure that in our previous experimental set-ups Player B was perceiving Player A's offers as unfair.

Method

Participants. In Experiment 3, 51 participants (21 females, mean age=34.06 SD±9.73) were recruited from AMT. Participants were paid an initial \$2.50 and an additional monetary bonus accrued during the task from one randomly realized trial from either the Baseline or Transfer phases. Informed consent was obtained from each participant in a manner approved by the University Committee on Activities Involving Human Subjects.

Task. The Ultimatum Game followed the same task parameters as those described in the previous experiments, only this time Player A acted as a proposer and Player B (the participant) acted as the responder. As in traditional Ultimatum Games, Player B could either accept the offer as is, or, reject the offer (with no option to compensate), which ensures that neither player receives a payout (Fig 5A). Effectively, rejecting is formalized as a way of enacting costly punishment on Player A.

Results

The pattern of results in this Ultimatum Game replicate the other experiments. After learning about a Receiver (another Player B in the Learning Phase) who strongly prefers to reject the offer and punish the perpetrator, participants exhibited a 12% increase in endorsing the punitive option when deciding for the self in the Transfer Phase (mean Response $\Delta=.12$ SD±.40; rmANOVA DV=Endorsement of Reverse, IV=Phase, $F(1,50)=9.01$, $p<0.004$, $\eta^2=.15$, Fig 5B). A comparison of the reinforcement learning models revealed the Fairness Model again

outperformed the other models (SI) indicating that participants were sensitive to the extent of fairness violation, but did not exhibit different learning rates for each offer type.

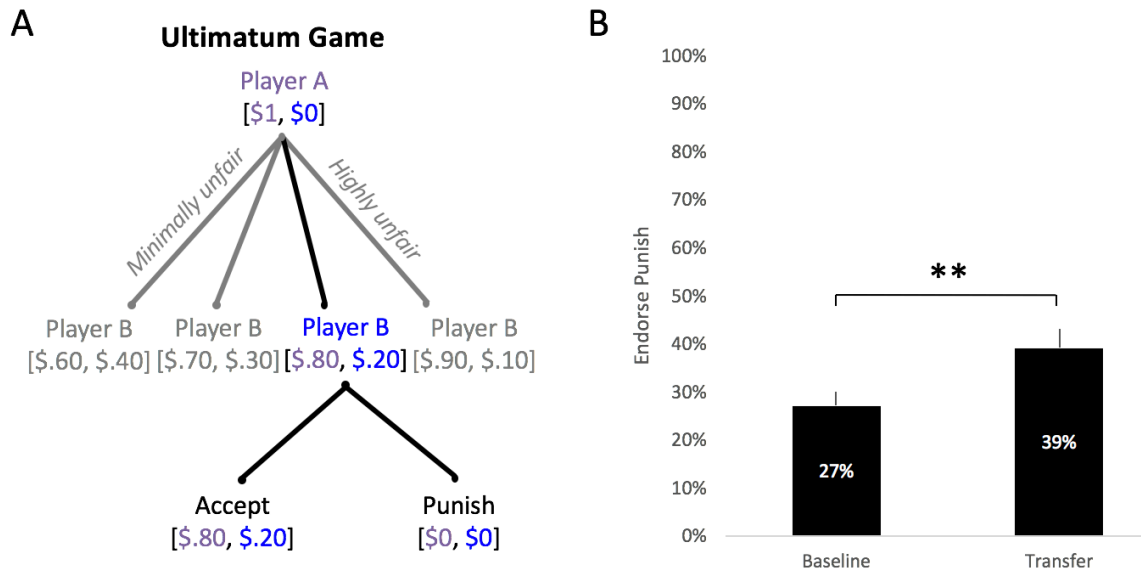


Fig 5 | Experiment 3. A) The same task structure described in the previous experiments—a Baseline, Learning and Transfer phase—was also employed, however, here subjects completed the Ultimatum game, where they could either choose to accept the offer from Player A as is, or reject the offer, thereby punishing Player A. **B)** Replicating the previous experiments, we found that after learning about a Receiver who wanted Player A to be punished, participants increased their own desire to punish by 12%.

