Observing Others Give & Take:

A Computational Account of Bystanders' Feelings and Actions

Joseph Marks[1^], Philipp Czech[1,2], Tali Sharot[1^]

1. Affective Brain Lab, Experimental Psychology, University College London, London, UK
2. Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany.

[^]Corresponding authors: t.sharot@ucl.ac.uk , joseph.marks.14@ucl.ac.uk

Word Count = 5542

Draft version, 13/11/19. This paper has not (yet) been peer reviewed or published and is not therefore the authoritative document of record.

Abstract

Social interactions influence people's feelings and behaviour. This is true not only when people are directly involved, but also when observing others interact. Little is known, however, about how others' interactions impact bystanders. Here, we developed a computational model that relates others' (un)selfish acts to observers' emotional reactions (which we call a 'feelings function') and punishment decisions ('punishment function'). Our models enabled us to quantify the impact of two social values: 'selfishness aversion' and 'inequity aversion' on feelings and punishment. The results revealed diminishing sensitivity to actions that violated these social norms. That is, small violations from equality and generosity had a disproportionately large impact on feelings and punishment. We then used our 'feelings function' to predict observers' feelings on out-of-sample trials and found that those estimated feelings were strongly correlated with observers' punishment decisions. This suggests that observers were often acting in accordance with their affective responses to selfishness and inequity when punishing. We further show that when affective responses indeed align with punishment decisions, participants feel better about their decisions. The study characterizes computational rules by which social interactions between other people are transformed into bystanders' reactions.

Keywords: Affect, Bystander, Decision Making, Punishment

Humans are social animals. We live among, and interact with, other humans daily. Social interactions can have a significant emotional impact on individuals (Carlsmith, Wilson, & Gilbert, 2008; Pillutla & Murnighan, 1996; Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003; Van't Wout, Kahn, Sanfey, & Aleman, 2006; Zheng, Yang, Jin, Qi, & Liu, 2017). For example, being the recipient of generosity will likely make us feel good, while being the recipient of selfish behaviour will likely make us feel bad. Because people care not only about their own well-being but also that of others (Crockett, Kurth-Nelson, Siegel, Dayan, & Dolan, 2014; Crockett, Siegel, Kurth-Nelson, Dayan, & Dolan, 2017; Dawes, Fowler, Johnson, Mcelreath, & Smirnov, 2007), it is likely that individuals will be impacted not only by direct interactions, but also by observing others interact. In other words, people may feel good when observing others behave generously and feel bad when observing others behave selfishly, even if they are not the recipient of the behaviour. Little is known, however, of how such observations are transformed into affective reactions and decisions to alter the status quo.

Here, we set out to characterize the computational rules by which observations are translated into feelings and action. This knowledge is important for two reasons. First, when developing public-policy it may be imperative to consider not only how a policy will impact people directly, but also how it may impact observing-third-parties indirectly (Posner & Sunstein, 2017; Sunstein, 2019a, 2019b). For example, stringent harassment laws may benefit not only those vulnerable to harassment but also third parties observing others being harassed. Second, it is theorized that feelings are important in governing choice (Charpentier, De Neve, Li, Roiser, & Sharot, 2016; Nelissen & Zeelenberg, 2009). Thus, if we are able to measure people's feelings when observing others interact we may be able to generate that person's utility function and use it to predict action. For example, predicting the likelihood that an observer will intervene when observing harassment.

To that end we recorded observers' explicit affective reactions and punishment choices in response to other people's decisions to allocate resources to themselves and another individual. We used a computational modelling approach to relate observers' affective reactions and punishment choices to others' behaviour, creating what we refer to as a 'feeling function' and a 'punishment function'. These functions allow for quantification of people's social values as they influence feelings and action. We pose that two dominant social values may play a role: 'selfishness aversion' and 'inequity aversion'. Because people are averse to selfish behaviour they may have a negative reaction when observing others allocate more to themselves than to others (Fehr & Fischbacher, 2004; Fehr & Gächter, 2002; Fliessbach et al., 2012; Rosas & Koenigs, 2014). Conversely, they may have a positive reaction when observing others allocate less to themselves than to others. However, inequity aversion (Gao et al., 2018; Sáez, Zhu, Set, Kayser, & Hsu, 2015; Yu, Calder, & Mobbs, 2014) may cause a negative reaction when observing unequal distribution of resources, regardless of whether this allocation resulted from generosity or selfishness.

We further examined whether observers' act in accordance with their feelings and what happens when observers' actions do not align with feelings. In particular, we asked observers how they felt about their decisions to punish, or refrain from punishing, the allocators. We hypothesized that observers will feel better when their affective response to others' behaviour is aligned with their own actions.

Results

We run two experiments with a total of sixty-seven participants. On each of 240 trials a participant observed what they believed was another participant (the 'allocator', whom was said to differ on each trial) make decisions about how to divide a sum of money between themselves and what they believed to be a third participant (who differed on each trial). On some trials the participants then had the opportunity to punish the allocator by giving some of the allocator's money back to the experimenter and subsequently indicated how they felt about their punishment decision. On other trials the participant rated how they felt about the allocations. In Experiment 1 the allocating agents chose how much money to take from someone who had received an endowment from the experimenter. In Experiment 2 the allocator was given an endowment and chose how much of it to give to another agent.

