

Advantageous and disadvantageous inequality aversion can be taught through vicarious learning of others' preferences

Shen Zhang , Oriel FeldmanHall, Sébastien Héту, A Ross Otto

State Key Laboratory of Cognitive Neuroscience and Learning & IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing, China • Department of Neurobiology, German Primate Center, Göttingen, Germany • Cognitive, Linguistics and Psychological Sciences, Brown University, Providence, United States • Carney Institute for Brain Sciences, Brown University, Providence, United States • Department of Psychology, Université de Montréal, Montréal, Canada • Centre Interdisciplinaire de Recherche sur le Cerveau et l'Apprentissage (CIRCA), Montréal, Canada • Department of Psychology, McGill University, Montréal, Canada

 https://en.wikipedia.org/wiki/Open_access

 Copyright information

eLife Assessment

This cleverly-designed and potentially **important** work supports our understanding regarding how and whether social behaviours promoting egalitarianism can be learned, even when implementing these norms entails a cost for oneself. However, the evidence supporting the major claims is currently **incomplete**, with major limitations being the statistical approach, the modelling, and over-interpretation. With a strengthening of the supporting evidence, this work will be of interest to a wide range of fields, including cognitive psychology/neuroscience, neuroeconomics, and social psychology, as well as policy making.

<https://doi.org/10.7554/eLife.102800.1.sa2>

Abstract

While enforcing egalitarian social norms is critical for human society, punishing social norm violators often incurs a cost to the self. This cost looms even larger when one can benefit from an unequal distribution of resources, a phenomenon known as advantageous inequity—for example, receiving a higher salary than a colleague with the identical role. In the Ultimatum Game, a classic testbed for fairness norm enforcement, individuals rarely reject (or punish) such unequal proposed divisions of resources because doing so entails a sacrifice of one's own benefit. Recent work has demonstrated that observing and implementing another's punitive responses to unfairness can efficiently alter the punitive preferences of an observer. It remains an open question, however, whether such contagion is powerful enough to impart advantageous inequity aversion to individuals—that is, can observing another's preferences to punish inequity result in increased enforcement of equality norms, even in the difficult case of AI? Using a variant of the Ultimatum Game in which participants are tasked

with responding to fairness violations on behalf of another ‘Teacher’—whose aversion to advantageous (versus disadvantageous) inequity was systematically manipulated—we probe whether individuals subsequently increase their punishment unfair after experiencing fairness violations on their own behalf. In two experiments, we found individuals can acquire aversion to advantageous inequity ‘vicariously’ through observing (and implementing) the Teacher’s preferences. Computationally, these learning effects were best characterized by a model which learns the latent structure of the Teacher’s preferences, rather than a simple Reinforcement Learning account. In summary, our study is the first to demonstrate that people can swiftly and readily acquire another’s preferences for advantageous inequity, suggesting in turn that behavioral contagion may be one promising mechanism through which social norm enforcement—which people rarely implement in the case of advantageous inequality—can be enhanced.

Introduction

Humans can learn how to navigate through the world by observing the actions of others. For example, individuals can learn complex motor skills by observing and imitating how experts coordinate their movements (Hayes et al., 2008). Observational learning can also transmit important information about social and moral norms, such as reciprocity (Engelmann & Fischbacher, 2009), cooperating with others (van Baar et al., 2019), or context in which punishment is considered appropriate (FeldmanHall et al., 2018). In some cases, an individual learns from others in a straightforward and unambiguous social context, where the tensions endemic to many moral dilemmas—e.g., benefit one-self versus the collective good—are not directly juxtaposed against one another. And yet, the social world is rarely straightforward, often ambiguous, and moral dilemmas that pit self-benefit over the collective good abound (FeldmanHall & Shenhav, 2019; Vives & Feldmanhall, 2018).

Consider the case of inequity for which individuals exhibit a strong distaste: concerns for fairness are well documented in adults (Güth et al., 1982; Sanfey et al., 2003), children (Fehr et al., 2008; McAuliffe & Dunham, 2017), primates (Brosnan & De Waal, 2003, 2014; Van Wolkenten et al., 2007) and even domesticated dogs (Essler et al., 2017). This aversion to inequity manifests perhaps most famously in the Ultimatum Game (UG), in which one player (the Proposer) decides how to split a sum of money with another player (the Receiver), a role typically assumed by the participant (Güth et al., 1982; Sanfey et al., 2003). A Receiver’s acceptance results in both parties receiving the offered money, whereas rejection results in neither party receiving any money—a form of costly punishment. A recurring observation supporting inequity aversion is that people tend to reject disadvantageous offers that unfairly benefit the other party (Brosnan & De Waal, 2014; Fehr & Schmidt, 1999).

At the same time, not all inequity is experienced in the same way. For example, when we stand to receive less than our “fair share”, such disadvantageous inequity (DI) engenders feelings of envy, anger, and/or disappointment (Heffner & FeldmanHall, 2022) which often manifests in punishment of unfair offers in the UG via rejection (McAuliffe et al., 2014, 2017; Pedersen et al., 2013; Pillutla & Murnighan, 1996). In contrast, when we stand to receive a favorable share of resources—albeit one that is unfairly distributed—these advantageous inequitable (AI) offers often engender feelings of guilt or shame (Gao et al., 2018). Despite these negative emotions, Receivers are much less willing to engage in costly punishment of offers that are advantageously inequitable (Civai et al., 2012; Hennig-Schmidt et al., 2008; Luo et al., 2018). In short, AI versus DI engender markedly different punishment preferences.

A growing body of developmental research demonstrates that this difference in punitive responses to AI versus DI manifests early in the developmental trajectory (Amir et al., 2023; Blake et al., 2015; Blake & McAuliffe, 2011; McAuliffe et al., 2017). Indeed, punishing AI

offers requires the sacrifice (larger) personal gain to achieve a fairer outcome—mirroring many moral dilemmas in which self-versus other-regarding interests are at odds—AI aversion likely necessitates complex cognitive abilities. More specifically, because the aversion to AI appears to arise from more abstract concerns about fairness (Tomasello, 2019), AI aversion is thought to impose considerable demands on more sophisticated cognitive processing (Gao et al., 2018). This may explain, in part, why the developmental trajectory of AI-averse preferences comes online much later relative to DI (McAuliffe et al., 2017).

Thus, one open question concerns how people acquire inequity-averse preferences. One influential framework posits that we often adapt our behaviors to people around us through a process of conformity (Cialdini & Goldstein, 2004). Put simply, loyally following the behaviors of another—a social contagion effect—is a powerful motivator of social behavior. Such behavioral contagion effects are observed in diverse decision-making domains such as valuation (Campbell-Meiklejohn et al., 2010), risk-taking (Suzuki et al., 2016), delay of gratification (Garvert et al., 2015), moral preferences (Bandura & McDonald, 1963; Vives et al., 2022), and social norms (Hertz, 2021). Recently, we demonstrated that punitive responses to DI can be ‘taught’ to participants in the context of the Ultimatum Game, whereby individuals’ preferences for rejecting disadvantageous unfair offers were strengthened as a result of observing another individual’s (a ‘Teacher’) desire to punish these offers (FeldmanHall et al., 2018).

At the same time, social contagion effects may be far less robust if the behavior demands sacrifice of self-benefit. Since this type of dynamic places behavioral contagion and desire for self-gain in opposition, it remains unclear whether AI aversion can also be acquired ‘vicariously’ through observing another’s preferences. Here, we investigate whether the act of observing the AI-averse preferences of another punitive Receiver enhances an individual’s aversion to AI, even if the rejection of such AI offers requires sacrificing self-benefit. Answering this question not only enriches our understanding of the nature of inequity aversion, but also enable us to better understand the mechanisms underpinning vicarious learning of moral preferences during social interactions.

One can imagine different possible computational mechanisms driving vicarious learning of inequity aversion. One possibility is that observational learning of moral preference is based on simple action-outcome contingencies—on this view, a simple but elegant Reinforcement Learning (RL; Burke et al., 2010; Diaconescu et al., 2020; FeldmanHall et al., 2018; Lindström et al., 2019) model formalizes how individuals adapt their behavior to recently observed outcomes. In its most basic form, RL makes choices on the basis of learned associations between actions and outcomes, and critically, actions are bound to the specific decision context—applied to the UG, the unfairness of the amount offered by the Proposer. However, during social interactions, people might not consider the behaviors of others as resulting from simple action-outcome associations but alternatively, construct and use models of other agents, representing their stable intentions, beliefs, and preferences (Anzellotti & Young, 2020). Accordingly, we also consider the possibility that moral preferences are immutable across contexts (Bail et al., 2018; Fehr & Schmidt, 1999; Taber & Lodge, 2012), which would suggest that moral preferences are not learned merely as associations (i.e. specific responses tied to different unfairness levels), but rather, through a deeper inference process which models the underlying fairness preferences of the observed individual. Here, across two experiments, we leverage a well-characterized vicarious learning paradigm (FeldmanHall et al., 2018; Son et al., 2019; Vives et al., 2022) to examine the conditions under which individuals are able to learn AI-averse preferences on the basis of exposure to another Receiver’s punitive preferences. In addition, to mechanistically probe how punitive preferences come to be valued in DI and AI contexts, we also characterize trial-by-trial acquisition of punitive behavior with computational models of choice.

Experiment 1

In Experiment 1, following the approach of previous experiments (FeldmanHall et al., 2018 [↗](#)), we test if the rejection of advantageously unfair offers can be learned on the basis of exposure to the preferences of another individual (the ‘Teacher’) who has AI-averse preferences. In a between subject design with three phases, participants interacted with other individuals in a repeated Ultimatum Games (**Figure 1a** [↗](#)). We assess contagion effects by measuring participants’ DI and AI aversion both before and after observing (and implementing) the preferences of a Teacher’s who exhibits inequity aversion in both DI and AI contexts (‘AI-DI-Averse’ condition; N =100) and inequity aversion only in a DI context (‘DI-Averse’ condition; N =100).

First, to assess participants’ baseline fairness preferences across inequity levels, in the Baseline Phase (**Figure 1c** [↗](#)) participants acted as a Receiver in several one-shot UGs, responding to offers ranging from extreme DI (e.g., the Proposer keeps 90 cents and offers 10 cents to the Receiver; a 90:10 split) to extreme AI (e.g., the Proposer keeps 10 cents and offers 90 cents to the Receiver; a 10:90 split). On each trial, participants interacted with a different Proposer, and unbeknownst to participants, the offers were pre-determined by the experimenters. Following the typical formulation of the UG (Güth et al., 1982 [↗](#)), participants made the choice between accepting versus rejecting each offer, and also rated the fairness of the offer.

Next, in the Learning Phase (**Figure 1d** [↗](#)), participants played a repeated vicarious UG as a third party, in which they accepted or rejected offers on behalf of another Receiver (termed the Teacher in this phase) such that the participant’s decisions did not impact their own payoff but would impact the payoffs to the Proposer and the Teacher. Critically, after each decision, participants received feedback whether the Teachers *would* have preferred acceptance versus rejection (i.e., punishment) of the offer. Thus, through trial-by-trial feedback, the Teacher can signal to participants their preference to punish the Proposer for making unfair offers. Critically, in the ‘DI-Averse’ condition, akin to the typical pattern of human preferences observed in the UG (Güth et al., 1982 [↗](#); Sanfey et al., 2003 [↗](#)), the Teacher’s preferences exhibited a strong aversion to DI, and thus routinely punishing unfair offers. Specifically, the Teacher was likely to reject DI offers (i.e. 90:10 and 70:30), but not AI offers (i.e. 30:70 and 10:90; see **Figure 1b** [↗](#) and **Table S1** [↗](#)). However, in the ‘AI-DI-Averse’ condition, the Teacher was likely to reject any unfair offer, regardless of whether it was DI or AI, manifesting typical DI-averse preferences as well as the less commonly observed aversion to AI (i.e., punishing advantageous offers). Feedback from the Teacher was also accompanied by fairness ratings consistent with their preferences (see **Figure 1b** [↗](#) and **Table S2** [↗](#)).

Finally, to examine contagion (or transmission) of the Teacher’s preferences to the participants, we assessed fairness preferences of the participants for a second time in a Transfer Phase (**Figure 1a** [↗](#)). This third phase was identical to the Baseline Phase and thus allowed us to quantify changes in participants’ fairness preferences before and after the Learning Phase.

