PEER-REVIEWED

Accepted in *Cognition* (in press)

The Role of Attention and Frames on Third-Party Punishment and Compensation Choices

Claudia Civai[1,2*], Valerio Capraro[3], and Luca Polonio[4]

[1] Department of Psychology and Neuroscience, School of Health and Medical Sciences, City St. George's, University of London, UK

[2] Division of Psychology, School of Health and Applied Sciences, London South Bank University, UK

[3] Department of Psychology, Universita' degli Studi di Milano – Bicocca, Italy

[4] Department of Economics, Management and Statistics, Universita' degli Studi di Milano – Bicocca, Italy

Word count, including abstract: 13400

Author Note

We have no known conflict of interest to disclose.

* Correspondence concerning this article should be addressed to Claudia Civai, Department of Psychology and Neuroscience, School of Health and Medical Sciences, City St. George's, University of London, Northampton Square, London EC1V 0HB, United Kingdom.

Email: Claudia.Civai@citystgeorges.ac.uk

Acknowledgments

The authors are grateful to Paige Johns, Vassilis Sideropoulos, Oliver Summer and Bindiya Thapa for their help in building the lab-based eye-tracking paradigm and for collecting the data, and to Ellis Keene and Ella Barry for their help in building the online eye-tracking paradigm.

The research was funded by an internal Seed Corn Funding grant awarded to the Claudia Civai from London South Bank University.

A non-peer review preprint of this work is available here https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4458455

The Role of Attention and Frames on Third-Party Punishment and Compensation Choices

Abstract

People often forgo their own self-interest to react to fairness and justice violations, even when not directly affected by the infraction. There are different ways to react to an injustice: some may prefer to punish the perpetrator, and others to compensate the victim. Here, our focus is on the role played by attention to determine these choices, investigating the relationship between attentional mechanisms and punishment/compensation in five preregistered experiments (N=1,157). Two eye-tracking experiments showed that people who focus more on the offender's payoff are more likely to punish, and when an exogenous stimulation increases the focus on the offender's payoff, people spend more to punish. An offender bias was also found, meaning that people, overall, prefer to focus on the offender's, rather than the victim's, payoff, and punish more than compensate. This was confirmed in three behavioural experiments, where people were exposed to either the offender's or the victim's payoff: when given the choice, people prefer to reveal the offender's payoff, and then punish; however, when randomly exposed to the victim's payoff, the preference for punishment disappears. Affective empathy boosts this effect: higher empathy leads to more punishment (or compensation) when the offender's (or victim's) payoff is revealed. These findings suggest that, whilst people have an intrinsic motivation to search for information that matches their preference (i.e., the offender's payoff and punishment), when exposed to an alternative piece of information (i.e., the victim's payoff), they modify their behaviour. Implications for understanding information bubbles and ways to overcome them are discussed.

Keywords: injustice perception; third-party punishment and compensation; attentional processes; information frames; choice process.

1. Introduction

Decades of psychological, neuroscientific, and behavioural economic research have demonstrated that, when we are exposed to unfairness or injustice, although we can certainly decide to do nothing at all, many of us decide to sacrifice something to counteract the violation (Fehr & Fischbacher, 2004). There are different ways to react to an injustice: some may prefer to punish the perpetrator, and others to compensate the victim. Whilst both behaviours signal a willingness to react to injustice and therefore, from this perspective, serve a similar purpose, they differ in their consequences, in that punishment harms, whereas compensation helps. For example, when it comes to choosing how public funds should be employed, a preference for punishment may lead people to prioritise the implementation of harsher policies for tackling violent crime or harsher prison regulations; on the other hand, a preference for compensation may be associated with higher support for programmes that help the victims of these violent crimes. Investigating the prevalence of one strategy over the other, as well as the psychological underpinnings and the individual differences that may explain these preferences, is important to understand the factors that drive moral decision-making and prosocial behaviour. Moreover, clarifying whether and how easily these choices can be manipulated can have important implications in different contexts. In fact, attention is a limited resource, and the media are constantly competing to capture a piece of it: this phenomenon, known as attention economy (Stroud, 2017), has fundamentally shaped how news is delivered and consumed. This almost relentless exposure to information, even when we are not actively seeking it (Weeks & Lane, 2020), likely influences our attitudes and judgments in ways that often go unnoticed. Understanding the fundamental attentional mechanisms that shape our socio-moral choices is therefore crucial for increasing awareness of how our perspectives are formed. For example, to what extent can our physical environment, including social media or online news outlets, shape our attitudes toward retributive (punish offenders) or restorative (compensate victims) justice?