Experiment 4

In Experiments 1-3 we found that learning about another who strongly prefers punishment as a means of restoring justice changes how readily one endorses punitive measures when they themselves are the victim. It is less clear, however, whether the acquisition of another's preferences requires active learning (i.e., implementation of the Receiver's preferences) or whether passive, observational learning (i.e., mere exposure to the Receiver's preferences) is enough to elicit a similar pattern of behavioral acquisition (i.e., a contagion effect). To examine this question, we ran a fourth experiment where participants either completed an active learning task (akin to those described in the previous experiments) or a passive learning task, in which participants simply observed a Receiver punitively responding to fairness violations.

In addition, a lingering question is whether certain phenotypic trait profiles result in greater acquisition of moral preferences. It is possible, for example, that the more receptive and

empathic an individual is (Preston 2013), the more likely one is to acquire another's moral values (Hoffman 2000). By this logic, those who are better at perspective taking or feeling empathic concern may be more readily able and motivated to simulate the desires of another (Hastings et al. 2000), which in turn could enable better learning and greater acquisition of another's punitive preferences. In Experiment 4 we tested this theory, positing that those high in trait empathy would demonstrate greater behavioral modification than those who are low in trait empathy.

Method

Participants. In Experiment 4, 200 participants (96 females, mean age=35.4 SD±10.5) were recruited from AMT and randomly selected to partake in either the active or passive learning task (a between-subjects design with N=100 per task, although five participants in the active learning condition failed to complete the experiment, resulting in N=95 for the active learning condition). Participants were paid an initial \$2.50 and an additional monetary bonus accrued during the task from one randomly realized trial from either the Baseline or Transfer phases. Informed consent was obtained from each participant in a manner approved by the University Committee on Activities Involving Human Subjects.

Task. The active learning condition was the same as the Punishment condition described in Experiment 1, with the addition of the Interpersonal Reactivity Index (Davis 1983) administered at the end of the experiment. This enabled us to collect individual trait measures of how well each participant engages in empathic concern (EC) and perspective taking (PT). The passive learning condition followed a similar structure as described in Experiment 1, with the following differences: During the Learning phase of the task, participants observed a Receiver making their own decisions about how to reapportion the money. For example, in the first trial of the Learning phase, Player A might offer the Receiver a [\$.90, \$.10] split. The Receiver—who adhered to an algorithm of selecting the Reverse option 90% of the time (as was done in all the previous experiments)—would then typically choose to Reverse the outcomes, such that Player A received [\$.10] and the Receiver [\$.90]. After watching the Receiver make her choice, the participant was asked what the Receiver had just decided. By having the participant key in a response, we were able to structurally match the passive learning condition to the active learning condition, while ensuring that participants were also engaged across the Learning

phase. How well a participant attended to and reported the Receiver's decisions was denoted as accuracy, such that accuracy was computed by comparing the participant's response on a given trial with what the Receiver had actually selected on that trial. Thus, feedback in the passive learning condition is akin to how well a participant paid attention to what the Receiver was deciding (rather than implementing what the Receiver desired, as was indexed in all previous experiments). At the end of the passive learning task, we also collected the Interpersonal Reactivity Index measure.

Results.

Baseline Behavior. As expected, we observed no differences in baseline behavior between the active and passive learning conditions (Table 3, Fig 6). As in the previous experiments, regardless of the condition, participants choose to Compensate at high rates (active learning: 72% endorsed Compensate, 17% endorsed Reverse, and 11% endorsed Accept; passive learning: 70% endorsed Compensate, 21% endorsed Reverse, and 9% endorsed Accept).

TABLE 3 | Mixed Effects logistic regression predicting endorsement of Reverse option in Experiment 3.

$$Reverse\ Rate_{i,t} = \beta_1 Condition + \varepsilon$$

DV	Coefficient (β)	Estimate (SE)	z-value	P value
Reverse	Intercept	0.168 (.02)	6.78	<0.001***
	Passive Learning	0.04 (.02)	1.76	0.08

*Note: Where Reverse (1=Reverse, otherwise 0) is indexed by subject and trial and condition is an indicator variable, such that the Active learning condition serves as the intercept. *** $p < .001$*

Transfer effects. In the active learning condition, learning about another who strongly prefers to punish the perpetrator resulted in a 12% increase in endorsing the punitive option (rmANOVA DV=Endorsement of Reverse, IV=Phase, $F(1,94)=14.2$, $p < 0.001$, $\eta^2=.13$, Fig 6), and we again found evidence that the Fairness Model best characterized overall learning (Table S6). In the passive learning condition, observing another person respond to offers with punishment also resulted in a 12% increase in endorsing the punitive option (rmANOVA DV=Endorsement of Reverse, IV=Phase, $F(1,99)=14.3$, $p < 0.001$, $\eta^2=.13$, Fig 6). To directly assess whether there were any significant differences in how readily one acquires preferences to punish during active learning versus mere exposure, we ran an independent samples t-test on Response Δ scores across the two tasks. We observed no differences in overall acquisition (independent t-test

between active and passive learning tasks, $t(193)=-0.07$, $p=.94$; Fig 6B) nor any differences at each level of fairness infraction (all $P_s>.25$), suggesting that mere exposure without implementation also results in acquiring the punitive preferences of another.