Preferences Across Baseline and Transfer Phases

Mirroring preferences typically observed in Western, Educated, Industrialized, Rich, and Democratic (WEIRD) participant populations (Henrich et al., 2006 [↗](#)), punishment choices in the Baseline Phase were DI-averse, but not AI-averse, as we observed similarly high rejection rates during the Baseline Phase for DI offers across the DI-Averse and the AI-DI-Averse conditions (**Figure 2** [↗](#), **Table S3** [↗](#); all P s < 0.001). As expected, the rejection rates for AI offers were much lower than DI offers (**Figure 2** [↗](#), e.g. 90:10 vs 10:90, all P s < 0.001, see **Table S3** [↗](#) for estimates of rejection rates). Consistent with these Rejection rates, participants rated DI offers as unfair (i.e.,

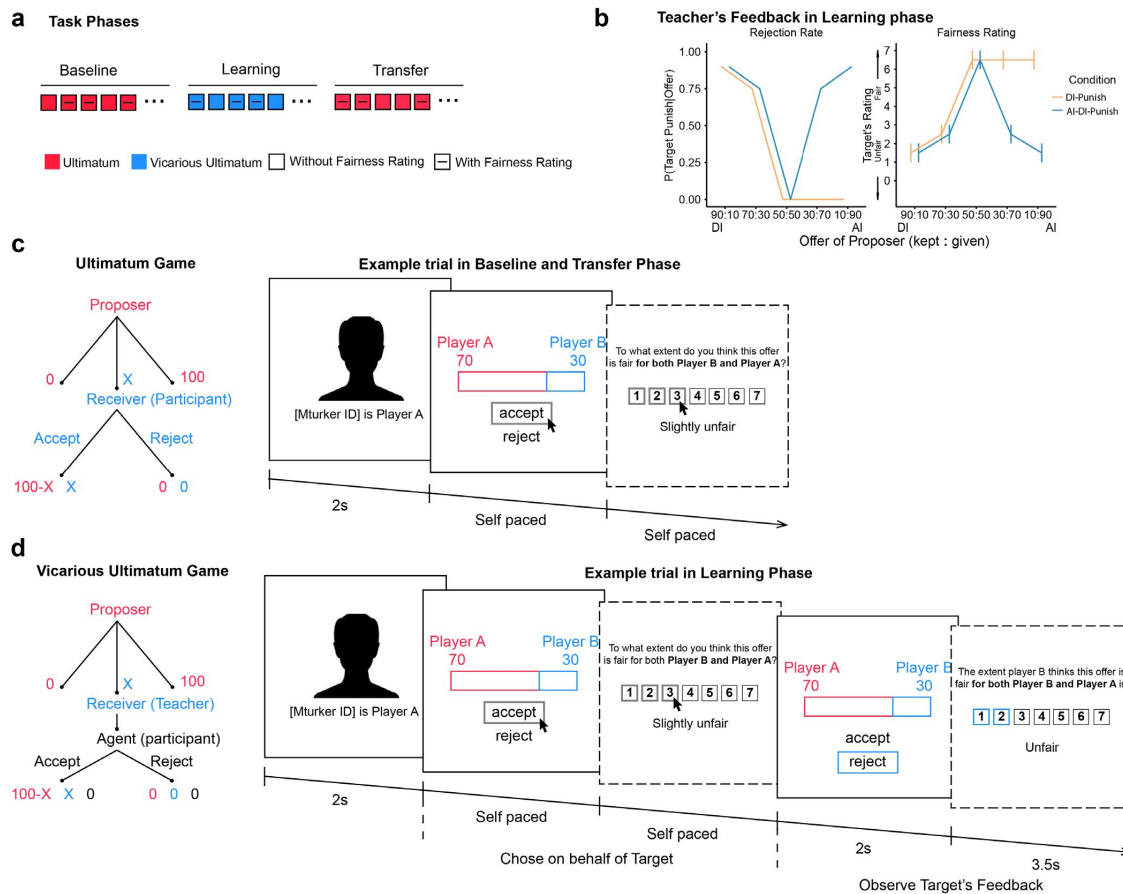


Figure 1.

a) Task Overview. Our main task consists of 3 phases. In the Baseline Phase participants acted as a Receiver, responding to offers of different inequity level and rated their perceived fairness towards the offers on three out of every five trials. While, in the subsequent Learning Phase, participants acted as an Agent, deciding on behalf of the Receiver (Teacher) and Proposer. Again, they rated the fairness on three out of every five trials. Finally, participants made choices in a Transfer Phase which was identical to the Baseline Phase. **b)** Preferences and Fairness Ratings governing the Teacher's feedback in the Learning Phase (See [Table S1](#) and [Table S2](#)). **c)** Baseline and Transfer phase, in which participants played the Ultimatum game as a Receiver, making choices on their own behalf. **d)** In the Learning phase, participants acted as a third party (the agent), making decisions on behalf of the Proposer and the Receiver (Teacher), playing a Vicarious Ultimatum game. In a Vicarious Ultimatum game, the Agent make decisions for the Receiver, if he/she rejects the proposed split, both the Proposer and the Receiver receive nothing. If he/she accepts, the Proposer and the Receiver are rewarded with the proposed split.

lower than the rating scale midpoint of 4, **Table S4** [↗](#); all P s < 0.001) and AI offers were rated as more fair than DI offers (all P s < 0.001)— despite the fact that the offers in the AI and DI contexts represent the same magnitude of inequity (e.g. 90:10 vs 10: 90 splits).

To examine whether exposure to the Teacher's punishment preferences in the Learning Phase resulted in changes in participants' fairness preferences, we examined changes in Rejection rates and Fairness rating between the Baseline and Transfer phases (**Figure 2** [↗](#)). First, in response to DI offers (90:10 and 70:30 splits), we observed robust increases in punishment rates after observing a Teacher who prefers punishment. That is, after exposure to the Teacher's punitive preferences in the Learning Phase, participants were more likely to reject offers that unfairly benefitted the Proposer in the Transfer Phase (**Figure 2a** [↗](#)), conceptually replicating our previous results (FeldmanHall et al., 2018 [↗](#)). These increases in punishment rates under DI contexts were statistically significant in both the AI-DI-Averse (**Table S5** [↗](#); 90:10 splits: β (SE)=0.14(0.03), p < 0.001; 70:30 splits: β (SE)=0.12(0.03), p < 0.001) and DI-Averse conditions (**Table S5** [↗](#), 90:10 splits: β (SE)=0.18(0.13), p < 0.001; 70:30 splits: β (SE)=0.14(0.03), p < 0.001). These preference changes were consistent with the stronger DI aversion exhibited by the Teacher (in both conditions) than participants in the Baseline Phase. However, we did not observe consistent fairness rating changes for DI offers, presumably because the fairness ratings given by the Teacher was similar to participants' baseline fairness ratings (**Figure 2b** [↗](#), **Table S6** [↗](#)) in both conditions.

Examining punishment choices in response to AI (30:70 and 10:90) splits, we observed that participants in the AI-DI-Averse condition increased their rates of rejection to AI offers (**Figure 2a** [↗](#)). That is, participants exposed to the Teacher's preferences to punish AI offers during the Learning Phase became significantly more likely, in the Transfer Phase, to reject unequal offers that stood to unfairly reap further monetary benefits for the participants (30:70 splits: β (SE)=0.09(0.03), p =0.002; 10:90 splits: β (SE)=0.12(0.03), p < 0.001; **Table S5** [↗](#)). Importantly, we did not observe transfer of AI-averse preferences in the DI-Averse condition, where the Teacher did not express a desire to punish the AI offers in the Learning phase (30:70 splits: p =0.666; 90:10 splits: p =0.280; **Table S5** [↗](#)). Mirroring changes in rejection rates, participants' fairness ratings also shifted towards those of the Teacher (**Figure 2b** [↗](#))— specifically, in the AI-DI-Averse condition, participants rated AI offers as more unfair in Transfer compared to Baseline Phase (**Table S6** [↗](#), 30:70 splits: β (SE)=-0.54(0.12), p < 0.001; 10:90 splits: β (SE)= 0.76(0.12), p < 0.001), while in the DI-Averse condition, participants rated AI offers as more fair (**Table S6** [↗](#), 30:70 splits: β (SE)=0.32(0.12), p < 0.001; 10:90 splits: β (SE)=0.28(0.12), p < 0.001). Exposure to the Teacher's behavior in response to “fair” offers (50:50 splits) did not appear to elicit changes in punishment rates in either the AI-DI-Averse (**Table S5** [↗](#), p =0.349) or the DI-Averse (**Table S5** [↗](#), p =0.517) condition, which suggests against the possibility that participants in the AI-DI-Averse condition were simply increasing punitive preferences, regardless of context. Finally, in line with the rejection rates results, for fair offers (**Table S6** [↗](#), 50:50 splits), we did not observe any change in fairness ratings in the AI-DI-Averse condition (**Table S6** [↗](#), p =0.662) and a slight increase in the DI-Averse condition (p =0.041).

Learning another's preferences

Having demonstrated that participants' preferences to reject unfair DI and AI offers were altered on the basis of exposure to the Teacher's preferences, we next examined the trial-by-trial changes in rejection rates during the Learning phase (**Figure 3a** [↗](#)). A mixed-effects logistic regression revealed a significant positive effect of trial number on rejection rates of DI Offers for the AI-DI-Averse condition (70:30 splits: β (SE)=0.36(0.09), p < 0.001; not significant in 90:10 splits: β (SE)=0.27(0.20), p =0.164; **Table S7** [↗](#)) and the DI-Averse condition (90:10 splits: β (SE)=0.63(0.20), p =0.001; 70:30 splits: β (SE)=0.31(0.09), p < 0.001), indicating an increase of rejection rates during the Learning phase. In contrast, examining responses to AI offers, we only observed a learning effect—that is, a rejection rate increase—in the AI-DI-Averse condition (30:70 splits: β (SE)=0.47(0.12), p < 0.001; 10:90 splits: β (SE)=0.77(0.18), p < 0.001) where the Teacher imparted AI-

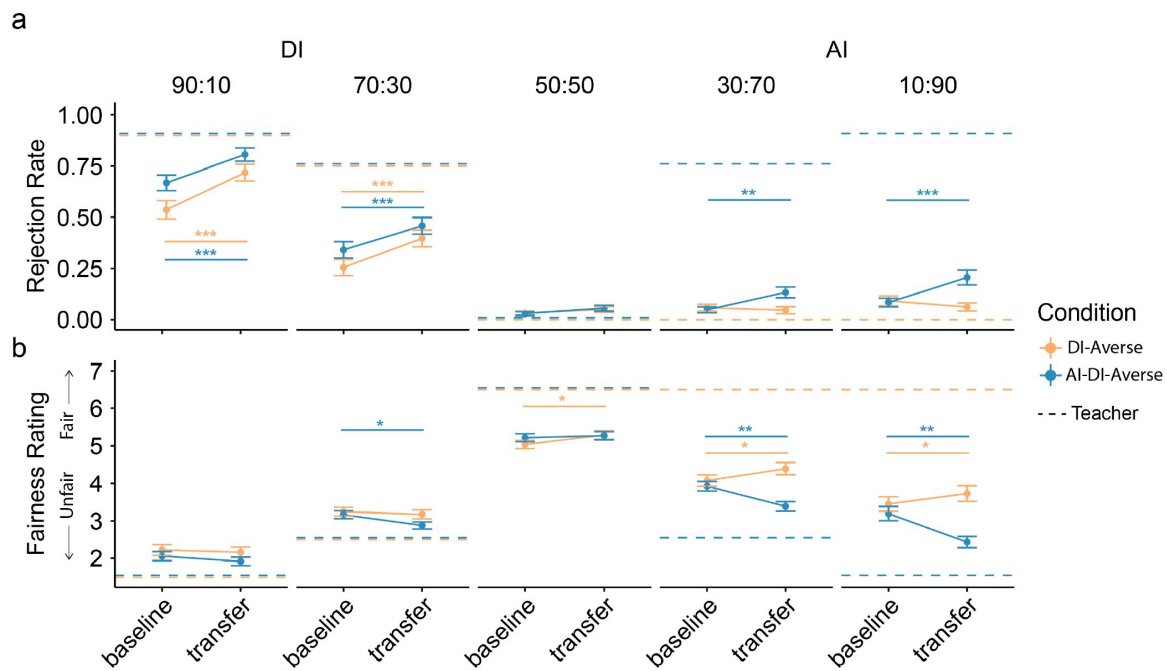


Figure 2.

Behavioral Contagion in Experiment 1

a) Rejection rates change significantly in DI offers for all conditions, while changes in AI offers were only evident in AI-DI-Averse Condition. **b)** Observing the Teacher's ratings of AI offers changed fairness ratings in all offer types, while the Teacher's behaviors in DI offers didn't. Dashed lines indicate behaviors of the Teacher. Error bars indicate standard error. (†indicates $p < 0.1$, *indicates $p < 0.05$, **indicates $p < 0.01$, ***indicates $p < 0.001$)

averse preferences, but not in DI-Averse condition (30:70 splits: $p=0.111$; 10:90 splits: $p=0.429$), where the Teacher did not. In other words, participants by and large appeared to adjust their rejection choices and fairness ratings in accordance with the Teacher's feedback in an incremental fashion.

Computational Models of Learning Punishment Preferences

Having established that rejection rates increased for both DI offers (in both the DI-Averse and AI-DI-Averse conditions) and AI offers (only in the AI-DI-Averse condition), we then sought to better understand the learning mechanisms underpinning trial-by-trial learning ([Figure 3a](#)). We used computational modelling to formalize two different sets of assumptions about how participants learn from Teachers' feedback. Under one account, a simple Reinforcement Learning (RL) model proposes that decision-makers learn punishment preferences by observing feedback resulting from actions made in response to specific offers. Previously we have found that this 'naïve' model provided a reasonable characterization of participants' trial-by-trial learning of DI-averse preferences ([FeldmanHall et al., 2018](#)). However, this model may fail to capture a critical facet of learning: participants' moral preferences may not be learned merely as associations—the type of response being tied to specific offers—but rather, through a deeper inference process which models the underlying fairness preferences of the Teacher. Accordingly, our alternative model assumes that the participant uses trial-by-trial feedback to infer the Teacher's underlying preferences concerning inequality, which may shift depending on the context (DI versus AI). Following [Fehr-Schmidt's \(1999\)](#) simple inequality aversion formalism, in our model—termed the Preference Inference model—the Teacher's aversion to DI is modeled by an 'Envy' parameter, while the 'Guilt' parameter captures the Teacher's aversion to AI (see Methods for model details). Critically, the RL model does not learn the Teacher's preferences per se, but the value of each action (accept or reject), independently for each offer type. In contrast, the Preference Inference model explicitly represents the extent of the Teacher's AI- and DI-aversion—i.e., their underlying preferences—by independently updating the envy and guilt parameters using trial-by-trial feedback from the Teacher.

We compared the goodness of fit of six different models to participants' choices in the Learning Phase: the Preference Inference model which learns the 'guilt' and 'envy' parameters experientially, a Static Preference model which assumes the 'guilt' and 'envy' are fixed over the course of learning (baseline model 1, see Methods), three variants of the RL model that make different assumptions about how action values are represented, and a baseline or 'null' model which assumed a fixed probability of each action (randomly choosing, baseline model 2). We fit each of the six models via maximum likelihood estimation, penalizing for model complexity (see Methods), and found that the Preference Inference model provided the best characterization of learning ([Figure 3b](#), [Table S8](#)), suggesting that participants were performing trial-by-trial inference of the Teacher's underlying inequity preferences, rather than simply learning reinforced associations between experienced offer types and actions. Even the Static Preference model, which does not assume any learning mechanism but rather, assumes fixed preferences with respect to DI and AI, provided a better characterization of learning than any of the three RL models which do assume incremental learning over paired associations.