1.1 Third-party punishment and compensation

Previous research has returned mixed findings on which action is preferred: some studies found that people prefer punishment over compensation (FeldmanHall et al, 2014; Stallen et al, 2018), whilst others found either the opposite, with compensation consistently preferred over punishment even in the presence of repeated unfairness by the hand of a single offender (Chavez & Bicchieri, 2013; Van Doorn & Browers, 2017; Van Doorn et al, 2018), or no clear preference (Hu et al, 2015; Civai et al, 2019). It has been suggested that punishment may be a more emotional and rewarding reaction compared to compensation (Kühne et al, 2015). Evidence of activation of neural reward areas, specifically ventral striatum, when participants chose to punish, seems to support this idea (Stallen et al, 2018). However, further findings showed that both punishment and compensation, compared to no reaction, are associated with a similarly higher activation of the striatum, and therefore interpreted as rewarding, depending on individual preferences (Civai et al, 2019). Other studies found that negative emotions are associated with these choices as well: anger at justice violations, and in particular moral outrage, predicts compensation as well as punishment (Lotz et al, 2011; Thulin & Bicchieri, 2016). However, a recent study showed that induced acute stress increases compensation and reduces punishment, an effect that is mediated by the activation of the emotional salience neural network (Wang et al, 2024). Compensation and punishment can also be perceived as a signal of trustworthiness, in that they show the willingness to act prosocially (Nelissen, 2008); interestingly, evidence shows that the perception of trustworthiness of an individual increases when they pick compensation, or helping behaviour, over punishment (Jordan et al, 2016). Overall, these contrasting results suggest that these preferences are underpinned by a complex interaction of psychological processes, where context may also play a role, as explained in the next paragraphs.

1.2 Attention and choice

The main goal of the present work is to investigate whether context-specific factors, such as the way in which the information is presented, also referred to as frame, play a role in determining preferences for punishment and compensation. Previous research has shown that manipulating people's top-down attention, which is intrinsic and goal-directed, and usually driven by endogenous motivational factors, by explicitly asking them to concentrate on the offender/victim, increases the preference for punishment/compensation (Gromet & Darley, 2009; Kühne et al, 2015; David et al, 2017). It remains less clear whether a more automatic, bottom-up, attentional mechanism, driven exogenously by the environment, may also affect the choice of strategy: if, through a subtle stimulus, attention were to be exogenously redirected towards the offender or the victim, would the choice be influenced accordingly? Some studies on moral decision-making and eye-tracking seem to suggest that this is the case. For example, Pärnamets and colleagues (2015) presented participants with morally charged statements, such as "murder is justifiable", and then manipulated the length of exposure to the two options available to participants to respond, i.e., "never" or "sometimes". The results showed that the option that was presented on the screen the longest was the one most likely to be chosen. Similarly, Ghaffari and Fiedler (2018) presented participants with statements such as "If I saw a stranger on the street struggling with her grocery bags, I would help her carry them", and then two options (e.g., "Only if I have time" and "I would usually help"). The results demonstrated that manipulating the position of the last visual fixation by interrupting the decision process and forcing a choice would affect said choice. Here, we aim to extend the investigation of the effects of implicit attention manipulation by testing whether redirecting attention exogenously and implicitly, while holding available information constant, can alter individuals' preferences to either punish or compensate.