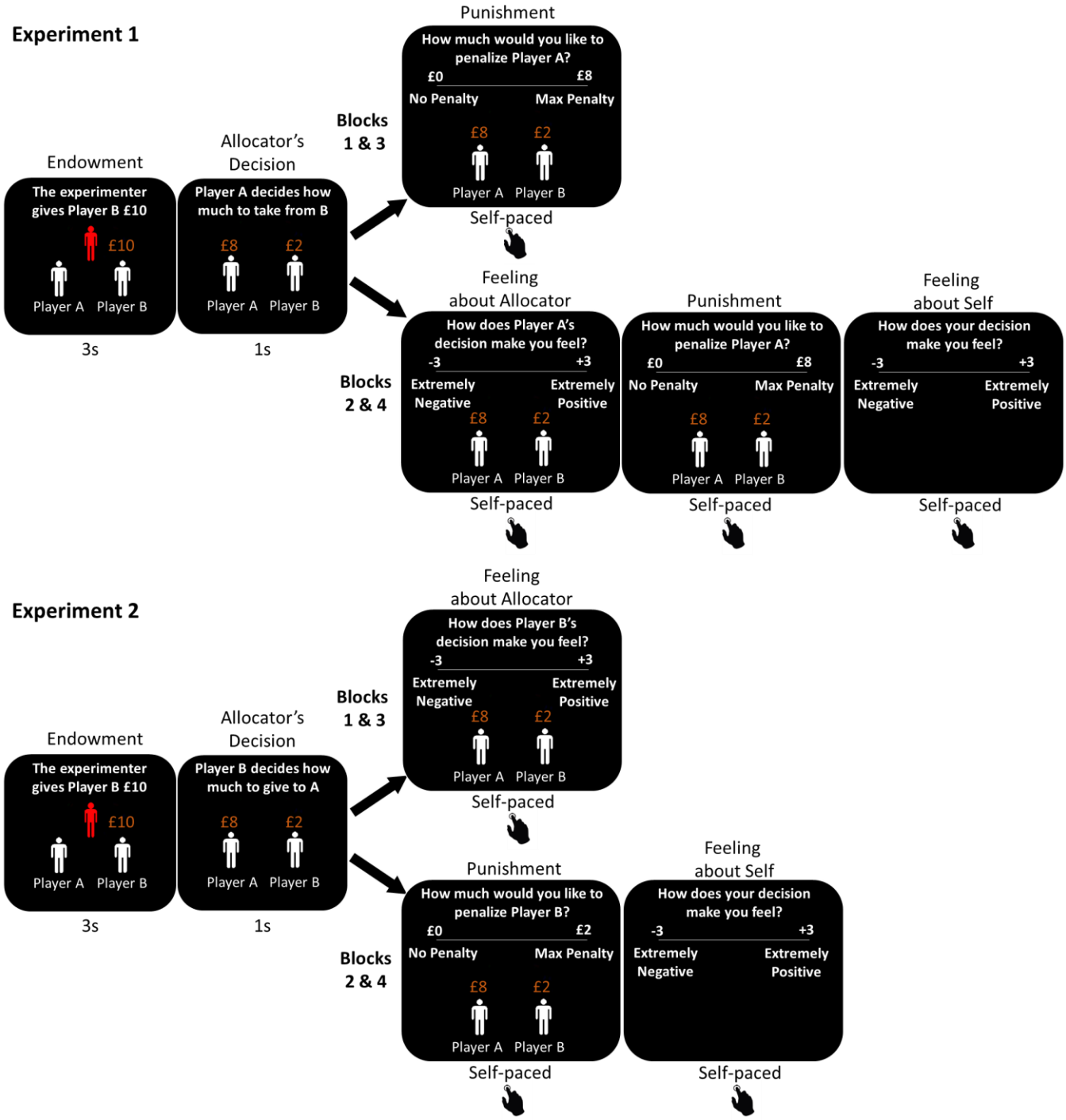

Figure 1. Task. Participants observed what they were led to believe were other participants' resource allocation decisions. On each trial, the Experimenter gave a financial endowment (£1 to £15, step size = £1) to Player B. **Experiment 1:** Player A, the allocator, could then take a portion of this money for themselves (10% to 100%, step size = 10%). In blocks 2 and 4 participants rated how they felt about the allocator's decision. In all blocks, participants decided if and by how much to punish the allocator. In blocks 2 and 4 participants subsequently rated how they felt about their punishment decision. **Experiment 2:** Player B was the allocator and could share a portion of their money with Player A (0% to 90%, step size = 10%). In blocks 1 and 3 participants rated how they felt about the allocator's decision. In blocks 2 and 4 participants decided if and by how much to punish the allocator and then rated how they felt about their punishment decision.

Observers' affective responses are influenced by observed selfishness and inequity. First, we tested whether observers were negatively affected by observing others act selfishly - even though the selfish behavior was not directed towards them - and positively when observing others split the resources equally or act generously (**Figure 2**). A one-way repeated-measures ANOVA revealed a significant effect of condition (selfish allocation, equal allocation, generous allocation) on feelings (Experiment 1: $F(2,30) = 47.66$, $p < .001$, $\eta_p^2 = 0.76$; Experiment 2: $F(2,33) = 41.66$, $p < .001$, $\eta_p^2 = 0.72$). Indeed, participants reported negative affect when watching the allocator take more than half the endowment (feeling rating significantly lower than zero: $t(31) = -11.57$, $p < .001$, $d = -2.05$) as well as keep more than half the endowment ($t(34) = -4.94$, $p < .001$, $d = -.83$). This is unsurprising and consistent with past findings showing that observers are often willing to pay a cost to punish selfish behaviour (Fehr & Gächter, 2002; Jordan, Hoffman, Bloom, & Rand, 2016). When the allocator split the money equally the observers reported positive affect (feeling rating significantly greater than zero: Experiment 1: $t(31) = 4.11$, $p < .001$, $d = .73$; Experiment 2: $t(34) = 9.84$, $p < .001$, $d = 1.66$).

Surprisingly however, feelings were not significantly positive when observing allocators act generously - both when observing allocators give more than half the endowment (rating not significantly different from zero: $t(34) = 1.64$, $p = .11$, $d = .27$) and when observing them take less than half the endowment ($t(31) = 1.82$, $p = .079$, $d = .32$). In fact, observers report feeling *worse* when observing allocators act generously by giving more than half the endowment than when observing even splits ($t(34) = 3.94$, $p < .001$, $d = .94$), and there was a marginable effect in the same direction when allocators took less than half compared to when observing even splits ($t(31) = 1.89$, $p = .068$, $d = .40$). Participants did report more positive feelings when observing generosity than selfishness (former case $t(31) = 6.84$, $p < .001$, $d = 1.79$; latter case $t(34) = 3.94$, $p < .001$, $d = .96$).

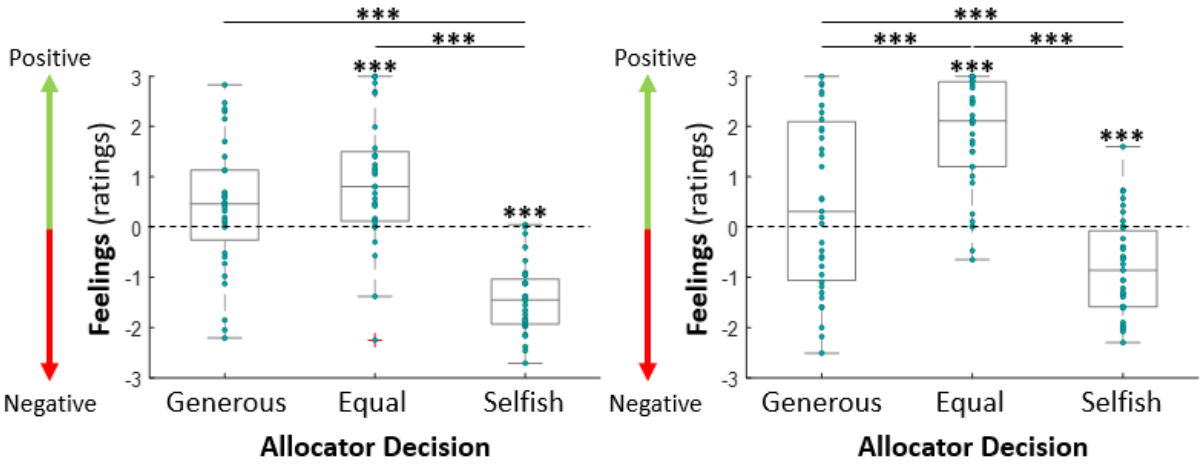

Figure 2. Observers are negatively affected when observing selfish behavior, positively affected when observing fair acts, and relatively unaffected when observing generosity. Observers report negative feelings when observing allocators act selfishly and positive feeling when observing equal splits. This is true both in Experiment 1 (left panel) and Experiment 2 (right panel). Feelings were not significantly positive nor negative when observing generosity. *** $p < .001$.

What is underlying this surprising pattern of results? We hypothesized that if participants were averse both to selfishness and inequity they may report feeling worse when observing unequal splits relative to even splits, even if unequal splits were a consequence of generous acts. To formally test this hypothesis, we quantified the influence of (un)selfishness and (in)equity on observers affect and characterized the computational rules by which features of observed acts are transformed into affective responses.

We operationalized selfishness (blue line, **Figure 3**) as the percentage of the endowment the allocator took/kept for themselves ranging from 0 (when the allocator was most generous, allocating nothing to themselves) to 100 (when the allocator was most selfish, allocating all to themselves). We operationalized inequity (orange line, **Figure 3**) as the absolute difference between the percentages of the endowment that each person was left with post allocation, ranging from 0 (when the split was 50/50) to 100 (when one person receives all and the other none).

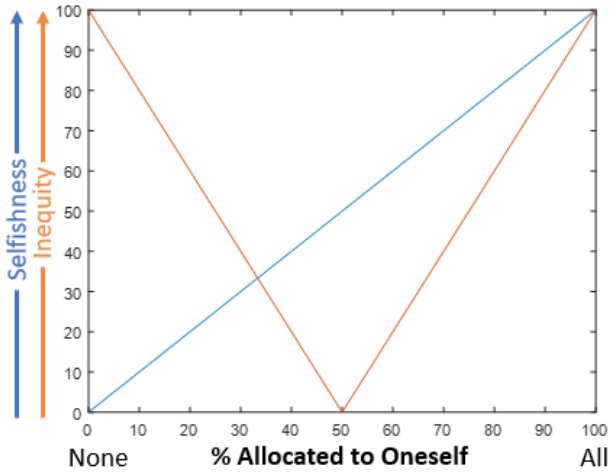

Figure 3. Operationalizing selfishness and inequity. Selfishness (blue line) is defined as the percentage of the endowment the allocator took/kept for themselves ranging from 0 (when the allocator was most generous, allocating nothing to themselves) to 100 (when the allocator was most selfish, allocating all to themselves). Inequity (orange line) is defined as the absolute difference between the percentages of the endowment each person was left with post allocation ranging from 0 (when the split was 50/50) to 100 (when one person receives all and the other none).