To examine the learning dynamics underpinning the best-fitting Preference Inference Model, we simulated Learning Phase choice behavior using participants' estimated parameter values ([Figure 3a](#), see Methods for details). The close correspondence between the simulated and observed learning curves indicates that the Preference Inference model captures the reinforcement-guided variations in punishment in DI-context and, crucially, the marked differences in learning between DI-Averse and AI-DI-Averse conditions in AI-context (30:70 and 10:90 splits). To better understand how the Preference Inference model accounts for these patterns of change in rejection rates, we examined how model-inferred aversion to DI ('envy') and AI ('guilt')—the two components of the Fehr-Schmidt inequality aversion model ([Fehr & Schmidt, 1999](#)) representing the latent structure of the Teachers' preferences—emerge as a function of exposure to the Teacher's

preferences. In both the DI-Averse and AI-DI-Averse conditions, the model inferred a similarly “envious” Teacher (**Figure 3c**), while the model only increased its estimate of the Teacher’s ‘guilt’ parameter (**Figure 3d**), in the AI-DI-Averse condition where the Teachers’ feedback exhibited AI aversion, mirroring the model’s—and participants’—shift in rejection rates over the course of the Learning Phase and further suggesting that the Preference Inference model captured critical aspect of the participants’ vicarious learning behaviors.

Relationship between Learning and Contagion Effects

We next investigated whether the extent of a participants’ vicarious learning was associated with greater contagion of punishment preferences. To do this, we examined whether participants’ changes in rejection rates between Transfer and Baseline, could be explained by the degree to which they vicariously learned, defined as the change in punishment rates between the first and last 5 trials of the Learning phase. Across offer types (**Figure 4**), we observed that participants who more strongly adapted their rejection rates to match the Teacher’s preferences in the Learning phase were more likely to adopt the Teachers’ preference while making their own choices in the Transfer phase. Specifically, for DI offers, these predictive relationships were strong in the AI-DI-Averse (90:10 splits: $\beta(SE)=0.22(0.12)$, $p=0.073$; 70:30 splits: $\beta(SE)=0.33(0.08)$, $p<0.001$; **Table S9**) and DI-Averse (90:10 splits: $\beta(SE)=0.48(0.09)$, $p<0.001$; 70:30 offers: $\beta(SE)=0.36(0.08)$, $p<0.001$; **Table S9**) conditions. However, for AI offers, we only observed a relationship in the AI-DI-Averse Condition (30:70 splits: $\beta(SE)=0.26(0.10)$, $p<0.010$; 10:90 splits: $\beta(SE)=0.27(0.07)$, $p<0.001$; **Table S9**) but not the DI-Averse condition (a trend in 30:70 splits: $\beta(SE)=0.25(0.13)$, $p=0.058$; 10:90 splits: $\beta(SE)=0.12(0.10)$, $p=0.196$; **Table S9**). As expected, we did not observe any predictive relationships between learning and transfer behavior in response to fair (50:50) offers in either condition (AI-DI-Averse: $\beta(SE)=0.10(0.20)$, $p=0.625$; DI-Averse: $\beta(SE)=0.12(0.18)$, $p=0.503$; **Table S9**). Taken together, the relationships between the degree of successful learning of the Teachers’ preferences and the magnitude of change in punishment rates (when participants acted on their own behalf) strongly suggests that participants’ changes in punitive preferences—particularly for AI offers—occurred as a result of exposure to another’s preferences.

Conceptually replicating our previous findings (FeldmanHall et al., 2018), Experiment 1 provides evidence for behavioral contagion of DI-averse preferences, but extends this work by revealing that AI-averse preferences—which are believed to be less mutable (Luo et al., 2018)—can be similarly shaped by exposure to another individual with manifesting a strong aversion to resource divisions that unfairly benefit them. These results suggest that individuals’ moral preferences can be learned even in cases where these preferences conflict with one’s own self-interest.

Computationally, this learning process was best characterized by an account that prescribes that individuals build a representation of others’ moral preferences about DI and AI (akin to the Fehr-Schmidt (1999) model of inequity aversion), rather than by a simple Reinforcement Learning account. This preference inference account predicts that individuals, when exposed to only a fraction of inequity-related punishment preferences, should generalize these inferred preferences to other similar inequity contexts. In Experiment 2, we sought to test this generalization hypothesis more directly, buttressing the idea that the learning and transfer of Inequity-averse preferences observed in Experiment 1 came about as a result of participants modeling the Teachers’ inequity-averse preferences.

Experiment 2

Experiment 2 provides a more stringent test of whether participants model the Teacher’s underlying inequity preferences. Specifically, if individual indeed learn the Teacher’s latent inequity-averse preferences in the Learning phase, we would expect that feedback-driven learning of Teacher’s punishment preferences on specific (moderate inequity) offers (30:70 splits) should

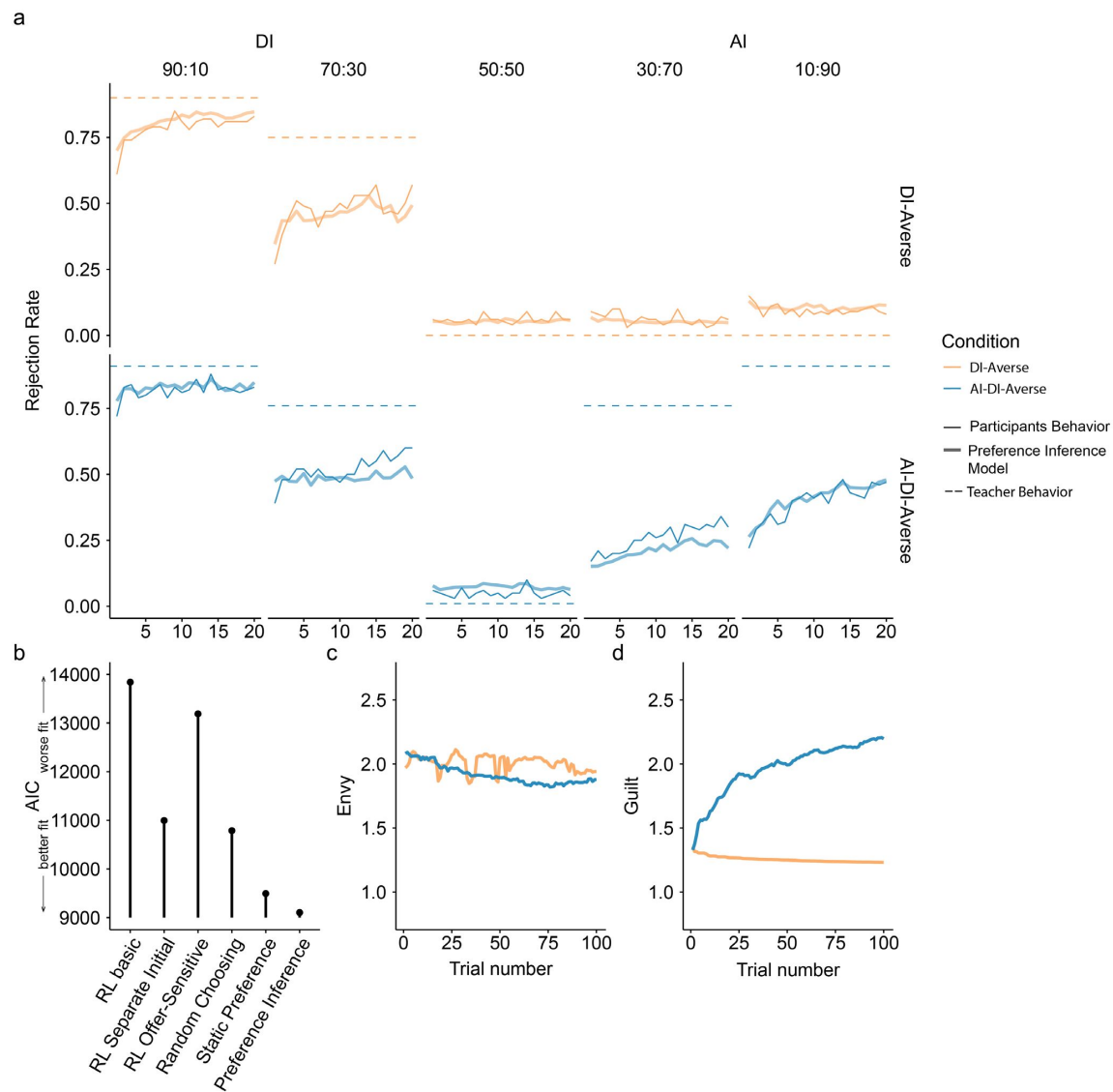


Figure 3.

Learning phase behavior in Experiment 1.

a) Rejection rate changes in Learning Phase. Rejection choices were summarized across participants. For DI Offers, rejection rate increased in both Conditions during learning. While rejection rate only changed (increase) in AI-DI-Averse Condition for AI offers. Solid thin lines denote participants' rejection choices, dashed lines denote the Teacher's preferences, and solid thick lines represent predictions of the (bestfitting) Preference Inference Model. **b)** Model comparison, demonstrating that the Preference Inference model provided the best fit to participant' Learning Phase behavior (AIC: Akaike Information Criterion) **c)** and **d)** Parameters updating for the Preference Inference model. The Preference Inference model captured significant rejection rate increasing in AI offers by updating the guilt parameter in a trial by trial manner.

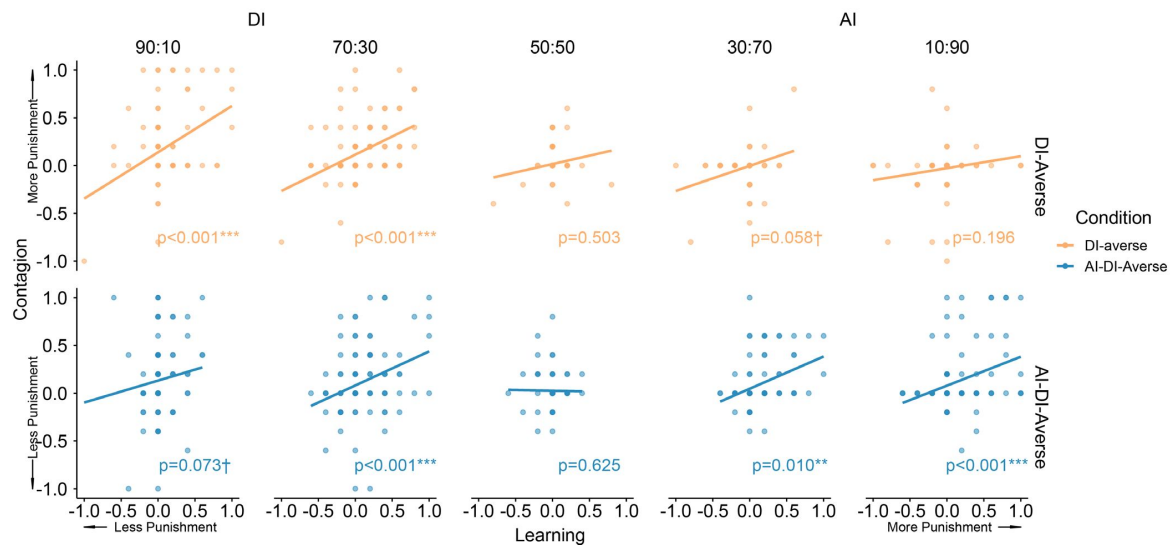


Figure 4.

Relationship between learning and contagion effects in Experiment 1.

Rejection rate changes in Learning phase was indexed by the averaged rejection rate difference between first five and the last five trials in Learning phase. On the DI side, the learning index can predict contagion in both AI-DI-Averse and DI-Averse conditions. while on the AI side, this effect is more salient in the AI-DI-Averse Condition than that in the DI-Averse Condition (†indicates $p < 0.1$, *indicates $p < 0.05$, **indicates $p < 0.01$, ***indicates $p < 0.001$)

generalize to offers in the same context (10:90 splits) without any direct experience of those offer types. Accordingly, to probe for this sort of generalization—a hallmark of the sort of latent structure learning we attribute to the behavior we observed in Experiment 1—we now eliminate feedback for extreme DI (90:10) and AI (10:90) offers from the Learning phase. If participants' punishment preferences are informed by modeling inferred inequity-averse preferences of the Teacher, we should expect to see these preferences transfer to participants' own fairness preferences in a similar, generalized manner in the Transfer Phase.

Mirroring Experiment 1, we employed two conditions governing the Teachers' preferences in the Learning Phase: the Teacher in the DI-Averse condition only exhibited strong punishment preferences concerning moderate DI offers (30:70 splits), while the Teacher in the AI-DI-Averse condition exhibited strong preferences in response to both moderate DI and moderate AI offers (30:70 and 70:30 splits).

Contagion Effects for Extreme Unfair Offers Suggest Generalization

In Experiment 2, we took the same analysis approach as in Experiment 1, examining changes in rejection rates between the Baseline Phase and the Transfer Phase (after participants experienced feedback with moderately unfair offers). Similar to what we observed in Experiment 1 (**Figure 5a** [↗](#)), participants increased their rates of rejection of extreme DI (i.e., 90:10) in the Transfer Phase, relative to the Baseline phase (AI-DI-Averse Condition: β (SE)=0.13(0.03), $p < 0.001$; DI-Averse Condition: β (SE)=0.18(0.03), $p < 0.001$; **Table S10** [↗](#)), suggesting that participants' learned (and adopted) DI-averse preferences, generalized from one specific offer type (70:30) to an offer types for which they received no Teacher feedback (90:10). Examining generalization across AI offers, we found a trend that participants in the AI-DI-Averse Condition increased their rejection rates of extreme AI offers (10:90) for which they did not receive any Teacher feedback in the Learning Phase (β (SE)=0.06(0.03), $p=0.060$). Furthermore, we saw decreases in rejection rates of extreme AI offers for participants in the DI-Averse Condition (β (SE)=-0.06(0.03), $p=0.035$). Mirroring the observed rejection rates (**Figure 5b** [↗](#)), participant rated these (untrained) extreme unfair DI offers (90:10) as less fair in the Transfer phase in both the AI-DI-Averse (β (SE)=-0.23(0.13), $p=0.076$; see **Table S11** [↗](#)) and DI-Averse conditions (β (SE)=0.29(0.13), $p=0.023$), and rated extreme AI offers (10:90) as less fair only in AI-DI-Averse Condition (β (SE)=- 0.84(0.13), $p < 0.001$).