1.3 Empathy and third-party punishment and compensation.

A secondary aim of this study was to investigate the effects of empathy, considered a trait that could explain some of the individual variance observed in participants' choices in terms of attentional focus (offender or victim) and action (punish or compensate). Empathy has been considered as one of the key mechanisms to mediate sensitivity to injustice (Decety & Yoder, 2015), and this trait has also been associated specifically with punishment and compensation preferences. Empathy is a multifaceted mechanism characterised by different components that allow us both to understand (cognitive empathy) and to share (affective empathy) others' feelings and emotional experience (Reniers et al, 2011). Affective empathy, specifically empathic concern (subscale of the Interpersonal Reactivity Index, (Davis, 1983)), was found to predict compensation (Leliveld et al, 2012; Hu et al, 2015) and costly altruistic behaviour (FeldmanHall et al, 2015). However, other findings suggest that this trait predicts punishment rather than compensation (Lu & McKeown, 2018), or even predicts both (Will et al, 2013); moreover, Hu et al (2020) found that, when instructed to focus on the offender, higher empathic individuals punished more than they compensated. On the other hand, the role of cognitive empathy has not been widely investigated: Decety and Yoder (2015) found that perspective taking, associated with cognitive empathy, predicted a higher sensitivity to injustice experienced by others, and Lu and McKeown (2008) found that higher perspective taking predicted compensation rather than punishment. Here, we wanted to control for this individual trait working as a potential explanatory variable for our behaviours of interest, as well as clarifying the role of different empathy components to address the contrasting findings regarding affective empathy and the gap in the literature regarding cognitive empathy.

1.4 Our contribution

We conducted five experiments to investigate the influence of context on third-party punishment and compensation preference specifically looking at how this choice is affected by the way in which the information is framed; in three of these five experiments, we also looked at the role played by affective and cognitive empathy. In all five experiments, we captured third-party preferences by measuring people's behaviour in a third-party game that we call Third-Party Justice Game (TPJG) (Stallen et al, 2018; Civai et al, 2019; Civai et al, 2020): in each trial of the game, participants are presented with two payoffs, player A's and player B's, one of which (player A's) can be much higher as a consequence of A taking from B. Participants need to decide whether to spend some of their own allocated money to either punish player A (the offender) or compensate player B (the victim).

1.4.1 Experiments 1-2: attentional attractors and choice

In the first two experiments, we used eye-tracking to investigate the relationship between attention and choice, and the effects of task-unrelated contextual elements, i.e., how payoffs are represented, on these choices. The idea is grounded on a rich body of literature linking attention and choice in different decision making tasks including value based decision-making (e.g., Armel et al, 2008; Krajbich et al, 2010; Pittarello et al, 2016; Teoh et al, 2020; Zonca et al, 2019), lotteries (e.g Arieli et al,2011; Alós-Ferrer et al. 2021; Alós-Ferrer & Ritschel 2022), and social games (e.g., Fiedler et al, 2013; Jiang et al, 2016; Marchiori et al, 2021; Polonio et al, 2015; Polonio & Coricelli, 2019; Stewart et al, 2016; Zonca et al, 2020a). These findings show that descriptive features of the choice options, such as the presence of salient stimuli that act as attractors (Devetag et al, 2016) or the complexity of the decision environment (Zonca et al, 2020b), can influence attentional patterns, and that measuring these patterns provides extremely useful insights into how people process information and eventually make decisions, which might be driven by these attentional processes (e.g., Milosavljevic et al, 2011; Ghaffari & Fiedler, 2018; see Rahal & Fiedler, 2019 for an overview of the benefits of eye-tracking in social psychological research). Here, we aim to add to this literature by testing whether including an attentional attractor within the visual presentation of the information impacts punishment/compensation preferences.