We then built seven computational models that aim to quantify the impact of selfishness and inequity on observers' feelings. Three models included both a selfishness aversion parameter and an inequity aversion parameter (1, 2, 3), while two models (4, 5) only included the former, and two models (6, 7) only the latter. To examine if (un)selfishness and (in)equity influenced participant's feelings in a non-linear fashion some models (1, 2, 4, 6) include a curvature parameter, ρ. All models included an Endowment term to account for the effect of the absolute amount of money available to the allocator.

No.	Model	BIC: Exp.1	BIC: Exp.2
1.	$\beta_0 + \beta_1 Selfishness^{\rho Selfishness} + \beta_2 Inequality^{\rho Inequality} + \beta_3 Endowment$	236.37	178.06

2.	$\beta_0 + \beta_1 Selfishness^{\rho} + \beta_2 Inequality^{\rho} + \beta_3 Endowment$	**196.33**	**156.92**
3.	$\beta_0 + \beta_1 Selfishness + \beta_2 Inequality + \beta_3 Endowment$	206.65	197.39
4.	$\beta_0 + \beta_1 Selfishness^{\rho} + \beta_2 Endowment$	264.89	247.58
5.	$\beta_0 + \beta_1 Selfishness + \beta_2 Endowment$	226.14	252.24
6.	$\beta_0 + \beta_1 Inequality^{\rho} + \beta_2 Endowment$	295.66	255.72
7.	$\beta_0 + \beta_1 Inequality + \beta_2 Endowment$	298.90	285.29

Table 2: Feelings model specifications.

Each of the above models were fit to each participant's (standardized) feelings ratings. Bayesian model comparisons indicated that the best-fitting model was Model 2 for both Experiments 1 and 2 (see **Table 1** for BIC values). This model includes both a selfishness aversion parameter (Experiment 1: β = -0.39, SE = .064, CI = [-0.51, -0.26]; Experiment 2: β = -0.15, SE = .024, CI = [-0.20, -0.11]) and an inequity aversion parameter (Experiment 1: β = -0.18, SE = .028, CI = [-0.24, -0.13]; Experiment 2: β = -.17, SE = .022, CI = [-0.22, -0.13]) indicating that selfishness aversion *and* inequity aversion both influence observer's affective responses. Moreover, the model included a curvature parameter (Experiment 1: ρ = .44, SE = .027, CI = [0.39, 0.49]; Experiment 2: ρ = .50, SE = .025, CI = [0.45, 0.55]) that translates to less sensitivity to each additional unit of selfishness or inequity (see Figure 4). That is to say, one unit has a greater impact on observer's response than the next unit and so forth.

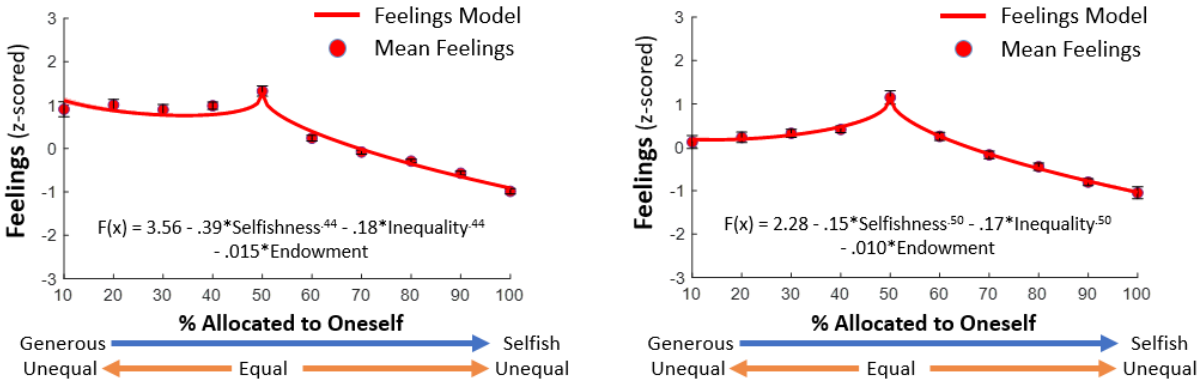

Figure 4. Modelling observers' feelings as a function of observed selfishness and inequity. Plotted is the winning model (Model 2), fit at the group-level, for Experiment 1 (left panel) and Experiment 2 (right panel). Participants' feelings ratings were z-scored before model-fitting to standardize responses. The predicted feelings from the model (red line) is overlaid on the mean observed feelings (red dots). Error bars represent SEM.

We used the above feelings function (which we built based on data from blocks 2 and 4 in Exp 1 and blocks 1 and 3 in Exp 2) to estimate how participants were likely feeling on blocks where we did not ask them about their feelings (that is blocks 1,3 in Exp 1 and 3,4 in Exp 2). We did

this simply by entering the selfishness value, inequity value and endowment value on that specific trial into the personalized feeling function of the participant – this generated the participant's predicted feeling on that trial. We then examined if the predicted feeling was related to the participant's decision to punish on that trial. If action is associated with affect the two values should correlate, and indeed they did (Mean Pearson Correlation Coefficient: Experiment 1: r = -0.68, p < .001. Experiment 2: r = -61, p < .001). Participants punished more heavily when the feelings function predicted a negative affective response and less heavily when the feelings function predicted a positive affective response.

Observer's decisions to punish are a function of both selfishness aversion and inequity aversion. Thus far we report that the observers' affective responses are influenced by observed selfishness and inequity. We next ask whether these same social values also drive observers' punishment behaviour. We expected this to be the case, because as we demonstrated above, estimated feelings generated by our feeling function were strongly related to participants' punishment decisions.

The results of a one-way repeated-measures ANOVA revealed the participants punished the three types of allocators differently (Experiment 1: $F(2,30) = 100.31$, $p < .001$, $\eta_p^2 = 0.87$; Experiment 2: $F(2,33) = 74.12$, $p < .001$, $\eta_p^2 = 0.82$). As one would expect, observers punished selfish allocators more than generous allocators (Experiment 1: $t(31) = 13.86$, $p < .001$, $d = 3.02$; Experiment 2: $t(34) = 8.82$, $p < .001$, $d = 2.18$) and allocators who split equally (Experiment 1: $t(31) = 13.81$, $p < .001$, $d = 2.91$; Experiment 2: $t(34) = 11.54$, $p < .001$, $d = 2.83$). There was no difference in the frequency of punishing generous allocators and those who split equally (Experiment 1: $t(31) = .41$, $p = .68$, $d = .041$; Experiment 2: $t(34) = 1.57$, $p = .13$, $d = .19$).

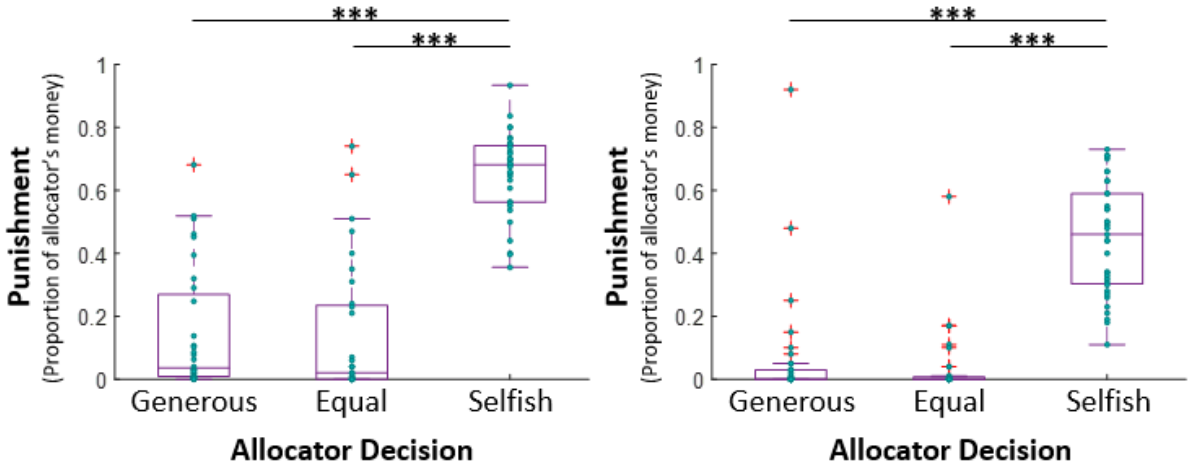

Figure 5. Observers' Punishment Choices. Participants punished selfish allocators significantly more than generous allocators and those who split resources equally in both Experiment 1 (left panel) and Experiment 2 (right panel). Error bars represent SEM, *** p < .001.