We further reasoned that if the contagion effects observed for extremely unfair offers (90:10 and 10:90 splits) resulted from the same latent structure learning process driving the observed contagion for moderately unfair offers (30:70 and 70:30 splits), we should expect the magnitude of contagion effects in moderately unfair offers to relate to the magnitude of contagion effects in extremely unfair offers. Examining changes in punishment rates in AI contexts—between the Transfer and Baseline phases—we observed that participants with larger contagion effects for 30:70 offers also exhibited larger contagion effects for 10:90 offers (**Figure S2** [↗](#)). This was observed both in rejection rate changes (AI-DI-Averse Condition: $p < 0.001$, DI-Averse Condition: $p < 0.001$) and perceived fairness changes (AI-DI-Averse Condition: $p < 0.001$, DI-Averse Condition: $p < 0.001$). In DI contexts, we found that the same relationships between responses to moderately unfair offers (70:30) and extremely unfair offers (90:10), both for changes in rejection rates (AI-DI-Averse Condition: $p < 0.001$; DI-Averse Condition: $p = 0.003$) and perceived fairness rating changes (AI-DI-Averse Condition: $p < 0.001$; DI-Averse Condition: $p < 0.001$).

In short, we find evidence that participants generalized across learning contexts, which in turn shaped their own punitive responses to extreme offers, both in the case of DI and AI offers. In other words, it appears that preferences acquired through contagion extends beyond mere associations between single offers and actions, and instead relies on a mechanism that infers the latent structure of the Teacher's fairness preferences.

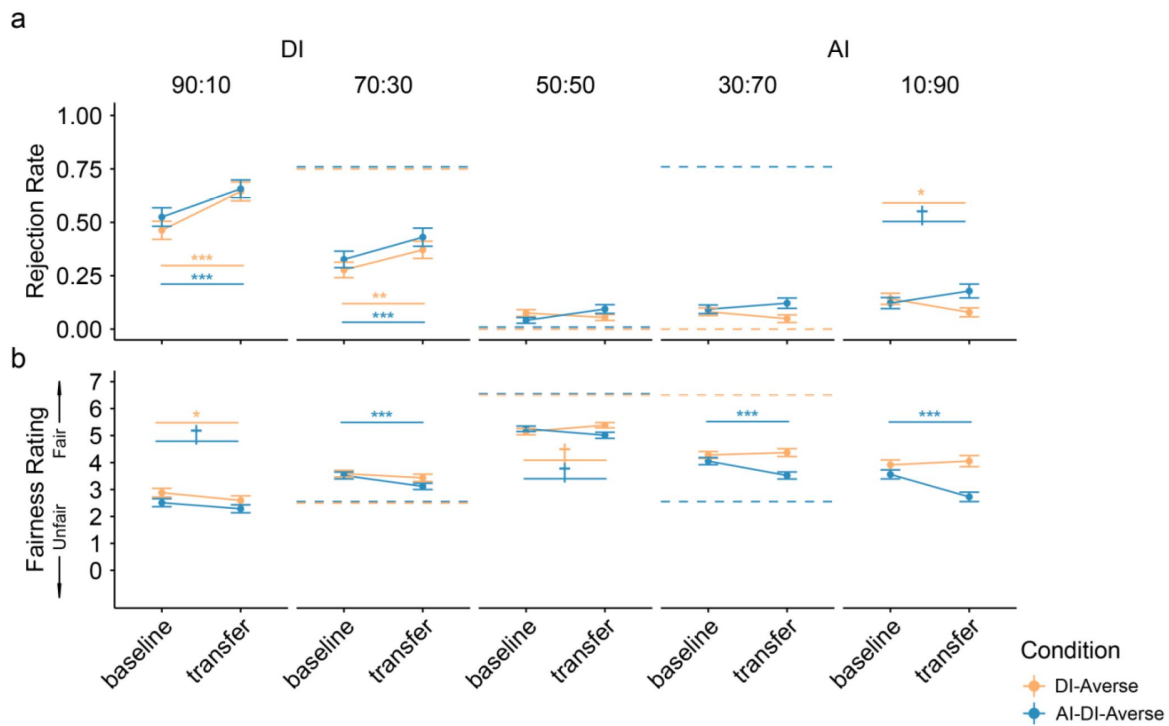


Figure 5

Baseline and Transfer Phase Behavior in Experiment 2.

a) Contagion in extremely unfair offers. Though no feedback was provided in the Learning phase for 90:10 or 10:90 splits, we observed generalization of punishment preferences in these types of offers. Dashed lines represent the Teacher's preferences. **b)** Fairness rating changes. We found significant changes from Baseline to Transfer phase in fairness rating for 90:10 in both AI-DI-Averse and DI-Averse Condition, but only in AI-DI-Averse Condition for 10:90 offers. Error bars represent standard errors (†indicates $p < 0.1$, *indicates $p < 0.05$, **indicates $p < 0.01$, ***indicates $p < 0.001$)

Preference Changes in the Learning Phase Suggest Generalization

A primary goal of Experiment 2 was to demonstrate that learning the Teacher's preferences with respect to moderately unfair offers generalized to extremely unfair offers, where no feedback from the Teacher was provided. Examining the time course of rejection rates in AI-contexts during the Learning phase ([Figure 6](#)) revealed that participants learned over time to punish mildly unfair 30:70 offers, and these punishment preferences generalized to more extreme offers (10:90). We observed a significant increase in rejection rates for 10:90 (AI) offers in the AI-DI-Averse Condition ([Figure 6](#), β (SE)=1.06(0.31), $p<0.001$; mixed-effects logistic regression, see [Table S12](#)), but not in the DI-Averse Condition (β (SE)=-0.54(0.40), $p=0.170$). We observed significant rejection rate increases for 90:10 (DI) offers in DI-Averse Condition (β (SE)=0.96(0.27), $p<0.001$) but not in the AI-DI-Averse Condition (β (SE)=0.36(0.26), $p=0.116$). Finally, we observed significant decreases in fairness ratings over time for 10:90 offers in the AI-DI-Averse Condition (β (SE)=-0.27(0.09), $p=0.004$; [Table S13](#)), again suggesting that participants generalized fairness ratings across offer types on the basis of inferred fairness preferences attributed to the Teacher. We did not observe this sort of generalization for fairness ratings in the DI-Averse condition, nor did we observe significant changes in fairness ratings over time in DI context ($ps > 0.131$ for all offers; see [Table S13](#)), perhaps owing to the smaller difference between the Teacher's preferences and participants' default (Baseline) preferences concerning DI offers. Furthermore, we reasoned that participants' learning in moderately unfair contexts (for which participants received feedback) should predict their learning in extremely unfair contexts (for which no feedback was provided). To test this, we examined individual differences in changes in punishment between the final 5 trials and the first 5 trials of the Learning phase, finding that learning for moderately unfair (30:70 and 70:30) offers predicted changes in punishment rates for extremely unfair (10:90 splits: $p < 0.001$ and 90:10 splits: $p < 0.001$) offers. In other words, participants who increased their punishment more in response to (reinforced) moderately unfair offers also exhibited changes in their preferences for (non-reinforced) rejection of extremely unfair offers.

Finally, following Experiment 1, we fit a series of computational models of Learning phase choice behavior, comparing the goodness-of-fit of the four best-fitting models from Experiment 1 (see Methods). As before, we found that the Preference Inference model best characterized participants' Learning Phase behavior ([Figure S1a](#), [Table S14](#)). This is unsurprising, given that this model—by virtue of learning the Teachers' underlying preferences— assumes the sort of generalization of preferences between offer types that Experiment 2 sought to test directly. We also simulated this model's Learning Phase behavior using participants' estimated parameter values and found that the model, mirroring participants' choices, exhibits clear incremental changes in rejection rates ([Figure 6](#); thick lines), both for offers where the model received explicit feedback (70:30) and for offers where the model received no feedback (90:10). In other words, like participants, the model generalizes the learned inequity-averse preferences to extreme AI offers (90:10), which stems from the model's trial-by-trial updating of the parameters governing the Teacher's preferences (see [Figures S1b and c](#)).

Discussion

While people tend to reject proposed resource allocations where they stand to receive less than their peers (so-called disadvantageous inequity; DI), they are markedly less averse to resource allocations where they stand to unfairly gain more than their peers (advantageous inequity or AI; [Blake et al., 2015](#); [Luo et al., 2018](#)). Here we considered the possibility that these complex, other-regarding preferences for fairness can be imparted merely by observing (and enacting) the preferences of another person. We investigated, in an Ultimatum Game setting, whether AI-averse

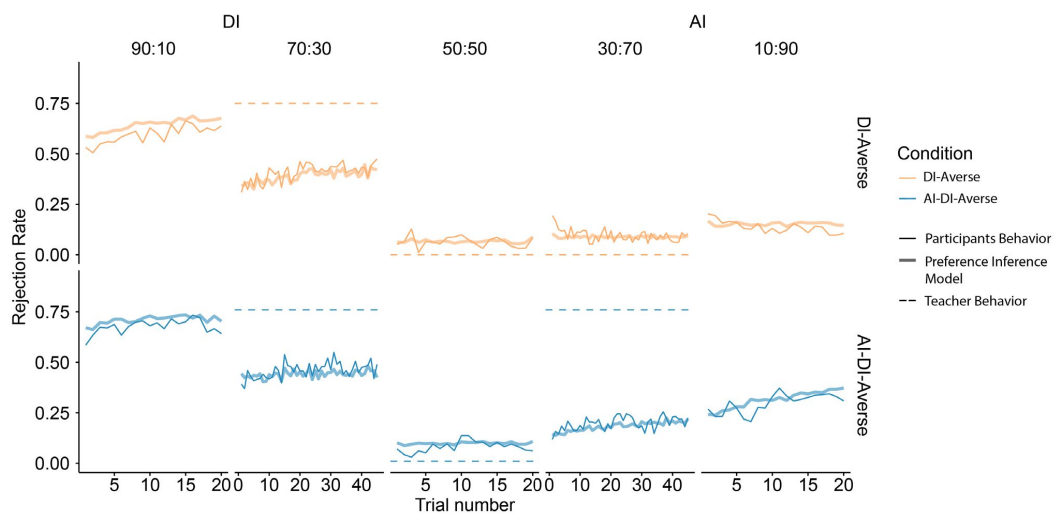


Figure 6.

Learning Phase Choice Behavior in Experiment 2.

Learning effect were documented in Extremely unfair offers. Rejection choices were summarized across subjects. Dashed lines indicate Rejection choice of the Teacher. The learning effect were evident for 90:10 offers in DI-Averse condition and 10:90 offers in AI-DI Averse condition. Thin solid lines represents participants' rejection choice, thick solid lines show the predictions of the Preference Inference Model, and the dashed lines indicate the Teacher's preferences (not observed by participants in 90:10 and 10:90 splits).

preferences can be shaped by learning and the preferences of another individual. We leveraged a well-characterized observational learning paradigm (FeldmanHall et al., 2018 [DOI](#)), exposing participants to another individual (the Teacher) exhibiting a strong preference for punishment of advantageously unfair offers, and probed whether these punishment preferences in turn transferred to participants making choices on their own. We found that participants' own AI-averse preferences shifted towards the preferences of the Teacher they just observed, and the strength of these contagion effects related to the degree of behavior change participants exhibited on behalf of the Teachers, suggesting that they internalized, at least somewhat, these inequity preferences.

While we observed apparent adoption of the Teachers' inequity-averse preferences, in the face of both AI and DI, previous work has outlined a number of important differences with respect to how individuals respond to the two sorts of inequity. Aversion to DI is thought to arise from negative emotions such as spite (McAuliffe et al., 2014 [DOI](#); Pillutla & Murnighan, 1996 [DOI](#)) engendered by consideration of one's standing relative to others, while AI aversion, in contrast, is thought to stem from concerns about fairness or inequality (McAuliffe et al., 2013 [DOI](#)). Hence, the expression of AI aversion may signal, and even enforce, egalitarian social norms. Developmental evidence supports this distinction between AI versus DI aversion. While DI aversion emerges at the tender age of 4, AI aversion does not appear to manifest until about 8 years (Blake et al., 2015 [DOI](#); Blake & McAuliffe, 2011 [DOI](#); McAuliffe et al., 2017 [DOI](#)). In fact, AI aversion—which entails trading off self-interest against a social norm enforcement—is not even commonly observed in adults (Blake et al., 2015 [DOI](#); Hennig-Schmidt et al., 2008 [DOI](#); Luo et al., 2018 [DOI](#)), suggesting that AI aversion is more difficult (and less likely) to be learned than DI aversion. That we found evidence that AI averse-preferences can be learned suggests that observational learning processes are a potent and promising means by which sophisticated fairness preferences can be imparted.

Mechanistically, we found that participants' feedback-based learning of punishment preferences was best characterized by a computational model that assumes individuals infer the Teacher's latent and structured preferences for punishment—rather than a simple Reinforcement Learning (RL) account assuming that individuals learn contextually-bound punishment preferences. To support the interpretation that individuals indeed 'model' the fairness preferences of others, in a second experiment we directly test whether participants can infer the Teachers' inequity-averse preferences across contexts. We found that participants generalized the Teacher's punitive preferences to other contexts that varied in their unfairness, and this occurred in the both AI and DI contexts, suggesting that the discovery of latent structure is instrumental for generalization.

The representation of others' beliefs in other interpersonal decision-making tasks has been previously formalized, computationally, by a Bayesian account of theory of mind (ToM) in which the hypothetical beliefs were described by a prior distribution, and participants update this distribution using Bayesian updating (Baker, Chris et al., 2009 [DOI](#); Jara-ettinger, 2019 [DOI](#)). This computational framework has been applied to describing behaviors in a group decision making task (Khalvati et al., 2019 [DOI](#)). In our Preference-inference model, on each trial, the learner makes a guess about the Teacher's inequity aversion parameters, then the Teacher's feedback is subsequently used to further constrain the range of parameter values which could conceivably produce the Teacher's observed feedback. The learner then updates their initial guess range. Conceptually, this learning mechanism is consistent with the sort of Bayesian updating.

One open question concerning the contagion effects is how the identities—and number—of teachers experienced in the Learning phase bear on the strength of the learning and contagion effects observed. For simplicity, our Learning Phase employed only one distinct Teacher, that included no identity information, which contrasts with many social interactions in daily life, which are almost always accompanied by identifying information or attributes concerning the other (Hester & Gray, 2020 [DOI](#)). In contrast, our interactions with others are profoundly influenced by the identities of others—for example, whether they are conservative versus liberal (Leong et al.,

2020 [↗](#)), whether they are inversus out-group members (Hein et al., 2010 [↗](#); Vives et al., 2022 [↗](#)). At the same time, the strength of social influence often increases with the number (or proportion) of individuals in a group expressing a particular preference (Cialdini & Goldstein, 2004 [↗](#); Son et al., 2019 [↗](#)). However, it may also be the case that social contagion effects require repeated interactions with the same individual (Tsvetkova & Macy, 2014 [↗](#)), which the contagion observed in the present paradigm corroborates. Accordingly, future work should aim to examine the influence of the teacher's identity—and its concordance with the learner's identity—as well as seek to understand the how the relative balance of repeated experience with identical teachers versus the number of distinct teachers modulates the strength of contagion effects in AI/DI punishment preferences.