1.4.2 Experiments 3-5: information frame, empathy, and choice

In experiments 3-5, we aimed to investigate the automatic effects on choice of redirecting attention toward different sources of information, as well as control for and further explore the role played by affective and cognitive empathy in explaining these choices. We used information selection to build the frame: in some trials, only the information on the offender's payoff was presented, whilst in other trials we only presented information on the victim's payoff, so that participants were exogenously driven to focus only on one side of the outcome (offender's or victim's). The idea is based on the abundance of evidence on framing effects (see, for example, Beratšová et al, 2016), whereby the amount and type of information presented influence our decisions, such as during news consumption: news framing literature shows the powerful effect of journalistic choices and information selection on the public's moral judgment and attitude formation around any narrated issue, from immigration (Lecheler et al, 2015), to corporate crisis (Kim & Cameron, 2011), to gun violence (Liu et al, 2019) (see Lecheler & De Vreese, 2019 for a review). Our study aims to add to this topic by understanding the effect of information framing driven by automatic attentional manipulation on the choice in a third-party punishment and compensation paradigm that uses stimuli with minimal amount of information to minimise the influence of top-down goal-directed attention; in addition to this, we aim to shed light on the role played by different empathy components on these choices.

Specific hypotheses for each experiment are reported in the sections below.

1.5 Transparency and openness.

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in all five experiments. The design, hypotheses, materials and method, and analysis plans for all five experiments were pre-registered on the OSF:

Experiment one: https://osf.io/q6jyd/?view_only=0b7a51b22709433d9bb3444447b8e270

Experiment two: https://osf.io/23xau/?view_only=1b82988685c94b6a9ef141a35ac10b91

Experiments three and four: https://osf.io/2pvxz/?view_only=254f9f55277942969f6db9660882838a

Experiment 5: https://osf.io/p8k6r/?view_only=0da89bfdcd5946baa7b01c4b4de1648f

The data and analysis scripts for all five experiments are available here: https://osf.io/5egx8/files/osfstorage?view_only=b8a20e04eeb345e486b19a58cd9ee224

2. Experiment One: Attentional Correlates and Lab-Based Eye-Tracking data

We ran the first lab-based experiment to evaluate the influence of task-unrelated contextual elements on participants' decisions and uncover the relationship between attention and choices in the TPJG task. To do so, we developed a new version of the TPJG, which consisted of two conditions: in one condition (coins), the offender's and victim's payoffs were represented as piles of coins, as in Stallen et al (2018) and Civai et al (2019); in a second condition (digits), the payoffs were represented as numbers. We used eye-tracking to understand whether the choice to punish or compensate could be predicted by the amount of attention participants directed towards the offender's or the victim's payoffs, and, if so, whether manipulating the amount of attention towards one target (i.e., the offender) would influence choice. The rationale behind this design is that, in the coins condition, the difference between the amounts obtained by the two players becomes a salient feature that is automatically encoded. This hypothesis is supported by evidence suggesting that visual salience results from an interaction between a stimulus and other stimuli: during the initial stage of visual processing, which is thought to be automatic and preattentive, salience is driven by simple sensory features such as differences in colour, size, and form (e.g., Jarvenpaa, 1990; Itti & Koch, 2000). Here, whenever the offender takes chips from the victim, their piles of coins grow higher whilst the victim's shrink; therefore, the offender's payoff always appears visually richer compared to that of the victim's. We expect this visual contrast to automatically attract the participant's attention toward the more visually prominent element in the scene (see Figure 1). In the digits condition, on the other hand, the difference between the amounts obtained by the two players cannot be automatically encoded, and there is no visual stimulus that may serve as an attractor. If the offender's payoff in the coins condition serves as an attractor, it should be the first piece of information to be examined. Additionally, we should observe an increase in the time spent by the participant looking at the offender's payoffs compared to the victim's (Devetag et al, 2016; Li & Camerer, 2022). Therefore, if there is an effect of attention on moral behaviour, this in turn should translate into a higher probability of punishing rather than compensating. Following this line of thoughts, we hypothesise that:

1. The difference in the duration of gaze directed towards the offender versus the victim is larger in the coins condition (higher difference in visual saliency) compared to the digits condition;
2. The more participants look at the offender's payoff, as opposed to the victim's, the more likely they are to punish, as opposed to compensate, and the more they spend on punishment;

3. Participants are more likely to punish, and spend more to punish, in the coins condition compared to the digits condition.