As portrayed in Figure 5, selfish decisions are punished more than other decisions. To formally examine whether selfishness aversion and/or inequity aversion underlie this behaviour we adopted the same approach we had above for analysing the feelings data. In particular, we fit each of the seven computational models described in Table 1 to each observer's (standardized) punishment choices. Bayesian model comparisons revealed that, as for feelings, Model 2

outperformed all other models in describing punishment decisions in both experiments (see Table 2 for BIC values). This model includes both a selfishness aversion parameter (Experiment 1: β = .57, SE = .087, CI = [0.40, 0.74]; Experiment 2: β = 0.052, SE = .011, CI = [0.031, 0.073]) and an inequity aversion parameter (Experiment 1: β = 0.17, SE = .025, CI = [0.12, 0.22]; Experiment 2: β = 0.028, SE = 0.0059, CI = [0.016, 0.040]) indicating that selfishness aversion *and* inequity aversion both influence observer's behaviour. Moreover, the model included a curvature parameter (Experiment 1: ρ = .42, SE = .025, CI = [0.37, 0.47]; Experiment 2: ρ = .81, SE = .041, CI = [0.73, 0.89]) (see Figure 5) indicating that one unit of selfishness or inequity had a greater impact on observer's behaviour than the next unit and so forth.

No.	Model	BIC: Exp.1	BIC: Exp.2
1.	$\beta_0 + \beta_1 Selfishness^{\rho Selfishness} + \beta_2 Inequality^{\rho Inequality} + \beta_3 Endowment$	209.01	207.41
2.	$\beta_0 + \beta_1 Selfishness^{\rho} + \beta_2 Inequality^{\rho} + \beta_3 Endowment$	**172.82**	**176.47**
3.	$\beta_0 + \beta_1 Selfishness + \beta_2 Inequality + \beta_3 Endowment$	179.39	180.33
4.	$\beta_0 + \beta_1 Selfishness^{\rho} + \beta_2 Endowment$	205.21	251.16
5.	$\beta_0 + \beta_1 Selfishness + \beta_2 Endowment$	200.31	225.24
6.	$\beta_0 + \beta_1 Inequality^{\rho} + \beta_2 Endowment$	306.12	299.77
7.	$\beta_0 + \beta_1 Inequality + \beta_2 Endowment$	304.60	306.46

Table 2: Punishment model specifications.

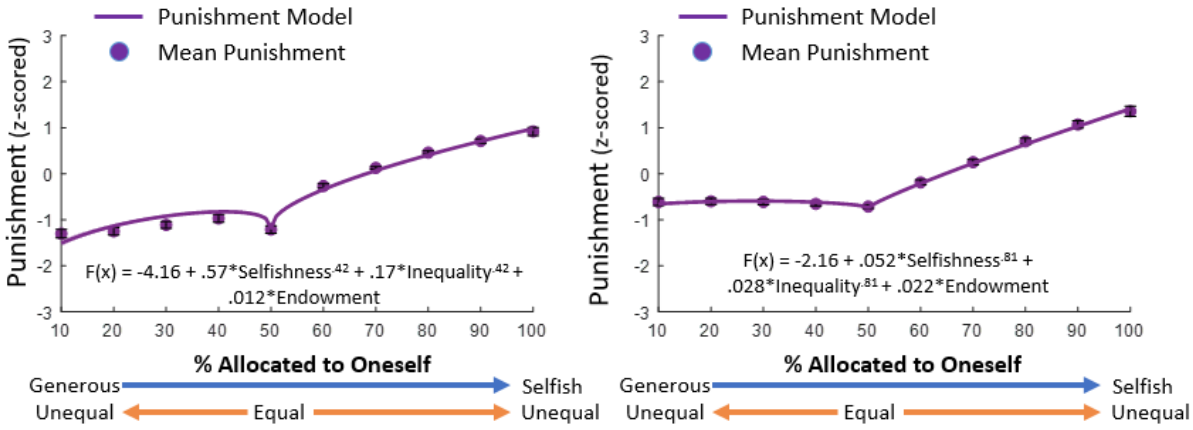

Figure 6. Observers' punishment decisions reflect selfishness aversion and inequity aversion. Plotted is the best-fitting model (Model 2), fit at the group-level, for Experiment 1(left panel) and Experiment 2 (right panel). Participants' punishment choices were z-scored before model-fitting to standardize responses. The predicted punishment from the model

(purple line) is overlaid on the mean observed punishment (purple dots). Error bars represent SEM.

Selfishness aversion and inequity aversion differentially impact feelings and actions. Our results suggest that selfishness and inequity are social values that influence affect and action when observing others give and take. However, it is possible that the two impact action and affect to different extents. Further, it is possible that the extent of the influence differs when observing others give rather than take. To formally test this, we calculated each participant's own feeling function and punishment function (analysis was conducted for participants for whom Model 2 provided a better fit to the data than a simple intercept model n = 64). We then examined the expected impact of selfishness on feelings and punishment decisions by setting the inequity value and intercept to zero. Specifically, we estimated the impact of selfishness on feelings/decisions for each level of selfishness and averaged those for each participant separately for feeling and action. We then did the same for the impact of inequity (i.e. setting the selfishness value to zero). Scores were entered into a 2 (response: feelings/punishment) by 2 (social norm: selfishness aversion/inequity aversion) repeated-measures ANOVA, with experiment framing (take/give) as a between-subjects factor.

The analysis revealed that selfishness aversion had a greater impact on participants' responses than inequity aversion (main effect: $F(1,62) = 11.92$, $p = .001$, $\eta_p^2 = .17$. The impact of selfishness aversion was greater when observing others take than give ($F(1,62) = 8.19$, $p = .005$, $\eta_p^2 = 0.12$), with no difference in inequity aversion under these two scenarios ($F(1,62) = 0.010$, $p = .92$, $\eta_p^2 < 0.001$). This interaction between social value and experiment framing (give/take) was significant ($F(1,62) = 8.67$, $p = .005$, $\eta_p^2 = .12$). Furthermore, inequity aversion had a stronger effect on feelings than decisions to punish ($F(1,62) = 8.86$, $p = .004$, $\eta_p^2 = 0.13$), while selfishness aversion had a marginally greater effect on decisions to punish than on feelings ($F(1,62) = 3.31$, $p = .073$, $\eta_p^2 = 0.051$). This interaction between response type and social value was significant ($F(1,62) = 4.22$, $p = .044$, $\eta_p^2 = 0.064$).

Together, these results suggest that the influence of selfishness aversion and inequity aversion is context dependent. First, selfishness aversion is stronger when observing an allocator take more than half the resources than keep more than half, despite the final resource distribution being the same. Second, inequity aversion impacts feelings more than action.

Observers' feel better about punishing allocators when those decisions align with their feelings about the allocation. The different extents to which inequity aversion and selfishness aversion impact feelings and action will at time result in discrepancies between how people feel about what they observe and whether they attempt to change the status quo. We hypothesized that when such discrepancies occur participants would not feel as good about their decisions to punish.

To test this we performed linear regressions for each participant that aim to account for participants' *feelings about their <u>own</u> punishment decisions* from (i) their *feelings about the <u>allocator's</u> decision*, (ii) the amount they punished, and (iii) the interaction between the two. This exercise required that for each trial we knew how the participant felt in response to the allocator's decision, whether and by how much they decided to punish and how they felt about their own decision. In Experiment 1 we indeed had 2 blocks were participants indicated all these variables (blocks 2,4). In experiment 2, however, participants rated how they felt about the allocators' decisions on two of the blocks (1,3), and made punishment decisions on the

other two blocks (2,4). Thus, we used our feelings function which was based on data from blocks 1,3 in Experiment 2 to estimate how participants felt about the allocator's decision on the punishment trials (Experiment 2: blocks 2,4) and entered these model-predicted feelings into the regression. We also controlled for the endowment amount which was added as an additional variable.