In summary, our study provides an initial demonstration that despite the desire for self-gain, we observe that people can swiftly and readily acquire another's preferences for advantageous inequity, even when it comes at a monetary cost to the self. Computationally, we find that this contagion of inequity averse-preferences occurs through representing the underlying structure of another's preferences, rather than a Reinforcement Learning-like process of learning simple context-action associations. Importantly, these inferred preferences are sufficient to induce individuals to change their preferences for punishing advantageous inequity, suggesting that social influence may be one promising route through which social norm enforcement—which is particularly uncommon in the case of AI—can be promoted.

Methods

Participants

We recruited US-based participants from Amazon Mechanical Turk (Crump et al., 2013 [↗](#)) for both Experiment 1 ($N=200$, M age = 37.53 ($SD=10.88$), 75 females) and Experiment 2 ($N=200$, M age=37.16 ($SD=11.79$) years old, 80 females). These sample sizes are based on our previous work employing the same per-condition sample sizes (FeldmanHall et al., 2018 [↗](#)) in a punishment-learning task. Participants provided informed consent in accordance with the McGill University Research Ethics Board. Participants were randomly assigned to either the AI-DI-Averse ($N = 100$) or the DI-Averse ($N = 100$) condition. As we replicated all key reported results when excluding participants who evidenced some disbelief that the Teacher (see below for procedural details), suggesting that these results are robust to potential disbelief about the task structure as described to them. In Experiment 2, we excluded the data of participants who failed to meet the requirements of each analysis due to missed trials (3 in AI-DI-Averse and 3 in DI-Averse Condition were removed in Contagion effects analysis; 2 in AI-DI-Averse , 5 in DI-Averse Condition were removed for the generalization analysis).

Punishment Learning Paradigm

We used a modified version of the Ultimatum Game (Güth et al., 1982 [↗](#)) to probe participants' fairness preference. In this task, the proposer, offer an allocation of a total amount (e.g. \$1 out of \$10). Then, another player, the Receiver, chooses to reject or accept the proposed allocation. If the Receiver accepts, both players receive the proposed amount, if they reject, both player receive nothing. Following our past work examining contagion effects (FeldmanHall et al., 2018 [↗](#)), our task consisted of 3 phases: Baseline, Learning and Transfer (see **Figure 1** [↗](#)).

In the Baseline and Transfer phases, participants responded to unfair offers as a receiver in multiple rounds of ultimatum game, and participants were informed that the Proposer on each trial is a different, randomly chosen individual with a unique (fictitious) participant ID (**Figure 1c** [↗](#)), which permitted us to measure participants' preferences and beliefs about fairness irrespective of any particular Proposer. We considered 5 unique offer levels (90:10, 70:30, 50:50,

30:70, 10:90) ranging from extreme DI to extreme AI, including ‘fair’ offers (50:50), which allowed us to measure participants’ baseline rejection tendency. We instructed participants that these Proposers were concurrently participating in this task and these were “real time” offers. Critically, participants were responding to offers made by fictitious players with predetermined offers to ensure that we could observe rejection preferences for each offer level. On 3/5 of Baseline and Transfer Phase trials, participants also rated the fairness of the offer on a 1-7 scale (1 being ‘Strongly unfair’; 7 being ‘Strongly fair’) after making an accept/reject choice. Participants experienced 5 trials of each offer type in both the Baseline and Transfer Phase.

In the Learning phase, participants played an Ultimatum Game, acting as a third party deciding to accept or reject Proposers’ offers for another receiver (the Teacher) receiving offers from Proposers (see [Figure 1d](#)). After the participant chose to accept or reject the offer, the option the Teacher would have preferred was revealed. In the AI-DI-Averse condition, the Teacher indicated preference for rejection of fair (50:50) and AI offers, and uniformly rated these AI offers as unfair (see [Figure 1b](#)). In the DI-Averse condition, the Teacher accepted all fair and AI offers and rated these offers as uniformly fair. In both the DI-Averse and AI-DI-Averse, DI offers were rejected and rated as unfair. The Teachers’ rejection rates and ratings in response to each offer type are provided in [Tables S1](#) and [S2](#).

Again, on 3/5 of trials, participants rated the fairness of each offer, and on these trials participants also saw the Teachers’ fairness rating of the offer. Like the Baseline and Transfer phases, participants saw a unique, randomly chosen Proposer on every trial, but were informed that there was a single, unvarying Teacher over the entire Learning Phase. In Experiment 1, participants experienced 20 trials for each of the 5 offer types considered (90:10, 70:30, 50:50, 30:70, 10:90). To generate each offer, we further added a uniformly distributed noise in the range of (-10, 10) to each offer type.

Experiment 2 followed the same procedure as Experiment 1 except for the following changes. First, and most importantly, in the Learning Phase participants did not experience feedback on extreme AI offers (10:90 splits). Second, we added to the Learning Phase 25 more trials each for offer type (70:30 and 30:70 splits), to afford greater opportunity to observe the Teacher’s preferences. Third, participants were instructed that Proposers’ offers (which were predetermined) were generated by previous participants in previous similar experiment. Finally, in all task phases, participants were required to respond within a 3-second deadline when making choices, and a 4-second deadline when providing fairness ratings.

Data Analysis

We used mixed-effects regression, implemented in the ‘lmer’ package for R ([Bates et al., 2015](#)) to estimate the effect of offer level and punishment condition upon contagion rates—the rejection rate change between the Baseline and Transfer phases. To do this the rejection rate change were modeled by interactions between Offer types (factors of five levels: 90:10, 70:30, 50:50, 30:70, 10:90, encoded as 5 columns) and Condition (AI-DI-Averse vs. DI-Averse, encoded as 2 columns), with random intercepts taken over participants (see [Table S7](#) for full coefficient estimates).

We estimated changes in punishment rates in the Learning Phase using mixed-effects logistic regressions in Learning. Specifically, rejection choices were predicted as a function of trial number, offer type, condition, and their resultant interactions, taking the 2-way interactions between Trial number and offer type (random slope) as random effects over participants. To estimate changes in fairness ratings, we estimated a linear mixed-effects model with the same terms.

Computational Models of Learning Phase Behavior

We considered 6 computational models of Learning Phase choice behavior, which we fit to individual participants' observed sequences of choices and Teacher feedback via Maximum Likelihood Estimation. Importantly, Models 1 (Random choosing) and 2 (Static preference) are baseline models which assume that rejection probabilities are fixed over time, while all other models allow for learning of rejection rates over time in accordance with Teacher feedback.

Model 1 (Random Choosing).

This model assumes that participants reject offers with a fixed probability governed by the parameter $p_{offertype}$ (one for each offer type; 5 free parameters in total).

$$P_{reject}(offertype) \sim \text{Bernoulli}(p_{offertype}) \quad (1)$$

Model 2 (Static preference):

This model that when choosing for others, the utility of accepting an offer, relative to rejection of the offer, is governed by the Fehr-Schmidt (FS) inequity aversion (Fehr & Schmidt, 1999 [↗](#); Luo et al., 2018 [↗](#)) function:

$$U_{accept}(offer) = offer - \alpha * \max(50 - offer, 0) - \beta * \max(offer - 50, 0) \quad (2)$$

$$U_{reject}(offer) = 0 \quad (3)$$

In this function, *offer* represents the share the Proposer give to the Receiver, α parameterizes the Teacher's disutility for accepting disadvantageous unfair offers (DI aversion), and β captures Teacher's disutility for accepting advantageous offers (AI aversion), which can each range from 0-1.

These action utilities were then transformed to choice probabilities using the softmax choice rule:

$$P_{reject}(offer) = \exp(\tau * U_{reject}) / (\exp(\tau * U_{reject}) + \exp(\tau * U_{accept})) \quad (4)$$

where the inverse temperature (τ) parameter captures decision noise, where a larger τ corresponds to a higher probability of choosing the action with higher utility, and as τ approaches 0, the two options are chosen with equal probability. In total, this model has 3 free parameters.

Model 3 (Basic RL):

Model 3 is a simple RL model following that used by FeldmanHall et al. (2018) [↗](#) only one learning rate, which represents and updates values of the two actions separately for each offer type, using a delta updating rule:

$$Q_{t+1}(action_t, offertype_t) = Q_t(action_t, offertype_t) + \eta * (R_t - Q_t(action_t, offertype_t)) \quad (5)$$

where $action_t$ is the action the participant chose (accept or reject) on the t -th trial, R_t is the reward on the t -th trial, which is defined as 1 (reward obtained) when the action taken was the same as the action the Teacher would have preferred, and 0 (no reward) otherwise. The softmax choice rule was used to translate these action values to predicted choice probabilities. This model has 2 free parameters.

Model 4 (Offer-Sensitive RL):

This RL model is a more complex variant of Model 3, and assumes a separate learning rate for each offer type (as in Model 3). In total, this model has 6 free parameters.

Model 5 (Offer-Sensitive RL with Separate Initial Values):

This RL model extends Model 3, assuming different initial action values for each offer type. Formally, this model treats $Q_0(\text{reject}, \text{offertype})$, ($\text{offertype} \in 90:10, 70: 30, 50: 50, 30: 70, 10: 90$) as free parameters with values between 0 and 1., resulting in 7 free parameters.

Model 6 (Preference Inference):

Model 6 posits that the participant infers the Fehr-Schmidt utility function (Equation 2) governing the Teacher's preferences, and updates their modeled 'guilt' (α) and 'envy' (β) parameters incrementally from feedback, under the assumption that the Teacher's indicated choices are made in accordance with each offer's Fehr-Schmidt utility (more formally, the Teacher rejects the offer when $U_{\text{accept}}(\text{offer}) < 0$). As α and β govern the disutility of unfair offers, the model infers the minimal value of α (or β) would lead to rejection of DI (or AI) offers, and similarly, the maximum values of α (or β) would lead to acceptance of DI (or AI) offers.

Accordingly, after observing that the Teacher prefers rejection in response to a DI offer, Equation 2 can be transformed to the following inequality:

$$U_{\text{accept}}(\text{offer}) = \text{offer} - \alpha * (50 - \text{offer}) < 0 \quad (6)$$

where the model can infer a lower bound of α , that would lead to the offer's rejection by solving (6):

$$\alpha > \frac{\text{offer}}{50 - \text{offer}} \quad (7)$$

The right side of (7) can be denoted as α_{lb} , and then that trial's estimate of α (denoted α_t) is updated as follows:

$$\alpha_{t+1} = \begin{cases} \alpha_t, & \text{if } \alpha_t > \alpha_{lb} \\ \alpha_t + \eta * (\alpha_{lb} - \alpha_t), & \text{otherwise} \end{cases} \quad (8)$$

The parameter η governs the rate at which the learner's estimate of the Teacher's α value is updated, and is constrained to the range [0, 5].

The updating procedure is similar when Teacher indicates acceptance of a DI offer, which implies that $U_{\text{accept}}(\text{offer}) > 0$, and the following inequality:

$$U_{\text{accept}}(\text{offer}) = \text{offer} - \alpha * (50 - \text{offer}) > 0 \quad (9)$$

which yields an upper bound for the envy parameter α

$$\alpha < \frac{\text{offer}}{50 - \text{offer}} \quad (10)$$

This upper bound, α_{ub} , is in turn used to update α_t :

$$\alpha_{t+1} = \begin{cases} \alpha_t, & \text{if } \alpha_t < \alpha_{lb} \\ \alpha_t + \eta * (\alpha_{ub} - \alpha_t), & \text{otherwise} \end{cases} \quad (10)$$

In the case of AI offers, the model employs the identical procedure to update the ‘guilt’ parameter β , and α is updated only in DI offers, while β is only updated in AI offers. Following [Luo et al. \(2018\)](#), α and β were restricted to the range of [0, 10]. The initial value of α and β are taken as free parameters in the range of [0, 10], resulting in a model with a total of 4 free parameters.

All models were fit via Maximum Likelihood Estimation, employing a nonlinear optimization procedure using 100 random start points in the parameter space in order to find the best-fitting parameter values for each participant. We then computed the Akaike Information Criterion (AIC; [Akaike, 1974](#)) to select the best-fitting models of Learning phase choice behavior, penalizing each model’s goodness-fit-score by its complexity (i.e., number of free parameters). See [Table S8](#) and [S14](#) for parameter estimates and goodness-of-fit metrics.

Acknowledgements

This work was funded by the European Union (ERC Starting Grant, NEUROGROUP, 101041799) and by an NSERC Discovery Grant, a New Researchers Startup Grant from the Fonds de Recherche du Québec - Nature et Technologies, and an infrastructure award from the Canadian Foundation for Innovation. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Supplementary Information

Figure S1.

Model comparison results in Experiment 2. **a)** AICs of the models considered in experiment 2. **b),c).** Updating of the ‘guilt’ and ‘envy’ parameters indicates the sanity of the Preference Inference model.

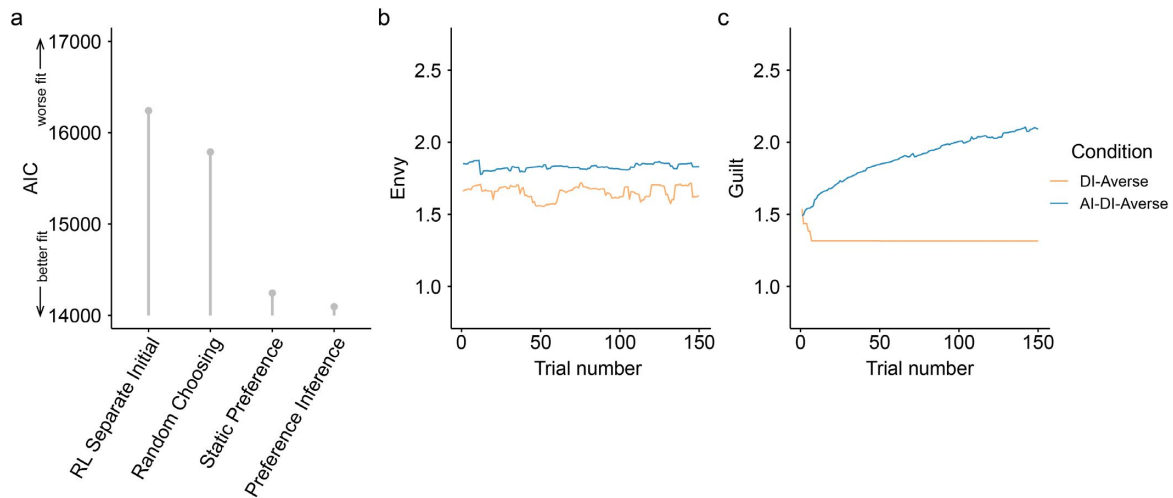


Figure S2.