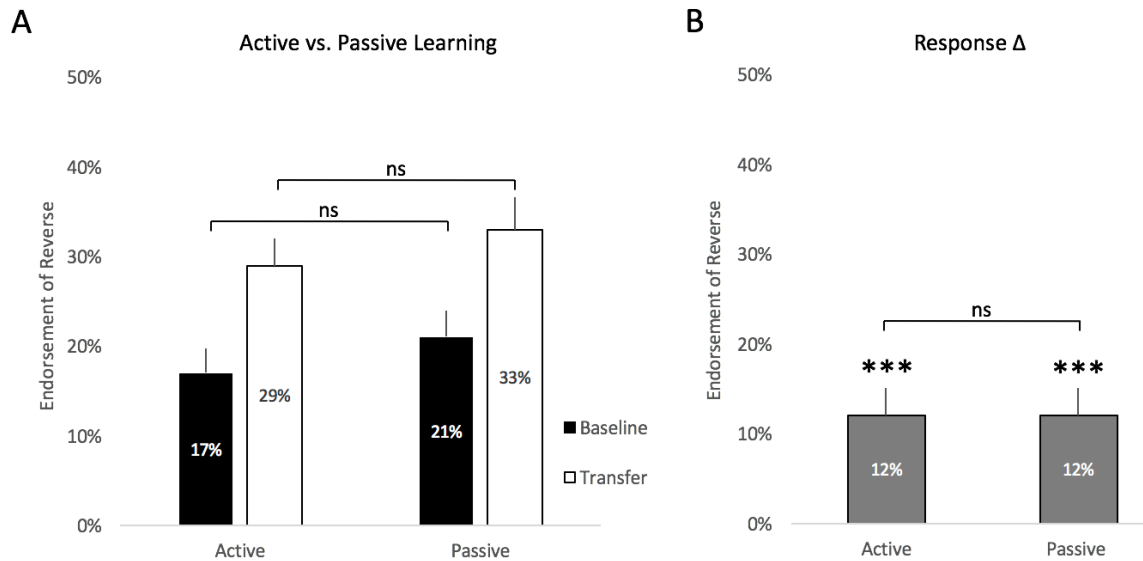


Fig 6 | Active versus Passive learning during Experiment 3. A) In the Baseline phase of the task, we observed no differences in how readily one endorsed Reverse (the punitive response) between the active and passive learning conditions. There were also no differences between conditions for the endorsement of Compensate (approximately 70%) or Accept (approximately 10%) options. **B).** While the overall increase in acquiring another’s punitive preferences was significant for both the active and passive conditions (all $P_s<0.001$), there were no differences in Response Δs across the two conditions—regardless of whether participants were required to actively learn about another’s preferences or merely observe another make choices for themselves.

Effects of Empathy on Learning. As would be expected, participants in the active and passive learning conditions exhibited similar empathic abilities—both in taking the perspective of another (active learning mean $PT=18.4$ $SD\pm6.15$; passive learning mean $PT=19.1$ $SD\pm5.27$, independent t -test, $t(193)=-.82$, $p=.42$), and in feeling empathic concern (active learning mean $EC=19.3$ $SD\pm6.12$; passive learning mean $EC=19.06$ $SD\pm6.97$, independent t -test, $t(193)=-.27$, $p=.79$). Surprisingly, we did not find evidence that either perspective taking or empathic concern abilities resulted in greater acquisition of another’s preferences (correlations between Response Δ x IRI subscales in both the active and passive conditions, all $P_s>.30$), suggesting that empathic abilities may not act as a proximal mechanism for conformist behavior under these circumstances.