The results revealed a significant interaction between participants' feelings about the allocation and the amount they punished (Experiment 1: $\beta = -.31$, SE = .038, $t(31) = -8.09$, $p < .001$, 95% CI [-0.38, -0.23]; Experiment 2: $\beta = -.35$, SE = .049, $t(34) = -7.31$, $p < .001$, 95% CI [-0.45, -0.26], **Figure 7**). The interaction was due to participants feeling better about punishing when they felt negatively about the allocation compared to when they felt positively about it.

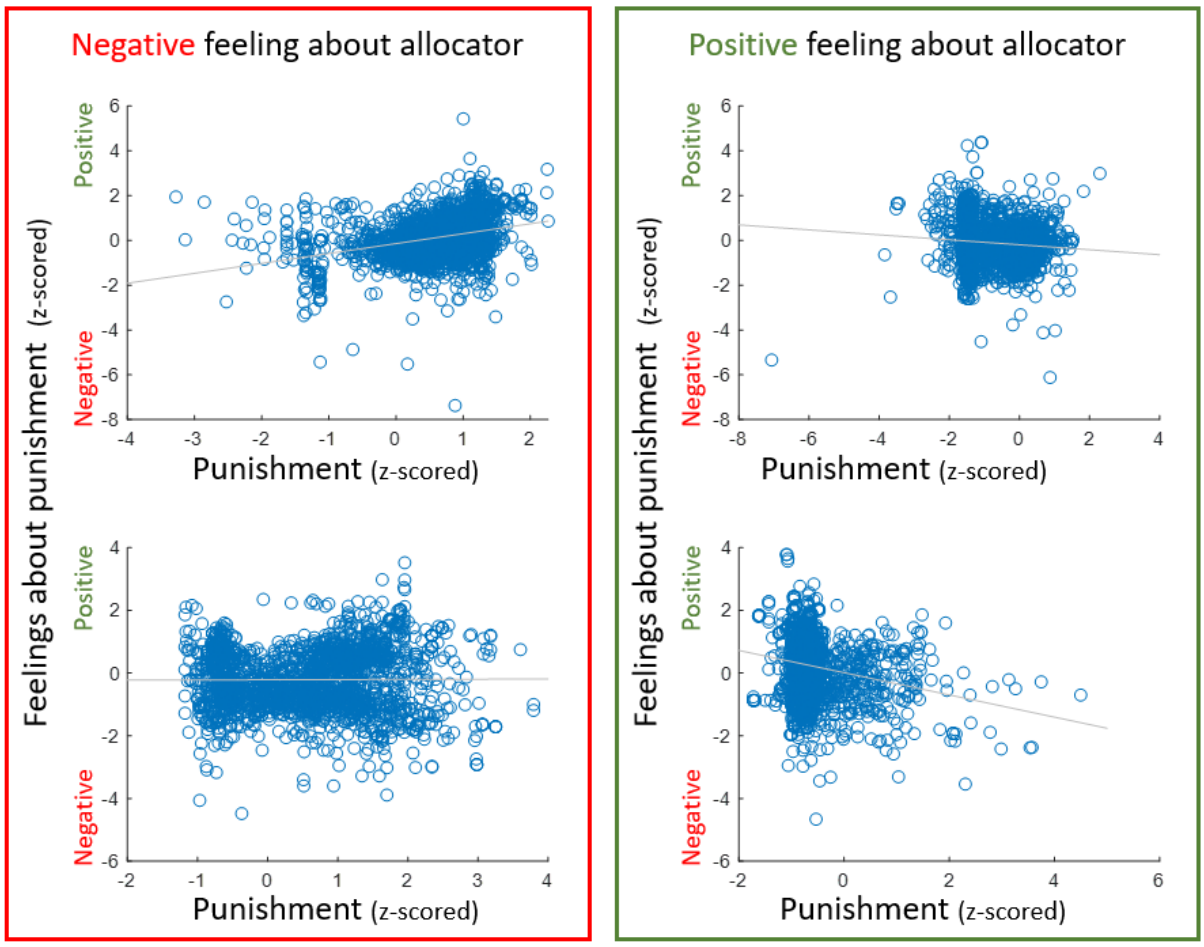

Figure 7. Participants' feelings about punishing depend on how they feel about the allocator. Plotted are participants feelings about their decision to punish (z-scored, y-axis) as a function of punishment magnitude (z-scored, x-axis). When participants felt negatively about the allocators' decision (left panel) greater punishment was related to more positive feelings about punishing in experiment 1 (top row) with no relationship in experiment 2 (bottom row). In contrast, when participant's felt positively about the allocators' decision (right panel) greater punishment was related to more negative feelings about punishing in experiment 1 (top row) and experiment 2 (bottom row). In both Experiments there was a significant interaction between amount punished and feelings about the allocation in explaining participants' feelings about punishing. Each dot represents data from one trial (these were entered into separate linear

regressions for each participant as described in the text). For experiment 2 we used the feeling function to predict how participants felt about the allocator's decision.

Discussion

In this study we characterize the rules by which other peoples' interactions are transformed into bystanders' reactions. First, we reveal diminishing marginal sensitivity to both selfishness and inequity. In particular, both observers' feeling and actions were governed by "selfishness aversion" (i.e., aversion to observing others allocate more resources to themselves rather than a third player) and "inequity aversion" (i.e., aversion to observing others allocate resources unequally, even when inequity is the consequence of generosity). Importantly, similar to the concept of 'diminishing marginal utility', according to which the first dollar earned has a greater impact on hedonic utility than the fifth (Kahneman & Tversky, 1979), each additional unit of observed selfishness/inequity had a smaller marginal impact than the last. While 'diminishing marginal utility' of rewards and losses on explicit feelings has been assumed by Prospect Theory (Kahneman & Tversky, 1979), only recently has direct evidence been reported (Charpentier et al., 2016). The current findings now suggest that this principle can be generalized beyond mere response to ones' own material rewards and losses, to responses to others' behaviour. These responses included both affective response and actions to alter the status quo.

Indeed, we report a strong association between observers' feelings and their actions. In particular, we used a 'feelings function' to predict observers' feelings in out-of-sample trials and found those estimated feelings to be correlated with observers' decisions to punish. The worse people felt about others' actions the more likely they were to punish them suggesting that observers' feelings influence decisions to change the status quo. Inequity, however, had a greater impact on feelings than on punishment choices. As a result, feelings and action did sometimes dissociate. In such cases, when observers' affect did not align with their action, the observers tended to feel worse about their own punishment decisions.

While it is well known that people engage in third-party punishment when others allocate selfishly (Balafoutas, Grechenig, & Nikiforakis, 2014; Bone, Silva, & Raihani, 2014; Charness, Cobo-Reyes, & Jiménez, 2008; De Quervain et al., 2004; Fehr & Gächter, 2002; Henrich et al., 2006; Jordan et al., 2016; Jordan, McAuliffe, & Rand, 2016; Raihani & Bshary, 2015; Henrich et al., 2010; Marlowe & Berbesque, 2008), bystanders explicit affective reaction to others' allocations were not previously reported (for hypothetical anticipated affect see Fehr & Fischbacher, 2004; and for affect manipulations see Nelissen & Zeelenberg, 2009). Our results show that observing selfishness elicits a negative affective reaction in observers. Interestingly, observing generosity - that is others allocate less than a fair share to themselves - elicited a more negative reaction than observing equal distributions. Thus, bystanders' affective reactions reflected the integration of "selfishness aversion", which would trigger a negative reaction in the former scenario and a positive reaction in the latter, and "inequity aversion", which would trigger a negative reaction in both (unequal) scenarios. Indeed, some studies have reported that participants punish others who behave generously (Herrmann, Thöni, & Gächter, 2008; Parks & Stone, 2010; Pfattheicher, Keller, & Knezevic, 2017; Pfattheicher, Landhäußer, & Keller, 2014; Pfattheicher & Schindler, 2015; Pleasant & Barclay, 2018; Sylwester, Herrmann, & Bryson, 2013). This, however, is observed predominantly in cultures with weak civic norms (Herrmann et al., 2008) and in people with a disposition to sadism (Pfattheicher, Keller, & Knezevic, 2017). Or results, suggest that while people do not often punish generous others, neither do they have a positive reaction to generous behaviour.