The study received ethical approval from the Ethics committee at the corresponding author's institution, with protocol number [MASKED FOR PEER REVIEW].

2.1 Method

2.1.1 Participants

Thirty-seven participants, mostly undergraduate students, took part in the experiment (28 females, 9 males, mean age = 28.5); one participant was excluded because of a technical error, and the data were not saved. Overall, the data from 36 participants were analysed. Sample size was determined before any data analysis based on effect size: specifically, d=0.54 is the effect size of an independent sample t-test comparing the difference between the percentage of punishing and compensating choices, i.e., our effect of interest, in a previous pilot (coins vs digits, between-participant design). Given that this analysis assumes complete independence between the measures (between-participant design), it is more conservative when estimating sample size from effect size.

2.1.2 Materials

The main structure of the TPJG was based on tasks already published (Stallen et al, 2018; Civai et al, 2019), and was similar to a Dictator Game as considered for example by Krupka and Weber (2013): two players A and B start each round sharing equally a sum of money and one of the players (player A, or offender) has the opportunity to take from the other (player B, or victim). To this basic game, we add third-party punishment/compensation: participants, who play as observers, see this new distribution, and must decide whether to do nothing or to spend their money to either punish the offender or to compensate the victim. Participants play multiple rounds, and they are told that, in each round, they would encounter a different pair of players, therefore making it a sequence of one-shot games. In each round, each player (A, B, and the participant) starts with 200 chips (equivalent to £2); player A can take some chips from player B, who is passive and cannot oppose the decision. At this point, the participant can decide whether they want to do nothing and "Leave" the round with their own 200 chips or react by spending some of their 200 chips to either punish A by taking some chips away ("Take from A") or compensate B by giving some chips ("Give to B"). Importantly, participants can only spend chips, and never gain any. If they decide to react, they are then asked how much they are willing to spend to either take from A or give to B: they can spend up to 100 chips, knowing that for every 10 chips they spend, player A will lose 30 (punish) or player B will gain 30 (compensate). Participants were told that no negative payoffs were allowed, and the minimum player A could get was 0 chips; they were given no further information on what players A and B knew about the situation and the possibility of being punished or compensated by a third party. See Figure 1 for the task structure. In this version of the game, the players' payoffs could be represented as coins or digits; the same rules applied to both conditions. The game involved deception: participants were told that, at the end of the game, one trial would have been selected to determine the payoff of all players in that round, and therefore that their choice would have made an impact on the final payoff of that pair of players A and B, whom they had played with in the selected trial. In reality, A and B were not real players, and the chips distributions were built by the experimenters.

Each condition (coins/digits) consisted of 64 trials, varying in the level of injustice: the offender could take 0 (fair), 25, 50, 75, or 100 coins from the victim. Each level of injustice was presented in 8 trials, except for the "fair" condition (0 coins taken), which appeared in 32 trials. This design ensured an equal distribution of fair and unfair trials, considering that participants might expect fairness to be the most common choice in these types of games (Stallen et al., 2018; Civai et al., 2019). The presentation of the offender's (player A) and the victim's (player B) payoff on either side of the screen (left or right) was counterbalanced: the offender's payoff was presented on the right (or left) 4 times per level of injustice, per condition. Responses were selected by pressing one of the arrow keys (left arrow = Give to B; down arrow = Take from A; right arrow = Leave). The arrows were presented on the screen for each trial as a reminder of the correct key to press. All trials were randomised. There was no time limit to respond, but participants were encouraged to answer as quickly as possible to avoid having some participants using lengthy deliberation when, in fact, we were interested in participants' intuitions and quick judgments. The task was built with Experiment Builder, the EyeLink 1000 software for stimulus presentation (S-R Research, Canada).