DISCUSSION

Similar to the old adage—‘we are like chameleons, we take our hue and the color of our moral character from those who are around us’ (Locke 1824, Chartrand and Bargh 1999)—here we found that fairness preferences to restore justice are shaped by the moral preferences of those around us. More specifically, participants who were exposed to another individual exhibiting a strong taste for punishment increased their own endorsement of punishment when later deciding how to restore justice for themselves. The acquisition of another’s fairness preferences was contingent on learning from social feedback, and recruited mechanisms akin to those deployed in basic reinforcement learning.

The convergent findings of these four experiments highlight the delicate nature of our moral calculus, revealing that preferences for fairness can be remarkably shaped by indirect, peripheral social factors. Rather than employing prescriptive judgments of justice in an absolutist manner (Turiel 1983), we observed that responses to fairness infractions are malleable even in the absence of social pressure (Milgram 1974) or perception of social consensus (Asch 1956, Kundu and Cummins 2013), and took as little as learning the preferences of one single individual (Latane 1981).

The fact that we observed behavioral modifications stemming from such subtle social dynamics extends the classic findings of individuals conforming under more obvious social pressure (Cialdini and Goldstein 2004) and is particularly surprising given that these foundational moral principles are assumed to be resolutely held (Piaget 1932, Tetlock 2003). Evidence of increasing one’s own valuation of punishment suggests that learning about how others navigate fairness violations may signal other, more socially accepted routes or prescriptions for restoring justice (Chudek and Henrich 2011, Rand and Nowak 2013). Accordingly, learning about another person’s preferences in the absence of explicit social demands appears to be powerful enough to indicate the existence of alternative social norms—which, as research has repeatedly shown—is a key factor in conformist behavior (Cialdini 1998, Cialdini and Goldstein 2004).

That exposure to someone who holds different moral values alters which social norm is relevant to the situation at hand, indicates that people naturally infer which behaviors are socially normative and adjust their own behavior accordingly. Indeed, given that participants interacted

with many different unfair players over the course of our task, it is highly unlikely that participants were merely increasing their rate of punishment in the belief that they would be able to change an unfair player's subsequent actions (and debriefing statements support this, see SI). Our results build on the foundational idea that how others behave—for example, picking up litter or denouncing bullying—can cue the endorsement of a new social norm, which can have pronounced effects on shifting perceptions of what is a desirable behavior (Cialdini et al. 1990, Paluck et al. 2016). Here, we find that this is the case even within the moral domain: individuals use another's moral preference as a barometer of what is socially acceptable or appropriate when deciding how to restore justice.

This effect—that when another person's valuation of punishment is powerfully expressed individuals readily upwardly adjust how much they value punishment themselves—appeared to be driven by a reinforcement mechanism. This suggests that punitive preferences are swiftly assimilated when there is a significant difference between another's moral values and one's own moral values (Daw and Doya 2006, Daw and Frank 2009). These behavioral modifications were influenced both when implementing another's preferences and through observational learning routes. In other words, active learning and mere exposure seems to similarly guide the acquisition of another's moral preferences. That both learning processes resulted in the alteration of how punitive value computations are represented and subsequently expressed, intimates that learning from others provides a potent cognitive mechanism by which moral preferences can be developed and shaped.

Historically, the notion that the human moral calculus should be stable in the face of social demands describes how moral attitudes are taken as ethical blueprints for action (Kant 1785, Forsyth 1980, Rawls 1994). That moral values are believed to play an integral and defining role in shaping a person's identity (Turiel 1983, Haidt 2001), explains why moral attitudes and preferences are often so sacredly held (Tetlock 2003). And yet the work here illustrates how the human moral experience can also be idiosyncratic and fickle. Continuing to tune our knowledge about how individuals come to understand the difference between right and wrong, acquire principles of fairness and harm, and ultimately make moral decisions, can not only help shed light on which factors are most influential in guiding moral choice, but also the computational mechanisms underpinning this complex learning process.

Context of Research. This research was born out of our prior work where we observed that when individuals are presented with non-punitive options for restoring justice, they strongly prefer not to punish the perpetrator—a finding that stood in stark contrast to much of the existing literature (FeldmanHall et al, 2014). Naturally, this led us to wonder which social contexts would spur people to punish. While there is good evidence that how other people value certain stimuli (e.g., another’s attractiveness, food preferences, risk profiles) can influence our own preferences for the same stimuli, much less work has focused on how another’s moral preferences can bias our own moral actions. Therefore, this set of experiments aimed to test whether social influence could alter our moral values. We were also interested in whether the process of acquiring another’s moral preferences would entail active learning (e.g., implementing the actual preferences of another) or mere exposure to another’s moral preferences. By fitting a series of learning models to our data, we were able to directly test if the acquisition of another’s moral preferences—in this case, one’s desire to punish—is a function of active learning or mere exposure.