In sum, we provide evidence that bystanders integrate multiple social values to govern decisions to change the status quo. The observed integration in bystanders is reminiscent of that observed previously in people who were personally affected by an interaction (Gao et al., 2018; Sáez et al., 2015). Bystanders actions may be adaptive in the long run as they could shape the behaviour of other individuals in a group whom may interact with the bystander in the future. Our results further suggest that bystanders' decisions are tightly related to their affective reactions, with initial deviations from a social norm having greater relative influence than subsequent deviations. When bystanders do not act in alignment with their affective response their emotional state suffers, perhaps providing a learning signal for the bystander to change their actions in the future.

Materials and Methods

Experiment 1

Participants. Thirty-two participants completed the experiment (17 females and 15 males, aged 20-32 years M = 25.19, SD = 3.81). All were students recruited from the UCL Division of Psychology and Language Sciences' online subject pool, enrolled in courses other than economics and/or psychology. Participants were paid £12 for completing the experiment. The experiment was approved by the departmental ethics committee at University College London.

Procedure and Task Design. Participants came into the lab and were told that they would be observing simple distribution games between anonymous and varying pairs of players on Amazon's Mechanical Turk. In reality, the other "players" were not participants but algorithms.

The experiment consisted of 4 blocks of 60 trials each (Figure 1). On each trial participants observed as one player (Player B) was given a financial endowment (3sc), ranging from £1 to £15 (step size = £1). The other player (Player A, the allocator) then decided how much of Player B's money to take for themselves (jittered duration 2-5sc). Participants were then shown the amount taken by Player A and the amount left for Player B (1sc). The amount taken varied on each trial from 10% to 100% of Player B's money (step size = 10%).

On even blocks (i.e. blocks 2, 4) participants were asked to indicate "How does Player A's decision make you feel?" from -3 ("extremely negative") to 3 ("extremely positive"; self-paced). In all blocks (i.e. 1, 2, 3, 4) they were then asked "By how much would you like to penalize Player A?" on a scale from £0 ("No penalty") to the amount taken by Player A (e.g. £8; "Penalize by maximum amount"; self-paced). Punishment was not costly to the participant. Neither did the participant nor Player B gain from the punishment. Rather the amount punished was deducted from Player's A reward. On even blocks (i.e. blocks 2, 4) participants were then asked "How does your decision make you feel?" on a scale from -3 ("extremely negative") to 3 ("extremely positive"; self-paced).

This design allowed us to build a 'feelings function', which specified the relationship between others' acts and the observer's emotional reactions. We modelled each participant's feelings function from their feeling ratings in blocks two and four and used this function to estimate how participants likely felt on trials in odd blocks (i.e. blocks 1, 3; blocks in which they did not provide a feelings rating). We were able to then test the validity of these estimated feelings by examining the relationship between the estimated feelings and participants' punishment choices in blocks one and three. If participants acted in accordance with their feelings, punishing more when they felt negatively about the allocator's decision, we should expect to see a negative correlation between estimated feelings and punishment choices.

The initial endowment was pseudo-randomized so that all 15 endowment amounts (£1-£15) were presented 4 times in each block. The amount taken by the allocator (10-100%, step-size = 10%) was also pseudo-randomized. Outcomes where the allocator took more than 50% of the endowment were presented 9 times each, those where the allocator took 50% or less were presented 3 times each. This was done so that the players' behaviour would seem realistic.

Finally, participants completed a debriefing questionnaire (see Supplementary Information) consisting of a funneled debrief which gave them an opportunity to report any suspicions that the online players were bots. They then provided demographic information and completed a series of standardized self-report questionnaires. Upon completion participants were told the purpose of the study and informed that the other players were bots. The approximate duration of the experiment was 1.5 hours.

Experiment 2

Participants. The recruitment procedure and compensation were the same as for Experiment 1. Thirty-five participants completed the experiment (22 females and 13 males, aged 18-61 years M = 27.91, SD = 11.18). The experiment was approved by the departmental ethics committee at University College London.

Procedure and Task Design. The experiment consisted of 4 blocks of 60 trials each (Figure 1). On each trial participants observed as one player (Player B) was given a financial endowment (3sc), ranging from £1 to £15 (step size = £1). Player B (the allocator) then decided how much of this money to give to Player A (jittered duration 2-5sc). Participants were then shown the amount given to Player A and the amount kept by Player B (1sc). The amount given varied on each trial from 0% to 90% of Player B's money (step size = 10%).

On odd blocks (i.e. blocks 1, 3) participants were asked to indicate "How does Player A's decision make you feel?" from -3 ("extremely negative") to 3 ("extremely positive"; self-paced). On even blocks (i.e. 2, 4) participants were asked "By how much would you like to penalize Player A?" on a scale from £0 ("No penalty") to the amount taken by Player A (e.g. £2; "Penalize by maximum amount"; self-paced). As in Experiment 1, punishment was not costly to the participant. After indicating their punishment decisions in blocks 2 and 4, participants were then asked "How does your decision make you feel?" on a scale from -3 ("extremely negative") to 3 ("extremely positive"; self-paced).

We used participants' feelings ratings in blocks 1 and 3 to build a feelings function and subsequently estimated how participants felt on trials in even blocks (i.e. blocks 2, 4; blocks in which they did not provide a feelings rating). We then assessed the validity of the estimated feelings by examining whether they were predictive of participants punishment choices in blocks 2 and 4. In contrast to Experiment 1, in which participants' feelings about allocators' decisions and feelings about their own punishment decisions were measured in the same blocks, in Experiment 2 they were measured in separate blocks. This allowed us to provide further validation of our feelings function by testing whether the estimated feelings were able to explain variation in how participants felt about their own punishment decisions.

The initial endowment the allocator received was pseudo-randomized so that all 15 endowment amounts (£1-£15) were presented 4 times in each block. Contrary to Experiment 1, the distribution of allocator decisions was uniform across all outcomes. Each possible allocator decision (i.e. allocator giving 0-90%, step-size = 10%) was observed 6 times.

Finally, participants completed the debriefing questionnaire (see Supplementary Information) consisting of a funneled debrief which gave them an opportunity to report any suspicions that the online players were bots. They also indicated, post-task, what they would have done if they had been in the role of Player B (the allocator) and what distribution decision they thought was fair. They then provided demographic information and completed a series of standardized self-

report questionnaires. Upon completion participants were told the purpose of the study and informed that the other players were bots. The approximate duration of the experiment was 1.5 hours.

Data Analyses Experiment 1 & 2

Behavioural Data Analysis. For each participant, mean feeling ratings and punishment values were calculated separately for trials in which the allocator allocated the resources selfishly, generously and equally. Selfish allocations were classified as those were the allocator took/kept more than half the endowment; generous allocations were classified as those were the allocator took/kept less than half the endowment; and equal allocations as those where the allocator split the endowment equally. For each type of allocation, we performed one-sample t-tests to assess whether feelings were significantly different from zero (i.e. significantly positive or negative). We then performed a one-way repeated-measures ANOVA to assess whether participants felt more positively about some allocator decisions than others and, when significant, we followed with pairwise comparisons. We also performed a one-way repeated-measures ANOVA to assess whether participants punished some allocator decisions more than others and, when significant, we followed with pairwise comparisons.

Computational Modeling. To quantify the influence of (un)selfishness and (in)equity on observers' feelings and punishment choices we fit seven models (see Tables 1 & 2), using the fitnlm function in Matlab (version 2018a), separately for each participant's standardized feelings ratings and standardized punishment choices. For each model, Bayesian information criterion (BIC; Schwarz, 1978) and adjusted R^2 were computed. Given that the models differed in their number of parameters, BIC (rather than R^2), which penalizes models with additional parameters, was used to compare models. After identifying the best fitting model, we re-fit the model separately for experiment one and experiment two, with all trials from all participants included to estimate the group-level parameters.

No.	Model
1.	$\beta_0 + \beta_1 Selfishness^{\rho Selfishness} + \beta_2 Inequality^{\rho Inequality} + \beta_3 Endowment$
2.	$\beta_0 + \beta_1 Selfishness^{\rho} + \beta_2 Inequality^{\rho} + \beta_3 Endowment$
3.	$\beta_0 + \beta_1 Selfishness + \beta_2 Inequality + \beta_3 Endowment$
4.	$\beta_0 + \beta_1 Selfishness^{\rho} + \beta_2 Endowment$
5.	$\beta_0 + \beta_1 Selfishness + \beta_2 Endowment$
6.	$\beta_0 + \beta_1 Inequality^{\rho} + \beta_2 Endowment$
7.	$\beta_0 + \beta_1 Inequality + \beta_2 Endowment$

Table 1: Model specifications.