2.1.3 Apparatus: Lab-Based Eye-tracking

To measure eye movement, we used desktop mount Eyelink 1000 (SR Research, Canada); we used a 13-point calibration, after which a validation phase was executed to make sure that the calibration had been accurate. To analyse eye-movements we defined two rectangular Regions of Interest (ROIs) with a size of 446 x 790 pixels (3.78 cm x 6.69 cm) including the label of the player (A or B) and the relative amount. The size of the ROIs does not change between conditions (coins vs digits). We discarded every fixation that was not located inside these two ROIs. More details on the eye-tracking apparatus and procedure can be found in the Supplementary Materials (SM 1.1).

2.1.4 Procedure

The recruitment was done via the Research Participation Scheme (RPS) system at the corresponding author's institution, and word of mouth. Once in the lab, participants were administered the task, including 6 initial practice trials, after giving consent to participate in the study. The task lasted on average 25 minutes.

Participants were either paid a show up fee (£5 Amazon voucher) or given 12 credits if they were students, plus a £5 Amazon voucher as bonus payment. At the end, participants were fully debriefed on the scope of the experiment and informed that they were not going to be paid according to one random trial, and to make up for the deception, they were paid more than they were expecting, i.e., a fixed bonus of £5 (Civai et al, 2020).

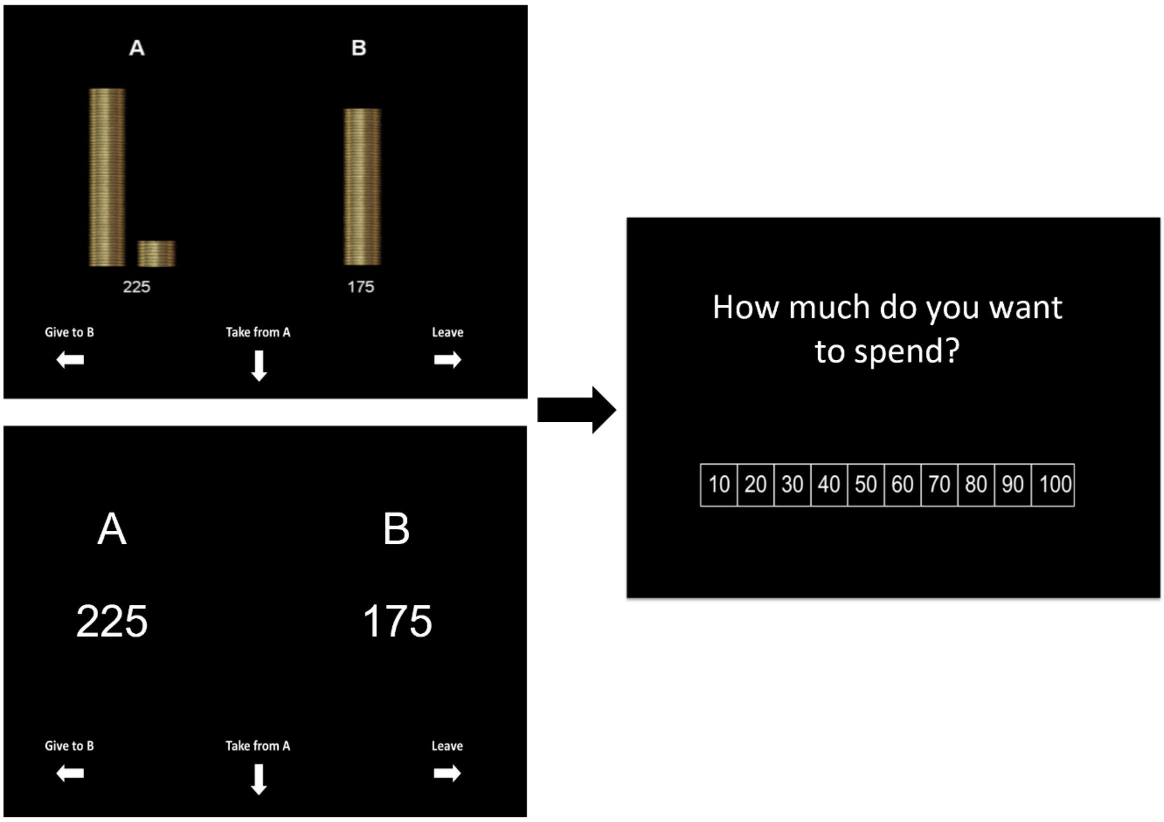

Figure 1. TPJG structure, as in experiment 1. First participants saw the players' payoffs, either as coins or digits, and had to choose one of the options ("Give to B", "Take from A" or "Leave"); after having selected their choice, if this was "Take from A" or "Give to B", they would be redirected to the amount screen, where they indicated how much they wanted to spend to either punish or compensate. If they selected "Leave", instead of the amount, the screen would tell them to "wait for the next trial". The arrows on the screen represented a reminder of the keys participants had to press to select that option (left arrow = give; down arrow = take; right arrow = leave.

2.1.5 Design and Analysis

This was a within-participant design. In one condition (coins), the payoff of the players (offender, or player A; victim, or player B) was represented as piles of coins; in a second condition (digits), the payoff of the players was represented as digits.

We measured:

- the percentage of dwell time (duration of gaze) on the offender's and on the victim's payoff;

- participants' choice (punish, compensate, or leave);

- the amount participants choose to spend to punish and to compensate (10 to 100, with increment of 10);

- in addition to these pre-registered outcome variables, we also considered the position of the first fixation, to seek support for the idea that coins indeed work as an exogenous attractor towards the offender.