REFERENCES

- Andrew Luttrell, R. E. P., Pablo Briñol, Benjamin C. Wagner (2016). "Making it moral: Merely labeling an attitude as moral increases its strength." Journal of Experimental Social Psychology **65**(82).
- Aramovich, N. P., B. L. Lytle and L. J. Skitka (2012). "Opposing torture: Moral conviction and resistance to majority influence." Social Influence **7**(1): 21-34.
- Asch, S. E. (1956). "Studies of Independence and Conformity .1. A Minority of One against a Unanimous Majority." Psychological Monographs **70**(9): 1-70.
- Bardsley, N. and R. Sausgruber (2005). "Conformity and reciprocity in public good provision." Journal of Economic Psychology **26**(5): 664-681.
- Bauman, C. W. and L. J. Skitka (2009). "In the Mind of the Perceiver: Psychological Implications of Moral Conviction." Moral Judgment and Decision Making **50**: 339-362.
- Baumeister, R. F. and M. R. Leary (1995). "The Need to Belong - Desire for Interpersonal Attachments as a Fundamental Human-Motivation." Psychological Bulletin **117**(3): 497-529.
- Bowles, S. and H. Gintis (2011). A Cooperative Species: Human Reciprocity and Its Evolution.
- Brosnan, S. F. and F. B. De Waal (2003). "Monkeys reject unequal pay." Nature **425**(6955): 297-299.
- Buhrmester, M., T. Kwang and S. D. Gosling (2011). "Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?" Perspectives on Psychological Science **6**(1): 3-5.
- Cameron, L. A. (1999). "Raising the stakes in the ultimatum game: Experimental evidence from Indonesia." Economic Inquiry **37**(1): 47-59.
- Campbell-Meiklejohn, D. K., D. R. Bach, A. Roepstorff, R. J. Dolan and C. D. Frith (2010). "How the Opinion of Others Affects Our Valuation of Objects." Current Biology **20**(13): 1165-1170.
- 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**(2): 284-299.
- Chartrand, T. L. and J. A. Bargh (1999). "The Chameleon effect: The perception-behavior link and social interaction." Journal of Personality and Social Psychology **76**(6): 893-910.
- Chudek, M. and J. Henrich (2011). "Culture-gene coevolution, norm-psychology and the emergence of human prosociality." Trends in Cognitive Sciences **15**(5): 218-226.
- Cialdini, R. B. and N. J. Goldstein (2004). "Social influence: Compliance and conformity." Annual Review of Psychology **55**: 591-621.
- Cialdini, R. B., R. R. Reno and C. A. Kallgren (1990). "A Focus Theory of Normative Conduct - Recycling the Concept of Norms to Reduce Littering in Public Places." Journal of Personality and Social Psychology **58**(6): 1015-1026.
- Cialdini, R. B. T., M.R. (1998). Social influence: Social norms, conformity and compliance. The handbook of social psychology. D. T. F. Gilbert, Susan T. New York, NY, McGraw-Hill.
- Crump, M. J. C., J. V. McDonnell and T. M. Gureckis (2013). "Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research." PLoS One **8**(3).
- Davis, M. H. (1983). "Measuring Individual Differences in Empathy: Evidence for a Multidimensional Approach." Journal of Personality and Social Psychology **44**(1): 113-126.
- Daw, N. D. and K. Doya (2006). "The computational neurobiology of learning and reward." Curr Opin Neurobiol **16**(2): 199-204.
- Daw, N. D. and M. J. Frank (2009). "Reinforcement learning and higher level cognition: introduction to special issue." Cognition **113**(3): 259-261.
- de Waal, F. (1996). Good natured: The origin of right and wrong in humans and other animals, Harvard University Press.

de Waal, F. (2006). Primates and Philosophers, How Morality Evolved, Princeton University Press.

Deutsch, M. (1955). "A Study of Normative and Informational Social Influences Upon Individual Judgment." The Journal of Abnormal and Social Psychology **51**(3): 14-14.

Dreber, A., D. G. Rand, D. Fudenberg and M. A. Nowak (2008). "Winners don't punish." Nature **452**(7185): 348-351.