To further tease apart the effects of selfishness and inequity aversion on feelings and punishment we used the winning model to estimate the impact of each social value while setting

the other value, and the intercept, to zero. Specifically, we estimated the mean impact of selfishness aversion across the various levels of observed (un)selfishness (i.e. allocator taking/keeping 10% to 100% of Player B's money with step size of 10%) on feelings and punishment separately for each participant. We did the same for inequity aversion (setting selfishness to zero). These scores were then entered into a 2 (response: feelings/punishment) by 2 (social norm: selfishness aversion/inequity aversion) repeated-measures ANOVA, with experiment framing (take/give) as a between-subjects factor.

Out-of-Sample Prediction. We next investigated whether observers acted in accordance with their affective responses to unselfishness and inequity when deciding whether to punish. If these social values are translated into affective responses and punishment decisions in a similar manner, we would expect the feelings function's estimated values to be correlated with participants' punishment decisions. To test this, we used the winning feelings model to estimate how each participant felt about allocators' decisions in blocks where feelings ratings were not collected (Blocks 1 and 3 in Experiment 1; Blocks 2 and 4 in Experiment 2) and then correlated these estimated feelings with the participant's punishment choices in these blocks. We then tested whether the overall correlation between estimated feelings and punishment choices was significantly different from zero by performing a one-sample t-test.

Next, we assessed whether these estimated feelings could help explain variation in how participants felt about their own punishment decisions. We hypothesised that how individuals felt about their punishment choices may depend on how they felt about allocator's decision. Specifically, we hypothesised that observers would feel better about their punishment decisions if they acted in accordance with their feelings about the allocator's decision. To test this, we needed to know how the participant felt in response to the allocator's decision, whether and by how much they decided to punish and how they felt about their own decision. In Experiment 1 we indeed had 2 blocks were participants indicated all these variables (blocks 2, 4). We thus performed a linear regression for each participant, with the participant's feelings about their own punishment decisions (standardized) entered as the dependent variable; and (i) feelings about the allocator's decision (standardized) (ii) punishment (standardized) (iii) the interaction between the two and (iii) the endowment amount as independent variables.

In Experiment 2, participants rated how they felt about the allocators' decisions on two of the blocks (1,3), and made punishment decisions on the other two blocks (2,4). Thus, we used our feelings function which was based on data from blocks 1,3 to estimate how participants felt about the allocator's decision on the punishment trials in blocks 2,4. We then performed the same analysis as above but entered the participant's *estimated* feelings about the allocator's decision into the regression models.

Acknowledgments. We thank Jingxiu Cheng for assistance with data collection. This work was funded by a Wellcome Trust Senior Research Fellowship (to T.S.).

References

Balafoutas, L., Grechenig, K., & Nikiforakis, N. (2014). Third-party punishment and counter-punishment in one-shot interactions. *Economics Letters*, *122*(2), 308–310. https://doi.org/10.1016/j.econlet.2013.11.028

Bone, J., Silva, A. S., & Raihani, N. J. (2014). Defectors, not norm violators, are punished by third-parties. *Biology Letters*, *10*(7). https://doi.org/10.1098/rsbl.2014.0388

Carlsmith, K. M., Wilson, T. D., & Gilbert, D. T. (2008). The Paradoxical Consequences of Revenge. *Journal of Personality and Social Psychology*, *95*(6), 1316–1324. https://doi.org/10.1037/a0012165

Charness, G., Cobo-Reyes, R., & Jiménez, N. (2008). An investment game with third-party intervention. *Journal of Economic Behavior and Organization*, *68*(1), 18–28. https://doi.org/10.1016/j.jebo.2008.02.006

Charpentier, C. J., De Neve, J. E., Li, X., Roiser, J. P., & Sharot, T. (2016). Models of Affective Decision Making: How Do Feelings Predict Choice? *Psychological Science*, *27*(6), 763–775. https://doi.org/10.1177/0956797616634654

Crockett, M. J., Kurth-Nelson, Z., Siegel, J. Z., Dayan, P., & Dolan, R. J. (2014). Harm to others outweighs harm to self in moral decision making. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(48), 17320–17325. https://doi.org/10.1073/pnas.1408988111

Crockett, M. J., Siegel, J. Z., Kurth-Nelson, Z., Dayan, P., & Dolan, R. J. (2017). Moral transgressions corrupt neural representations of value. *Nature Neuroscience*, *20*(6), 879–885. https://doi.org/10.1038/nn.4557

Dawes, C. T., Fowler, J. H., Johnson, T., Mcelreath, R., & Smirnov, O. (2007). *Egalitarian motives in humans*. *Nature*, *446*(7137), 794-796. https://doi.org/10.1038/nature05651

De Quervain, D. J. F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004). The neural basis of altruistic punishment. *Science*, *305*(5688), 1254–1258. https://doi.org/10.1126/science.1100735

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*(2), 63–87. https://doi.org/10.1016/S1090-5138(04)00005-4

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

Fliessbach, K., Phillipps, C. B., Trautner, P., Schnabel, M., Elger, C. E., Falk, A., & Weber, B. (2012). Neural responses to advantageous and disadvantageous inequity. *Frontiers in Human Neuroscience*, *6*, 165. https://doi.org/10.3389/fnhum.2012.00165

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*(33), E7680–E7689. https://doi.org/10.1073/pnas.1802523115

Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., … Ziker, J. (2010). Markets, religion, community size, and the evolution of fairness and punishment. *Science*, *327*(5972), 1480–1484. https://doi.org/10.1126/science.1182238

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., … Ziker, J. (2006). Costly punishment across human societies. *Science*, *312*(5781), 1767–1770. https://doi.org/10.1126/science.1127333

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science*, *319*(5868), 1362–1367. https://doi.org/10.1126/science.1153808

Jordan, J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature*, *530*(7591), 473–476. https://doi.org/10.1038/nature16981

Jordan, J., McAuliffe, K., & Rand, D. (2016). The effects of endowment size and strategy method on third party punishment. *Experimental Economics*, *19*(4), 741–763. https://doi.org/10.1007/s10683-015-9466-8

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis Of Decision Under Risk. *Econometrica 47*(2), 363-391. https://doi.org/10.2307/1914185

Marlowe, F. W., & Berbesque, J. C. (2008). More "altruistic" punishment in larger societies. *Proceedings of the Royal Society B: Biological Sciences*, *275*(1634), 587–590. https://doi.org/10.1098/rspb.2007.1517

Nelissen, R. M. A., & Zeelenberg, M. (2009). Moral emotions as determinants of third-party punishment: Anger, guilt, and the functions of altruistic sanctions. *Judgment and Decision Making*, *4*(7), 543–553.

Parks, C. D., & Stone, A. B. (2010). The desire to expel unselfish members from the group. *Journal of Personality and Social Psychology*, *99*(2), 303–310. https://doi.org/10.1037/a0018403

Pfattheicher, S., Keller, J., & Knezevic, G. (2017). Sadism, the intuitive system, and antisocial punishment in the public goods game. *Personality and Social Psychology Bulletin*, *43*(3), 337–346. https://doi.org/10.1177/0146167216684134

Pfattheicher, S., Landhäußer, A., & Keller, J. (2014). Individual Differences in Antisocial Punishment in Public Goods Situations: The Interplay of Cortisol with Testosterone and Dominance. *Journal of Behavioral Decision Making*, *27*(4), 340–348. https://doi.org/10.1002/bdm.1811

Pfattheicher, S., & Schindler, S. (2015). Understanding the Dark Side of Costly Punishment: The Impact of Individual Differences in Everyday Sadism and Existential Threat. *European Journal of Personality*, *29*(4), 498–505. https://doi.org/10.1002/per.2003

Pillutla, M. M., & Murnighan, J. K. (1996). Unfairness, anger, and spite: Emotional rejections of ultimatum offers. *Organizational Behavior and Human Decision Processes*, *68*(3), 208–224. https://doi.org/10.1006/obhd.1996.0100

Pleasant, A., & Barclay, P. (2018). Why Hate the Good Guy? Antisocial Punishment of High Cooperators Is Greater When People Compete To Be Chosen. *Psychological Science*, *29*(6), 868–876. https://doi.org/10.1177/0956797617752642

Posner, E. A., & Sunstein, C. R. (2017). Moral Commitments in Cost-Benefit Analysis. *Virginia Law Review*, *103*. Retrieved from https://heinonline.org/HOL/Page?handle=hein.journals/valr103&id=1859&div=47&collection=journals

Raihani, N. J., & Bshary, R. (2015). Third-party punishers are rewarded, but third-party helpers even more so. *Evolution*, *69*(4), 993–1003. https://doi.org/10.1111/evo.12637

Rosas, A., & Koenigs, M. (2014). Beyond "utilitarianism": Maximizing the clinical impact of moral judgment research. *Social Neuroscience*, *9*(6), 661–667. https://doi.org/10.1080/17470919.2014.937506

Sáez, I., Zhu, L., Set, E., Kayser, A., & Hsu, M. (2015). Dopamine modulates egalitarian behavior in humans. *Current Biology*, *25*(7), 912–919. https://doi.org/10.1016/j.cub.2015.01.071

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*(5626).