In Experiment 2, Rejection rates changes in the learning phase (indexed by the rejection rate difference between the first and the last five trials in each offer type) predict the rejection rate change from baseline to transfer phase (the contagion).

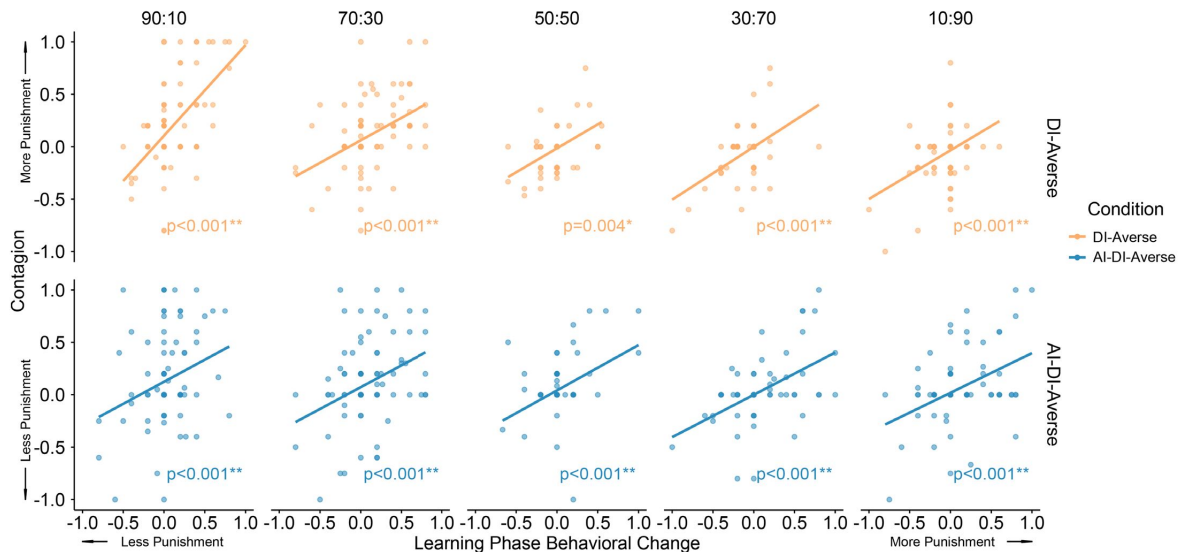


Table S1.

Reinforcement rates governing the Teacher's feedback in the Learning Phase. For example, a rate 90% for 90:10 offers indicates that on 90% of trials with that offer type, the Teacher indicated they would have preferred rejection of that offer.

Offer type	AI-DI-Averse	DI-Averse
90:10	90%	90%
70:30	75%	75%
50:50	0%	0%
30:70	75%	0%
10:90	90%	0%

Table S2.

Fairness rating of the Teacher in the Learning Phase in Experiment 1. The teacher rated 90:10 offers as Strongly unfair (1) or Unfair (2) randomly in both conditions.

Offer type	AI-DI-Averse	DI-Averse
90:10	1 or 2	1 or 2
70:30	2 or 3	2 or 3
50:50	6 or 7	6 or 7
30:70	2 or 3	6 or 7
10:90	1 or 2	6 or 7

Table S3.

Baseline Phase Choice Behavior in Experiment 1. Linear mixed model coefficients (fixed effects) indicating effects of condition and Offer type on the baseline Rejection rate.

Predictor	Estimate	SE	t	p
AI-DI-Averse×Offer 90:10	0.67	0.03	23.29	<0.001
AI-DI-Averse×Offer 70:30	0.34	0.03	11.85	<0.001
AI-DI-Averse×Offer 50:50	0.03	0.03	0.98	0.329
AI-DI-Averse×Offer 30:70	0.05	0.03	1.60	0.109
AI-DI-Averse×Offer 10:90	0.08	0.03	2.86	0.004
DI-Averse×Offer 90:10	0.54	0.03	18.69	<0.001
DI-Averse×Offer 70:30	0.25	0.03	8.85	<0.001
DI-Averse×Offer 50:50	0.03	0.03	1.05	0.296
DI-Averse×Offer 30:70	0.06	0.03	1.95	0.051
DI-Averse×Offer 10:90	0.09	0.03	3.14	0.002

Table S4.

Baseline Phase Rating Behavior in Experiment 1. Linear mixed model coefficients (fixed effects) indicating effects of condition and Offer type on the baseline Fairness Ratings (testing if the coefficient is equal to 4).

Predictor	Estimate	SE	t	p
AI-DI-Averse×Offer 90:10	2.06	0.14	-13.92	<0.001
AI-DI-Averse×Offer 70:30	3.16	0.14	-5.99	<0.001
AI-DI-Averse×Offer 50:50	5.22	0.14	8.71	<0.001
AI-DI-Averse×Offer 30:70	3.92	0.14	-0.55	0.583
AI-DI-Averse×Offer 10:90	3.19	0.14	-5.78	<0.001
DI-Averse×Offer 90:10	2.22	0.14	-12.75	<0.001
DI-Averse×Offer 70:30	3.24	0.14	-5.42	<0.001
DI-Averse×Offer 50:50	5.04	0.14	7.42	<0.001
DI-Averse×Offer 30:70	4.08	0.14	0.55	0.583
DI-Averse×Offer 10:90	3.45	0.14	-3.96	<0.001

Table S5.

Contagion effects in rejection rates in Experiment 1, computed as the difference between Transfer and Learning Phase rejection rates. Linear mixed model coefficients (fixed effects) indicating the effects of Condition and Offer Type on the Rejection rate changes from Baseline to Transfer phase.

Predictor	Estimate	SE	t	p
AI-DI-Averse×Offer 90:10	0.14	0.03	4.97	<0.001
AI-DI-Averse×Offer 70:30	0.12	0.03	4.25	<0.001
AI-DI-Averse×Offer 50:50	0.03	0.03	0.94	0.349
AI-DI-Averse×Offer 30:70	0.09	0.03	3.10	0.002
AI-DI-Averse×Offer 10:90	0.12	0.03	4.47	<0.001
DI-Averse×Offer 90:10	0.18	0.03	6.56	<0.001
DI-Averse×Offer 70:30	0.14	0.03	5.12	<0.001
DI-Averse×Offer 50:50	0.02	0.03	0.65	0.517
DI-Averse×Offer 30:70	-0.01	0.03	-0.43	0.666
DI-Averse×Offer 10:90	-0.03	0.03	-1.08	0.280

Predictor	Estimate	SE	t	p
AI-DI-Averse×Offer 90:10	-0.14	0.12	-1.15	0.252
AI-DI-Averse×Offer 70:30	-0.28	0.12	-2.32	0.021
AI-DI-Averse×Offer 50:50	0.05	0.12	0.44	0.662
AI-DI-Averse×Offer 30:70	-0.54	0.12	-4.42	<0.001
AI-DI-Averse×Offer 10:90	-0.76	0.12	-6.23	<0.001
DI-Averse×Offer 90:10	-0.06	0.12	-0.49	0.623
DI-Averse×Offer 70:30	-0.07	0.12	-0.57	0.566
DI-Averse×Offer 50:50	0.25	0.12	2.05	0.041
DI-Averse×Offer 30:70	0.32	0.12	2.59	0.010
DI-Averse×Offer 10:90	0.28	0.12	2.29	0.022

Table S6.

Contagion effects in Fairness ratings in Experiment 1, computed as the difference between Transfer and Learning Phase fairness ratings. Linear mixed model coefficients (fixed effects) indicating the effects of Condition and Offer Type on the Fairness rating changes.

Table S7.

Mixed-effects logistic regression examining Rejection choices (Reject vs. Accept) during the Learning Phase in Experiment 1, as a function of the interactions between Condition, Offer Type, and Trial Number.

Predictor	Estimate	SE	p
AI-DI-Averse×Offer 90:10	5.32	0.54	<0.001
AI-DI-Averse×Offer 70:30	0.18	0.29	0.533
AI-DI-Averse×Offer 50:50	-5.47	0.41	<0.001
AI-DI-Averse×Offer 30:70	-2.73	0.35	<0.001
AI-DI-Averse×Offer 10:90	-1.34	0.39	<0.001
DI-Averse×Offer 90:10	4.59	0.52	<0.001
DI-Averse×Offer 70:30	-0.31	0.29	0.288
DI-Averse×Offer 50:50	-5.61	0.43	<0.001
DI-Averse×Offer 30:70	-5.50	0.44	<0.001
DI-Averse×Offer 10:90	-6.20	0.53	<0.001
AI-DI-Averse×Offer 90:10×Trial Number	0.27	0.20	0.164
AI-DI-Averse×Offer 70:30×Trial Number	0.36	0.09	<0.001
AI-DI-Averse×Offer 50:50×Trial Number	0.06	0.14	0.653
AI-DI-Averse×Offer 30:70×Trial Number	0.47	0.12	<0.001
AI-DI-Averse×Offer 10:90×Trial Number	0.77	0.18	<0.001
DI-Averse×Offer 90:10×Trial Number	0.63	0.20	0.001
DI-Averse×Offer 70:30×Trial Number	0.31	0.09	<0.001
DI-Averse×Offer 50:50×Trial Number	0.12	0.15	0.423
DI-Averse×Offer 30:70×Trial Number	-0.30	0.19	0.111
DI-Averse×Offer 10:90×Trial Number	-0.22	0.28	0.429

Table S8.

Summary of Model Comparison in Experiment 1.

Model	nLL	AIC	Mean Parameter Estimates						
Preference Inference	3765.02	9130.04	α_0	β_0	η	τ			
			2.01	1.38	0.14	0.83			
Static preference	4146.96	9493.93	α	β	τ				
			2.01	1.58	0.67				
RL Basic	6519.68	13839.36	η	τ					
			0.28	14.27					
RL Separate Initial	4097.73	10995.45	η	τ	V_1	V_2	V_3	V_4	V_5
			0.06	35.14	0.72	0.44	0.11	0.20	0.26
RL Offer Sensitive	5393.92	13187.84	η_1	η_2	η_3	η_4	η_5	τ	
			0.35	0.16	0.74	0.44	0.48	33.68	
Random Choosing	4393.93	10787.85	p_1	p_2	p_3	p_4	p_5		
			0.80	0.50	0.05	0.16	0.24		

Predictor	Estimate	SE	p
AI-DI-Averse×Offer 90:10	0.13	0.03	<0.001
AI-DI-Averse×Offer 70:30	0.08	0.03	0.003
AI-DI-Averse×Offer 50:50	0.03	0.03	0.316
AI-DI-Averse×Offer 30:70	0.06	0.03	0.052
AI-DI-Averse×Offer 10:90	0.08	0.03	0.003
DI-Averse×Offer 90:10	0.14	0.03	<0.001
DI-Averse×Offer 70:30	0.12	0.03	<0.001
DI-Averse×Offer 50:50	0.02	0.03	0.528
DI-Averse×Offer 30:70	-0.00	0.03	0.907
DI-Averse×Offer 10:90	-0.03	0.03	0.300
AI-DI-Averse×Offer 90:10×Learning	0.22	0.12	0.073
AI-DI-Averse×Offer 70:30×Learning	0.33	0.08	<0.001
AI-DI-Averse×Offer 50:50×Learning	0.10	0.20	0.625
AI-DI-Averse×Offer 30:70×Learning	0.26	0.10	0.010
AI-DI-Averse×Offer 10:90×Learning	0.27	0.07	<0.001
DI-Averse×Offer 90:10×Learning	0.48	0.09	<0.001
DI-Averse×Offer 70:30×Learning	0.36	0.08	<0.001
DI-Averse×Offer 50:50×Learning	0.12	0.18	0.503
DI-Averse×Offer 30:70×Learning	0.25	0.13	0.058
DI-Averse×Offer 10:90×Learning	0.12	0.10	0.196

Table S9.

Relationships between Learning Phase Behavior and Contagion in Experiment 1. Linear Mixed model coefficients (fixed effects) indicating the effects of Condition, Offer Type, and Learning index (Rejection rate change between first and last five trials of Learning phase) on the Contagion effect (Rejection rate changes from Baseline to Transfer phase).

Table S10.

Contagion effects in Experiment 2. Linear mixed model coefficients (fixed effects) indicating the effects of Condition and Offer Type on the Rejection rate changes (from Baseline phase to Transfer phase).

Predictor	Estimate	SE	p
AI-DI-Averse×Offer 90:10	0.13	0.03	<0.001
AI-DI-Averse×Offer 70:30	0.10	0.03	<0.001
AI-DI-Averse×Offer 50:50	0.05	0.03	0.078
AI-DI-Averse×Offer 30:70	0.03	0.03	0.345
AI-DI-Averse×Offer 10:90	0.06	0.03	0.060
DI-Averse×Offer 90:10	0.18	0.03	<0.001
DI-Averse×Offer 70:30	0.09	0.03	0.002
DI-Averse×Offer 50:50	-0.02	0.03	0.492
DI-Averse×Offer 30:70	-0.03	0.03	0.280
DI-Averse×Offer 10:90	-0.06	0.03	0.035

Table S11.

Contagion effect of Fairness rating in Experiment 2. Linear mixed model coefficients (fixed effects) indicating the effects of Condition and Offer Type on the Fairness rating changes from Baseline to Transfer phase.