Falk, A., E. Fehr and U. Fischbacher (2008). "Testing theories of fairness - Intentions matter." Games and Economic Behavior **62**(1): 287-303.

Fehr, E., H. Bernhard and B. Rockenbach (2008). "Egalitarianism in young children." Nature **454**(7208): 1079-U1022.

Fehr, E. and U. Fischbacher (2004). "Third-party punishment and social norms." Evolution and Human Behavior **25**(2): 63-87.

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

FeldmanHall, O., P. Sokol-Hessner, J. J. Van Bavel and E. A. Phelps (2014). "Fairness violations elicit greater punishment on behalf of another than for oneself." Nature Communications **5**.

Forsyth, D. R. (1980). "A Taxonomy of Ethical Ideologies." Journal of Personality and Social Psychology **39**(1): 175-184.

Fowler, J. H. and N. A. Christakis (2010). "Cooperative behavior cascades in human social networks." Proc Natl Acad Sci U S A **107**(12): 5334-5338.

Gintis, H. and E. Fehr (2012). "The social structure of cooperation and punishment." The Behavioral and brain sciences **35**(1): 28-29.

Guth, W., R. Schmittberger and B. Schwarze (1982). "An Experimental-Analysis of Ultimatum Bargaining." Journal of Economic Behavior & Organization **3**(4): 367-388.

Haidt, J. (2001). "The emotional dog and its rational tail: a social intuitionist approach to moral judgment." Psychol Rev **108**(4): 814-834.

Haidt, J., K. S. G. (2010). Handbook of Social Psychology. Chapter 22: Morality. Hoboken, NJ, John Wiley & Sons.

Haney, C., Banks, W. C., & Zimbardo, P. G. (1973). "Study of prisoners and guards in a simulated prison." Naval Research Reviews **9**(30): 1-17.

Hastings, P. D., C. Zahn-Waxler, J. Robinson, B. Usher and D. Bridges (2000). "The development of concern for others in children with behavior problems." Developmental Psychology **36**(5): 531-546.

Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis, R. McElreath, M. Alvard, A. Barr, J. Ensminger, N. S. Henrich, K. Hill, F. Gil-White, M. Gurven, F. W. Marlowe, J. Q. Patton and D. Tracer (2005). "Models of decision-making and the coevolution of social preferences." Behavioral and Brain Sciences **28**(6): 838-855.

Hoffman, M. (2000). Empathy and Moral Development: implications for caring and justice, Cambridge University Press.

Hornsey, M. J., L. Majkut, D. J. Terry and B. M. McKimmie (2003). "On being loud and proud: Non-conformity and counter-conformity to group norms." British Journal of Social Psychology **42**: 319-335.

Horton, J. J., D. G. Rand and R. J. Zeckhauser (2011). "The online laboratory: conducting experiments in a real labor market." Experimental Economics **14**(3): 399-425.

Izuma, K. and R. Adolphs (2013). "Social Manipulation of Preference in the Human Brain." Neuron **78**(3): 563-573.

Jones, W. K. (1994). "A theory of social norms." University of Illinois Law Review **3**.

Jordan, J. J., M. Hoffman, P. Bloom and D. G. Rand (2016). "Third-party punishment as a costly signal of trustworthiness." Nature **530**(7591): 473-+.

Kant, I. (1785). Grounding for the Metaphysics of Morals, Hackett.

Kouchaki, M. (2011). "Vicarious Moral Licensing: The Influence of Others' Past Moral Actions on Moral Behavior." Journal of Personality and Social Psychology **101**(4): 702-715.

Kundu, P. and D. D. Cummins (2013). "Morality and conformity: The Asch paradigm applied to moral decisions." Social Influence **8**(4): 268-279.

Latane, B. (1981). "The Psychology of Social Impact." American Psychologist **36**(4): 343-356.

Locke, J. (1824). The works of John Locke, in nine volumes. London,, Printed for C. and J. Rivington etc.

Lotz, S., T. G. Okimoto, T. Schlosser and D. Fetchenhauer (2011). "Punitive versus compensatory reactions to injustice: Emotional antecedents to third-party interventions." J Exp Soc Psychol **47**(2): 477-480.

Mason, W. and S. Suri (2012). "Conducting behavioral research on Amazon's Mechanical Turk." Behav Res Methods **44**(1): 1-23.