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Sunstein, C. R. (2019a). Cost-Benefit Analysis, Who's Your Daddy? *Journal of Benefit Cost Analysis*, *7*(1), 107–120. https://doi.org/10.1017/bca.2016.1

Sunstein, C. R. (2019b). The Cost-Benefit Revolution. Cambridge, MA: MIT Press.

https://doi.org/10.7551/mitpress/11571.001.0001
Sylwester, K., Herrmann, B., & Bryson, J. J. (2013). Homo homini lupus? Explaining antisocial punishment. *Journal of Neuroscience, Psychology, and Economics*, *6*(3), 167–188. https://doi.org/10.1037/npe0000009
Van't Wout, M., Kahn, R. S., Sanfey, A. G., & Aleman, A. (2006). Affective state and decision-making in the Ultimatum Game. *Experimental Brain Research*, *169*(4), 564–568. https://doi.org/10.1007/s00221-006-0346-5
Yu, R., Calder, A. J., & Mobbs, D. (2014). Overlapping and distinct representations of advantageous and disadvantageous inequality. *Human Brain Mapping*, *35*(7), 3290–3301. https://doi.org/10.1002/hbm.22402
Zheng, Y., Yang, Z., Jin, C., Qi, Y., & Liu, X. (2017). The influence of emotion on fairness-related decision making: A critical review of theories and evidence. *Frontiers in Psychology*, *8*(SEP), 1–10. https://doi.org/10.3389/fpsyg.2017.01592

Supplementary Information

Part 1 (funnelled debrief).

Participants were asked:

Were the instructions for all four sessions clear? (Yes/No). If No, please specify.

- 100% of the participants answered yes.

Were you at any stage confused or didn't know what to do to perform the task? (Yes/No). If Yes, please specify.

- 100% of the participants answered no.

What do you think the purpose of this experiment is?

- 16 participants reported that they thought the purpose of the study was related to fairness judgements; 5 participants thought the purpose was to examine the link between emotion and decision making; 2 participants suggested that we were interested in both fairness and emotion; 3 participants said they didn't know; and 6 participants provided some other response.

Did you participate in a similar version of this experiment (observation of distribution games with option to interact)? (Yes/No). If Yes, do you know what the purpose of the experiment was? If Yes, did your previous experience affect the decisions you made in today's experiment? (Yes/No).

- 31 participants said 'no', 1 participant said 'yes'. The participant who said 'yes' reported that they did not know the purpose of the experiment.

Have you read or learned about similar games previously? (Yes/No). If Yes, do you know which games? If Yes, did your knowledge affect the decisions you made in this game? (Yes/No).

- 30 participants said 'no', 2 participants said 'yes'. The participants who said 'yes' could not name a specific game and neither thought their knowledge affected the decisions they made in the current experiment.

Overall, was the study what you had expected? If No, please explain why:

- 28 participants said 'yes', 4 participants said 'no'. Explanations did not relate to any of the main research questions of this study.

Which of the following factors contributed to your decision to penalize player A (multiple answers possible)?

Answer	**Count**
Minimize Player A's payoff	8

Have equal payoffs for Player A and Player B	22
Punish Player A	14
Have not penalized player A at all	0
Other	4

Which of the following factors contributed to your motivation to penalize player A (multiple answers possible)?

Answer	**Count**
Empathy for player B	17
Aversion against player A	9
Care for fairness	22
Have not penalized player A at all	0
Other	0

Overall, do you think your reasons to penalize player A changed throughout the task? (Yes/No). If Yes, please specify.

- 26 participants said ‘no’, 6 said ‘yes’. Those who said yes suggested that their reasons may have changed because they wanted to be fair on the given trial or because they were using a new strategy.

Did you, at any point throughout the experiment, think that the experimenter had deceived you in any way? (Yes/No). If Yes, please specify.

- 25 participants said ‘no’, 7 said ‘yes’. 6 out of the 7 who said ‘yes’ reported that they suspected that they were not playing the game online with real people.

(Excluding the participants who suspected that they were not playing the game online with real people from the analyses did not affect any of the results reported in the main text.)

Finally, participants were asked if they had any final comments for the researchers.

Experiment 2

Part 1 (funnelled debrief).

Participants were asked:

Were the instructions for all four sessions clear? (Yes/No). If No, please specify.

- 100% of the participants answered yes.

Were you at any stage confused or didn’t know what to do to perform the task? (Yes/No). If Yes, please specify.

- 33 of the participants answered ‘no’, 2 answered ‘yes’. One of those who answered ‘yes’ was unsure whether they were seeing the same players in every trial, while the other was unsure what was meant by the feelings questions.

What do you think the purpose of this experiment is?

- 16 participants reported that they thought the purpose of the study was related to fairness judgements and/or decision making; 6 participants thought the purpose was to examine the link between emotion and decision making; 2 participants suggested that we were interested in both fairness and emotion; 2 participants said they didn’t know; and 9 participants provided some other response.

Have you read or learned about similar experiments before? (Yes/No).

- 32 participants said ‘no’, 3 participants said ‘yes’.

Have you previously participated in a similar version of this experiment before? (Yes/No). If Yes, do you know which games? If Yes, Do you know what the purpose of the previous experiment you participated in was?

- 34 participants said ‘no’, 1 participant said ‘yes’. The participants who said ‘yes’ said they thought the previous study was related to autism.

Did you, at any point throughout the experiment, think that the experimenter had deceived you in any way? (Yes/No). If Yes, please specify.

- 27 participants said ‘no’, 8 said ‘yes’. 6 out of the 8 who said ‘yes’ reported that they suspected that they were not playing the game online with real people.

(Excluding the participants who suspected that they were not playing the game online with real people from the analyses did not affect any of the results reported in the main text.)

Imagine you were assigned to be Player A and given £10 by the experimenter. How much of that amount would you share with Player B?

- Mean = £4.03, SD = 1.76, Range = 0-5

What do you think is the fairest amount for Player A to give to Player B in this situation?

- Mean = £4.69, SD = .90, Range = 1-5

Did you penalize the Player As who kept all or most of the money for themselves? If Yes, what motivated your decision to penalize the Player As who kept all or most of the money for themselves?

- 27 participants said ‘yes’, 1 said ‘no’, 7 said ‘sometimes’. Of those who said ‘yes’, 15 cited fairness as the main motivator driving their punishment, 7 said selfishness or greediness, 4 said to create equality, 1 said both unfairness and selfishness, and 5 provided some other moral argument.

Did you penalize the Player As who gave away all or most of the money? If Yes, what motivated your decision to penalize the Player As who gave away all or most of the money?

- 6 participants said ‘yes’, 25 said ‘no’, 4 said ‘sometimes’. Of those who said ‘yes’, 6 said they punished because the allocator made a bad or stupid decision, 1 said that over generosity might make the co-player feel guilty, 1 said it was unfair, 1 cited equality, and 1 did not give a reason.

One participant in experiment one had to restart the task, and one participant in experiment two did not complete all of the trials in one of the feelings blocks, due to technical errors. These participants were included in the analyses reported in the main text as we were still able to fit models to their data. Excluding them did not affect any of our results.