Predictor	Estimate	SE	p
AI-DI-Averse×Offer 90:10	-0.23	0.13	0.076
AI-DI-Averse×Offer 70:30	-0.41	0.13	0.001
AI-DI-Averse×Offer 50:50	-0.24	0.13	0.056
AI-DI-Averse×Offer 30:70	-0.53	0.13	<0.001
AI-DI-Averse×Offer 10:90	-0.84	0.13	<0.001
DI-Averse×Offer 90:10	-0.29	0.13	0.023
DI-Averse×Offer 70:30	-0.15	0.13	0.228
DI-Averse×Offer 50:50	0.24	0.13	0.060
DI-Averse×Offer 30:70	0.09	0.13	0.498
DI-Averse×Offer 10:90	0.14	0.13	0.272

Predictor	Estimate	SE	p
AI-DI-Averse×Offer 90:10	2.35	0.50	<0.001
AI-DI-Averse×Offer 70:30	-0.86	0.35	0.013
AI-DI-Averse×Offer 50:50	-3.61	0.30	<0.001
AI-DI-Averse×Offer 30:70	-3.32	0.34	<0.001
AI-DI-Averse×Offer 10:90	-1.95	0.44	<0.001
DI-Averse×Offer 90:10	1.48	0.49	0.003
DI-Averse×Offer 70:30	-1.38	0.35	<0.001
DI-Averse×Offer 50:50	-4.47	0.34	<0.001
DI-Averse×Offer 30:70	-4.76	0.37	<0.001
DI-Averse×Offer 10:90	-5.45	0.53	<0.001
AI-DI-Averse×Offer 90:10×Trial Number	0.36	0.26	0.167
AI-DI-Averse×Offer 70:30×Trial Number	0.13	0.08	0.093
AI-DI-Averse×Offer 50:50×Trial Number	0.56	0.24	0.019
AI-DI-Averse×Offer 30:70×Trial Number	0.10	0.10	0.300
AI-DI-Averse×Offer 10:90×Trial Number	1.06	0.31	<0.001
DI-Averse×Offer 90:10×Trial Number	0.96	0.27	<0.001
DI-Averse×Offer 70:30×Trial Number	0.26	0.08	0.001
DI-Averse×Offer 50:50×Trial Number	-0.14	0.27	0.593
DI-Averse×Offer 30:70×Trial Number	-0.37	0.13	0.003
DI-Averse×Offer 10:90×Trial Number	-0.54	0.40	0.170

Table S12.

Mixed-effects logistic regression examining Rejection choices (Reject vs. Accept) during the Learning Phase in Experiment 2, as a function of the interactions between Condition, Offer Type, and Trial Number.

Table S13.

Mixed-effects regression examining Fairness ratings during the Learning Phase in Experiment 2, as a function of the interactions between Condition, Offer Type, and Trial Number.

Predictor	Estimate	SE	p
AI-DI-Averse×Offer 90:10	2.40	0.16	<0.001
AI-DI-Averse×Offer 70:30	3.22	0.11	<0.001
AI-DI-Averse×Offer 50:50	5.09	0.09	<0.001
AI-DI-Averse×Offer 30:70	3.50	0.12	<0.001
AI-DI-Averse×Offer 10:90	2.58	0.18	<0.001
DI-Averse×Offer 90:10	2.72	0.16	<0.001
DI-Averse×Offer 70:30	3.59	0.12	<0.001
DI-Averse×Offer 50:50	5.41	0.09	<0.001
DI-Averse×Offer 30:70	4.45	0.12	<0.001
DI-Averse×Offer 10:90	4.08	0.18	<0.001
AI-DI-Averse×Offer 90:10×Trial Number	0.05	0.06	0.424
AI-DI-Averse×Offer 70:30×Trial Number	-0.04	0.02	0.071
AI-DI-Averse×Offer 50:50×Trial Number	0.12	0.08	0.126
AI-DI-Averse×Offer 30:70×Trial Number	-0.10	0.03	0.002
AI-DI-Averse×Offer 10:90×Trial Number	-0.27	0.09	0.004
DI-Averse×Offer 90:10×Trial Number	-0.09	0.06	0.131
DI-Averse×Offer 70:30×Trial Number	-0.02	0.02	0.333
DI-Averse×Offer 50:50×Trial Number	0.16	0.08	0.054
DI-Averse×Offer 30:70×Trial Number	0.01	0.03	0.828
DI-Averse×Offer 10:90×Trial Number	0.07	0.09	0.448

Table S14.

Summary of Model Comparison in Experiment 2.

Model	nLL	AIC	Mean Parameter Estimates						
Preference Inference	6250.45	14100.90	α_0	β_0	η	τ			
			1.75	1.48	0.07	0.82			
Static preference	6522.31	14244.63	α	β	τ				
			1.74	1.53	0.78				
RL Separate Initial	6720.48	16240.97	η	τ	V_1	V_2	V_3	V_4	V_5
			0.01	38.16	0.58	0.39	0.15	0.23	0.26
Random Choosing	6894.55	15789.11	p_1	p_2	p_3	p_4	p_5		
			0.62	0.43	0.08	0.15	0.22		

References

- Akaike H. (1974) **A New Look at the Statistical Model Identification** *IEEE Transactions on Automatic Control* **19**:716–723 <https://doi.org/10.1109/TAC.1974.1100705>
- Amir D. *et al.* (2023) **Computational signatures of inequity aversion in children across seven societies** *Journal of Experimental Psychology: General* **152**:2882–2896 <https://doi.org/10.1037/xge0001385>
- Anzellotti S., Young L. L. (2020) **The acquisition of person knowledge** *Annual Review of Psychology* **71**:613–634 <https://doi.org/10.1146/annurev-psych-010419-050844>
- Bail C. A., Argyle L. P., Brown T. W., Bumpus J. P., Chen H., Fallin Hunzaker M. B., Lee J., Mann M., Merhout F., Volfovsky A. (2018) **Exposure to opposing views on social media can increase political polarization** *Proceedings of the National Academy of Sciences of the United States of America* **115**:9216–9221 <https://doi.org/10.1073/pnas.1804840115>
- Baker, Chris L., Saxe R., Tenenbaum J. B. (2009) **Citation Accessed Action Understanding as Inverse Planning** *Cognition* **113**:329–349
- Bandura A., McDonald Frederick J. (1963) **Influence of social reinforcement and the behavior of models in shaping children's moral judgment** *Influence of Social Reinforcement and the Behavior of Models in Shaping Children's Moral Judgment* **67**:274–281
- Bates D., Mächler M., Bolker B., Walker S. (2015) **Fitting Linear Mixed-Effects Models Using lme4** *Journal of Statistical Software* **67** <https://doi.org/10.18637/jss.v067.i01>
- Blake P. R., McAuliffe K. (2011) **“I had so much it didn't seem fair” Eight-year-olds reject two forms of inequity** *Cognition* **120**:215–224 <https://doi.org/10.1016/j.cognition.2011.04.006>
- Blake P. R. *et al.* (2015) **The ontogeny of fairness in seven societies** *Nature* **528**:258–261 <https://doi.org/10.1038/nature15703>
- Brosnan S. F., De Waal F. B. M. (2003) **Monkeys reject unequal pay** *Nature* **425**:297–299 <https://doi.org/10.1038/nature01963>
- Brosnan S. F., De Waal F. B. M. (2014) **Evolution of responses to (un)fairness** *Science* **346** <https://doi.org/10.1126/science.1251776>
- Burke C. J., Tobler P. N., Baddeley M., Schultz W. (2010) **Neural mechanisms of observational learning** *Proceedings of the National Academy of Sciences* **107**:14431–14436 <https://doi.org/10.1073/pnas.1003111107>
- Cialdini R. B., Goldstein N. J. (2004) **Social Influence: Compliance and Conformity** *Annual Review of Psychology* **55**:591–621 <https://doi.org/10.1146/annurev.psych.55.090902.142015>
- Civai C., Crescentini C., Rustichini A., Rumiati R. I. (2012) **Equality versus self-interest in the brain: Differential roles of anterior insula and medial prefrontal cortex** *NeuroImage* **62**:102–112 <https://doi.org/10.1016/j.neuroimage.2012.04.037>

- Crump M. J. C., McDonnell J. V., Gureckis T. M. (2013) **Evaluating Amazon’s Mechanical Turk as a Tool for Experimental Behavioral Research** *PLoS ONE* **8** <https://doi.org/10.1371/journal.pone.0057410>
- Diaconescu A. O., Stecy M., Kasper L., Burke C. J., Nagy Z., Mathys C., Tobler P. N. (2020) **Neural arbitration between social and individual learning systems** *eLife* **9**:1–35 <https://doi.org/10.7554/eLife.54051>
- Engelmann D., Fischbacher U. (2009) **Indirect reciprocity and strategic reputation building in an experimental helping game** *Games and Economic Behavior* **67**:399–407 <https://doi.org/10.1016/j.geb.2008.12.006>
- Essler J. L., Marshall-Pescini S., Range F. (2017) **Domestication Does Not Explain the Presence of Inequity Aversion in Dogs** *Current Biology* **27**:1861–1865 <https://doi.org/10.1016/j.cub.2017.05.061>
- Fehr E., Bernhard H., Rockenbach B. (2008) **Egalitarianism in young children** *Nature* **454**:1079–1083 <https://doi.org/10.1038/nature07155>
- Fehr E., Schmidt K. M. (1999) **A Theory of Fairness , Competition , and Cooperation** *The Quarterly Journal of Economics* **114**:817–868
- FeldmanHall O., Otto A. R., Phelps E. A. (2018) **Learning moral values: Another’s desire to punish enhances one’s own punitive behavior** *Journal of Experimental Psychology: General* **147**:1211–1224 <https://doi.org/10.1037/xge0000405>
- FeldmanHall O., Shenhav A. (2019) **Resolving uncertainty in a social world** *Nature Human Behaviour* <https://doi.org/10.1038/s41562-019-0590-x>
- Gao X., Yu H., Sáez I., Blue P. R., Zhu L., Hsu M., Zhou X. (2018) **Distinguishing neural correlates of context-dependent advantageous- and disadvantageous-inequity aversion** *Proceedings of the National Academy of Sciences of the United States of America* **115** <https://doi.org/10.1073/pnas.1802523115>
- Garvert M. M., Moutoussis M., Kurth-Nelson Z., Behrens T. E. J., Dolan R. J. (2015) **Learning-Induced Plasticity in Medial Prefrontal Cortex Predicts Preference Malleability** *Neuron* **85**:418–428 <https://doi.org/10.1016/j.neuron.2014.12.033>
- Güth W., Schmittberger R., Schwarze B. (1982) **An experimental analysis of ultimatum bargaining** *Journal of Economic Behavior & Organization* **3**:367–388 [https://doi.org/10.1016/0167-2681\(82\)90011-7](https://doi.org/10.1016/0167-2681(82)90011-7)
- Hayes S. J., Ashford D., Bennett S. J. (2008) **Goal-directed imitation: The means to an end** *Acta Psychologica* **127**:407–415 <https://doi.org/10.1016/j.actpsy.2007.07.009>
- Heffner J., FeldmanHall O. (2022) **A probabilistic map of emotional experiences during competitive social interactions** *Nature Communications* **13**:1–11 <https://doi.org/10.1038/s41467-022-29372-8>
- Hein G., Silani G., Preuschhoff K., Batson C. D., Singer T. (2010) **Neural responses to ingroup and outgroup members’ suffering predict individual differences in costly helping** *Neuron* **68**:149–160 <https://doi.org/10.1016/j.neuron.2010.09.003>

- Hennig-Schmidt H., Li Z. Y., Yang C. (2008) **Why people reject advantageous offers-Non-monotonic strategies in ultimatum bargaining. Evaluating a video experiment run in PR China** *Journal of Economic Behavior and Organization* **65**:373–384 <https://doi.org/10.1016/j.jebo.2005.10.003>
- Henrich J. *et al.* (2006) **Costly punishment across human societies** *Science* **312**:1767–1770 <https://doi.org/10.1126/science.1127333>
- Hertz U. (2021) **Learning how to behave: Cognitive learning processes account for asymmetries in adaptation to social norms** *Proceedings of the Royal Society B: Biological Sciences* **288** <https://doi.org/10.1098/rspb.2021.0293>
- Hester N., Gray K. (2020) **The Moral Psychology of Raceless, Genderless Strangers** *Perspectives on Psychological Science* **15**:216–230 <https://doi.org/10.1177/1745691619885840>
- Jara-ettinger J. (2019) **ScienceDirect Theory of mind as inverse reinforcement learning** *COBEHA* **29**:105–110 <https://doi.org/10.1016/j.cobeha.2019.04.010>
- Khalvati K., Park S. A., Mirbagheri S., Philippe R., Sestito M., Dreher J.-C., Rao R. P. N. (2019) **Modeling Other Minds: Bayesian Inference Explains Human Choices in Group Decision Making** *Science Advances* <https://doi.org/10.1101/419515>
- Leong Y. C., Chen J., Willer R., Zaki J. (2020) **Conservative and liberal attitudes drive polarized neural responses to political content** *Proceedings of the National Academy of Sciences* **117**:27731–27739 <https://doi.org/10.1073/pnas.2008530117>
- Lindström B., Golkar A., Jangard S., Tobler P. N., Olsson A. (2019) **Lindstrom et al. - 2019—Social threat learning transfers to decision making in humans.pdf** *Proceedings of the National Academy of Sciences of the United States of America* **116**:4732–4737 <https://doi.org/10.1073/pnas.1810180116>
- Luo Y. *et al.* (2018) **Early childhood investment impacts social decision-making four decades later** *Nature Communications* **9** <https://doi.org/10.1038/s41467-018-07138-5>
- McAuliffe K., Blake P. R., Kim G., Wrangham R. W., Warneken F. (2013) **Social influences on inequity aversion in children** *PLoS ONE* **8** <https://doi.org/10.1371/journal.pone.0080966>
- McAuliffe K., Blake P. R., Steinbeis N., Warneken F. (2017) **The developmental foundations of human fairness** *Nature Human Behaviour* **1** <https://doi.org/10.1038/s41562-016-0042>
- McAuliffe K., Blake P. R., Warneken F. (2014) **Children reject inequity out of spite** *Biology Letters* **10**:1–5 <https://doi.org/10.1098/rsbl.2014.0743>
- McAuliffe K., Dunham Y. (2017) **Fairness overrides group bias in children’s second-party punishment** *Journal of Experimental Psychology: General* **146**:485–494 <https://doi.org/10.1037/xge0000244>
- Pedersen E. J., Kurzban R., McCullough M. E. (2013) **Do humans really punish altruistically? A closer look** *Proceedings of the Royal Society B: Biological Sciences* **280** <https://doi.org/10.1098/rspb.2012.2723>
- Pillutla M. M., Murnighan J. K. (1996) **Unfairness, anger, and spite: Emotional rejections of ultimatum offers** *Organizational Behavior and Human Decision Processes* **68**:208–224 <https://doi.org/10.1006/obhd.1996.0100>