Mathew, S. and R. Boyd (2011). "Punishment sustains large-scale cooperation in prestate warfare." Proceedings of the National Academy of Sciences of the United States of America **108**(28): 11375-11380.

Milgram, S. (1963). "Behavioral Study of Obedience." J **67**: 371-378.

Milgram, S. (1974). Obedience to Authority: An Experimental View, Harper and Row, Publishers, Ince.

Moll, J., R. Zahn, R. de Oliveira-Souza, F. Krueger and J. Grafman (2005). "Opinion: the neural basis of human moral cognition." Nature reviews. Neuroscience **6**(10): 799-809.

Murty, V. P., O. FeldmanHall, L. E. Hunter, E. A. Phelps and L. Davachi (2016). "Episodic memories predict adaptive value-based decision-making." Journal of Experimental Psychology-General **145**(5): 548-558.

Nook, E. C., D. C. Ong, S. A. Morelli, J. P. Mitchell and J. Zaki (2016). "Prosocial Conformity: Prosocial Norms Generalize Across Behavior and Empathy." Pers Soc Psychol Bull.

Paluck, E. L., H. Shepherd and P. M. Aronow (2016). "Changing climates of conflict: A social network experiment in 56 schools." Proceedings of the National Academy of Sciences of the United States of America **113**(3): 566-571.

Paolacci, G., J. Chandler and P. G. Ipeirotis (2010). "Running experiments on Amazon Mechanical Turk." Judgment and Decision Making **5**(5): 411-419.

Peysakhovich, A., M. A. Nowak and D. G. Rand (2014). "Humans display a 'cooperative phenotype' that is domain general and temporally stable." Nature Communications **5**.

Peysakhovich, A. and D. G. Rand (2016). "Habits of Virtue: Creating Norms of Cooperation and Defection in the Laboratory." Management Science **62**(3): 631-647.

Piaget, J. (1932). The Moral Judgment of the Child. London, Routledge & Kegan Paul.

Pillutla, M. M. and J. K. Murnighan (1996). "Unfairness, anger, and spite: Emotional rejections of ultimatum offers." Organizational Behavior and Human Decision Processes **68**(3): 208-224.

Preston, S. D. (2013). "The origins of altruism in offspring care." Psychol Bull **139**(6): 1305-1341.

Rai, T. S. and A. P. Fiske (2011). "Moral Psychology Is Relationship Regulation: Moral Motives for Unity, Hierarchy, Equality, and Proportionality." Psychological Review **118**(1): 57-75.

Rand, D. G. and M. A. Nowak (2013). "Human cooperation." Trends in Cognitive Sciences **17**(8): 413-425.

Rawls, J. (1994). "Theory of Justice." Voprosy Filosofii(10): 38-52.

Skitka, L. J., C. W. Bauman and E. G. Sargis (2005). "Moral conviction: Another contributor to attitude strength or something more?" Journal of Personality and Social Psychology **88**(6): 895-917.

Spitzer, M., U. Fischbacher, B. Herrnberger, G. Gron and E. Fehr (2007). "The neural signature of social norm compliance." Neuron **56**(1): 185-196.

Straub, P. G. and J. K. Murnighan (1995). "An Experimental Investigation of Ultimatum Games - Information, Fairness, Expectations, and Lowest Acceptable Offers." Journal of Economic Behavior & Organization **27**(3): 345-364.

Sturgeon, N. (1985). Moral explanations. Morality, reason, and truth. D. C. D. Zimmerman, Rowman & Allanheld.

Sutton, R. S. B., A.G. (1998). Reinforcement Learning: An Introduction.

Suzuki, S., E. L. Jensen, P. Bossaerts and J. P. O'Doherty (2016). "Behavioral contagion during learning about another agent's risk-preferences acts on the neural representation of decision-risk." Proc Natl Acad Sci U S A.

Tetlock, P. E. (2003). "Thinking the unthinkable: sacred values and taboo cognitions." Trends Cogn Sci **7**(7): 320-324.

Turiel, E. (1983). The Development of Social Knowledge: Morality and Convention. Cambridge, UK, Cambridge University Press.

Wood, W. (2000). "Attitude change: persuasion and social influence." Annu Rev Psychol **51**: 539-570.

Zaki, J., J. Schirmer and J. P. Mitchell (2011). "Social influence modulates the neural computation of value." Psychol Sci **22**(7): 894-900.