- Sanfey A. G., Rilling J. K., Aronson J. A., Nystrom L. E., Cohen J. D. (2003) **The neural basis of economic decision-making in the Ultimatum Game** *Science* **300**:1755–1758 <https://doi.org/10.1126/science.1082976>
- Son J. Y., Bhandari A., FeldmanHall O. (2019) **Crowdsourcing punishment: Individuals reference group preferences to inform their own punitive decisions** *Scientific Reports* **9**:1–15 <https://doi.org/10.1038/s41598-019-48050-2>
- Suzuki S., Jensen E. L. S., Bossaerts P., O’Doherty J. P. (2016) **Behavioral contagion during learning about another agent’s risk-preferences acts on the neural representation of decision-risk** *Proceedings of the National Academy of Sciences* **113**:3755–3760 <https://doi.org/10.1073/pnas.1600092113>
- Taber C. S., Lodge M. (2012) **Motivated skepticism in the evaluation of political beliefs** (2006) *Critical Review* **24**:157–184 <https://doi.org/10.1080/08913811.2012.711019>
- Tomasello M. (2019) **Becoming human: A theory of ontogeny** *Belknap*
- Tsvetkova M., Macy M. W. (2014) **The Social Contagion of Generosity** *PLOS ONE* **9** <https://doi.org/10.1371/journal.pone.0087275>
- van Baar J. M., Chang L. J., Sanfey A. G. (2019) **The computational and neural substrates of moral strategies in social decision-making** *Nature Communications* **10** <https://doi.org/10.1038/s41467-019-09161-6>
- Van Wolkenten M., Brosnan S. F., De Waal F. B. M. (2007) **Inequity responses of monkeys modified by effort** *Proceedings of the National Academy of Sciences of the United States of America* **104**:18854–18859 <https://doi.org/10.1073/pnas.0707182104>
- Vives M. L., Cikara M., FeldmanHall O. (2022) **Following Your Group or Your Morals? The In Group Promotes Immoral Behavior While the Out-Group Buffers Against It** *Social Psychological and Personality Science* **13**:139–149 <https://doi.org/10.1177/19485506211001217>
- Vives M. L., Feldmanhall O. (2018) **Tolerance to ambiguous uncertainty predicts prosocial behavior** *Nature Communications* **9**:25–27 <https://doi.org/10.1038/s41467-018-04631-9>

Author information

Shen Zhang

State Key Laboratory of Cognitive Neuroscience and Learning & IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing, China, Department of Neurobiology, German Primate Center, Göttingen, Germany

For correspondence: Shen.Zhang@mail.bnu.edu.cn

Oriel FeldmanHall

Cognitive, Linguistics and Psychological Sciences, Brown University, Providence, United States, Carney Institute for Brain Sciences, Brown University, Providence, United States

Sébastien Héту

Department of Psychology, Université de Montréal, Montréal, Canada, Centre Interdisciplinaire de Recherche sur le Cerveau et l'Apprentissage (CIRCA), Montréal, Canada

A Ross Otto

Department of Psychology, McGill University, Montréal, Canada

Editors

Reviewing Editor

Clare Press

University College London, London, United Kingdom

Senior Editor

Christian Büchel

University Medical Center Hamburg-Eppendorf, Hamburg, Germany

Reviewer #1 (Public review):

Summary:

Zhang et al. addressed the question of whether advantageous and disadvantageous inequality aversion can be vicariously learned and generalized. Using an adapted version of the ultimatum game (UG), in three phases, participants first gave their own preference (baseline phase), then interacted with a "teacher" to learn their preference (learning phase), and finally were tested again on their own (transfer phase). The key measure is whether participants exhibited similar choice preferences (i.e., rejection rate and fairness rating) influenced by the learning phase, by contrasting their transfer phase and baseline phase. Through a series of statistical modeling and computational modeling, the authors reported that both advantageous and disadvantageous inequality aversion can indeed be learned (Study 1), and even be generalised (Study 2).

Strengths:

This study is very interesting, it directly adapted the lab's previous work on the observational learning effect on disadvantageous inequality aversion, to test both advantageous and disadvantageous inequality aversion in the current study. Social transmission of action, emotion, and attitude have started to be looked at recently, hence this research is timely. The use of computational modeling is mostly appropriate and motivated. Study 2, which examined the vicarious inequality aversion in conditions where feedback was never provided, is interesting and important to strengthen the reported effects. Both studies have proper justifications to determine the sample size.

Weaknesses:

Despite the strengths, a few conceptual aspects and analytical decisions have to be explained, justified, or clarified.

INTRODUCTION/CONCEPTUALIZATION

(1) Two terms seem to be interchangeable, which should not, in this work: vicarious/observational learning vs preference learning. For vicarious learning, individuals observe others' actions (and optionally also the corresponding consequence resulting directly from their own actions), whereas, for preference learning, individuals predict, or act on behalf of, the others' actions, and then receive feedback if that prediction is correct or not. For the current work, it seems that the experiment is more about preference learning and

prediction, and less so about vicarious learning. The intro and set are heavily around vicarious learning, and later the use of vicarious learning and preference learning is rather mixed in the text. I think either tone down the focus on vicarious learning, or discuss how they are different. Some of the references here may be helpful: Charpentier et al., *Neuron*, 2020; Olsson et al., *Nature Reviews Neuroscience*, 2020; Zhang & Glascher, *Science Advances*, 2020

EXPERIMENTAL DESIGN

(2) For each offer type, the experiment "added a uniformly distributed noise in the range of $(-10, 10)$ ". I wonder what this looks like? With only integers such as 25:75, or even with decimal points? More importantly, is it possible to have either 70:30 or 90:10 option, after adding the noise, to have generated an 80:20 split shown to the participants? If so, for the analyses later, when participants saw the 80:20 split, which condition did this trial belong to? 70:30 or 90:10? And is such noise added only to the learning phase, or also to the baseline/transfer phases? This requires some clarification.

(3) For the offer conditions (90:10, 70:30, 50:50, 30:70, 10:90) - are they randomized? If so, how is it done? Is it randomized within each participant, and/or also across participants (such that each participant experienced different trial sequences)? This is important, as the order especially for the learning phase can largely impact the preference learning of the participants.

STATISTICAL ANALYSIS & COMPUTATIONAL MODELING

(4) In Study 1 DI offer types (90:10, 70:30), the rejection rate for DI-AI averse looks consistently higher than that for DI averse (ie, the blue line is above the yellow line). Is this significant? If so, how come? Since this is a between-subject design, I would not anticipate such a result (especially for the baseline). Also, for the LME results (eg, Table S3), only interactions were reported but not the main results.

(5) I do not particularly find this analysis appealing: "we examined whether participants' changes in rejection rates between Transfer and Baseline, could be explained by the degree to which they vicariously learned, defined as the change in punishment rates between the first and last 5 trials of the Learning phase." Naturally, the participants' behavior in the first 5 trials in the learning phase will be similar to those in the baseline; and their behavior in the last 5 trials in the learning phase would echo those at the transfer phase. I think it would be stronger to link the preference learning results to the change between the baseline and transfer phase, eg, by looking at the difference between alpha (beta) at the end of the learning phase and the initial alpha (beta).

(6) I wonder if data from the baseline and transfer phases can also be modeled, using a simple Fehr-Schmidt model. This way, the change in alpha/beta can also be examined between the baseline and transfer phase.

(7) I quite liked Study 2 which tests the generalization effect, and I expected to see an adapted computational modeling to directly reflect this idea. Indeed, the authors wrote, "[...] given that this model [...] assumes the sort of generalization of preferences between offer types [...]". But where exactly did the preference learning model assume the generalization? In the methods, the modeling seems to be only about Study 1; did the authors advise their model to accommodate Study 2? The authors also ran simulation for the learning phase in Study 2 (Figure 6), and how did the preference update (if at all) for offers (90:10 and 10:90) where feedback was not given? Extending/Unpacking the computational modeling results for Study 2 will be very helpful for the paper.

<https://doi.org/10.7554/eLife.102800.1.sa1>

Reviewer #2 (Public review):**Summary:**

This study investigates whether individuals can learn to adopt egalitarian norms that incur a personal monetary cost, such as rejecting offers that benefit them more than the giver (advantageous inequitable offers). While these behaviors are uncommon, two experiments demonstrate that individuals can learn to reject such offers through vicarious learning - by observing and acting in line with a "teacher" who follows these norms. The authors use computational modelling to argue that learners adopt these norms through a sophisticated process, inferring the latent structure of the teacher's preferences, akin to theory of mind.

Strengths:

This paper is well-written and tackles a critical topic relevant to social norms, morality, and justice. The findings, which show that individuals can adopt just and fair norms even at a personal cost, are promising. The study is well-situated in the literature, with clever experimental design and a computational approach that may offer insights into latent cognitive processes. Findings have potential implications for policymakers.

Weaknesses:

Note: in the text below, the "teacher" will refer to the agent from which a participant presumably receives feedback during the learning phase.

(1) Focus on Disadvantageous Inequity (DI): A significant portion of the paper focuses on responses to Disadvantageous Inequitable (DI) offers, which is confusing given the study's primary aim is to examine learning in response to Advantageous Inequitable (AI) offers. The inclusion of DI offers is not well-justified and distracts from the main focus. Furthermore, the experimental design seems, in principle, inadequate to test for the learning effects of DI offers. Because both teaching regimes considered were identical for DI offers the paradigm lacks a control condition to test for learning effects related to these offers. I can't see how an increase in rejection of DI offers (e.g., between baseline and generalization) can be interpreted as speaking to learning. There are various other potential reasons for an increase in rejection of DI offers even if individuals learn nothing from learning (e.g. if envy builds up during the experiment as one encounters more instances of disadvantageous fairness).

(2) Statistical Analysis: The analysis of the learning effects of AI offers is not fully convincing. The authors analyse changes in rejection rates within each learning condition rather than directly comparing the two. Finding a significant effect in one condition but not the other does not demonstrate that the learning regime is driving the effect. A direct comparison between conditions is necessary for establishing that there is a causal role for the learning regime.

(3) Correlation Between Learning and Contagion Effects:

The authors argue that correlations between learning effects (changes in rejection rates during the learning phase) and contagion effects (changes between the generalization and baseline phases) support the idea that individuals who are better aligning their preferences with the teacher also give more consideration to the teacher's preferences later during generalization phase. This interpretation is not convincing. Such correlations could emerge even in the absence of learning, driven by temporal trends like increasing guilt or envy (or even by slow temporal fluctuations in these processes) on behalf of self or others. The reason is that the baseline phase is temporally closer to the beginning of the learning phase whereas the generalization phase is temporally closer to the end of the learning phase. Additionally, the interpretation of these effects seems flawed, as changes in rejection rates do not necessarily indicate closer alignment with the teacher's preferences. For example, if the

teacher rejects an offer 75% of the time then a positive 5% learning effect may imply better matching the teacher if it reflects an increase in rejection rate from 65% to 70%, but it implies divergence from the teacher if it reflects an increase from 85% to 90%. For similar reasons, it is not clear that the contagion effects reflect how much a teacher's preferences are taken into account during generalization.

(4) Modeling Efforts: The modelling approach is underdeveloped. The identification of the "best model" lacks transparency, as no model-recovery results are provided, and fits for the losing models are not shown, leaving readers in the dark about where these models fail. Moreover, the reinforcement learning (RL) models used are overly simplistic, treating actions as independent when they are likely inversely related (for example, the feedback that the teacher would have rejected an offer provides feedback that rejection is "correct" but also that acceptance is "an error", and the later is not incorporated into the modelling). It is unclear if and to what extent this limits current RL formulations. There are also potentially important missing details about the models. Can the authors justify/explain the reasoning behind including these variants they consider? What are the initial Q-values? If these are not free parameters what are their values?

(5) Conceptual Leap in Modeling Interpretation: The distinction between simple RL models and preference-inference models seems to hinge on the ability to generalize learning from one offer to another. Whereas in the RL models learning occurs independently for each offer (hence to cross-offer generalization), preference inference allows for generalization between different offers. However, the paper does not explore RL models that allow generalization based on the similarity of features of the offers (e.g., payment for the receiver, payment for the offer-giver, who benefits more). Such models are more parsimonious and could explain the results without invoking a theory of mind or any modelling of the teacher. In such model versions, a learner learns a functional form that allows to predict the teacher's feedback based on said offer features (e.g., linear or quadratic form). Because feedback for an offer modulates the parameters of this function (feature weights) generalization occurs without necessarily evoking any sophisticated model of the other person. This leaves open the possibility that RL models could perform just as well or even show superiority over the preference learning model, casting doubt on the authors' conclusions. Of note: even the behaviourists knew that as Little Albert was taught to fear rats, this fear generalized to rabbits. This could occur simply because rabbits are somewhat similar to rats. But this doesn't mean little Alfred had a sophisticated model of animals he used to infer how they behave.

(6) Limitations of the Preference-Inference Model: The preference-inference model struggles to capture key aspects of the data, such as the increase in rejection rates for 70:30 DI offers during the learning phase (e.g. Figure 3A, AI+DI blue group). This is puzzling.

Thinking about this I realized the model makes quite strong unintuitive predictions that are not examined. For example, if a subject begins the learning phase rejecting the 70:30 offer more than 50% of the time (meaning the starting guilt parameter is higher than 1.5), then overlearning the tendency to reject will decrease to below 50% (the guilt parameter will be pulled down below 1.5). This is despite the fact the teacher rejects 75% of the offers. In other words, as learning continues learners will diverge from the teacher. On the other hand, if a participant begins learning to tend to accept this offer (guilt < 1.5) then during learning they can increase their rejection rate but never above 50%. Thus one can never fully converge on the teacher. I think this relates to the model's failure in accounting for the pattern mentioned above. I wonder if individuals actually abide by these strict predictions. In any case, these issues raise questions about the validity of the model as a representation of how individuals learn to align with a teacher's preferences (given that the model doesn't really allow for such an alignment).

<https://doi.org/10.7554/eLife.102800.1.sa0>