Since this was a repeated measure design, we planned to use mixed models to account for the fixed effects of predictors and the random effect (intercept) of participants. The following models were employed:

- Model 1.1: a linear mixed model (lmer function in lme4 R package; Bates et al (2014)) to test whether the condition (coins vs digits) predicted the difference between dwell time on the offender's and the victim's payoffs, i.e., a vector obtained by subtracting the percentage of dwell time on the victim from the percentage of dwell time on the offender (Hypothesis 1). Full results for this model are reported in Figure 2a.
- Models 1.2: We then employed two logistic mixed models (glmer function in lme4 R package) to test whether the difference between dwell time on the offender's and the victim's

payoffs (Hypothesis 2) and the condition (coins vs digits) (Hypothesis 3) predicted the likelihood to punish and the likelihood to compensate for each trial. Full results for these models are reported in Figure 2b.

- Models 1.3: Additionally, two linear mixed models were also employed to test for the predictive effects of the difference between dwell time on the offender's and the victim's payoffs and the condition on the amount spent to punish and compensate. Full results for these models are reported in Figure 2c.

In all the models, we included the amount of chips taken by the offender as predictor, to control for the effect of the level of injustice as a manipulation check: if the manipulation worked, participants were expected to be more likely to punish/compensate, rather than leave, and to spend more to punish/compensate, as the level of injustice increased. All variables were standardised before running the models.

In order to get a measure of the effect of interest to use in G*Power to inform the subsequent online experiment, as an exploratory analysis, we calculated the proportion of punishment and compensation and the average amount spent across the levels of injustice for each participant and, using the statistical analysis software JASP, we performed a paired samples t-test to test whether the difference between the proportion of punishment and compensation, and the average amount spent to punish, are larger in the coins condition compared to digits. Results are reported in the SM (1.5).

2.2 Results

Descriptive statistics and the results of the manipulation check are reported in the SM (1.2 and 1.3).

Hypothesis 1: The difference in the duration of gaze directed towards the offender versus the victim is larger in the coin condition compared to the digits condition. Model 1.1 showed that the difference between the duration of gaze towards the offender and the victim was significantly predicted by the condition ($\beta = 0.12$, s.e. = 0.03, $t(4569) = 4.16$, $p < .001$) and the level of injustice ($\beta = 0.11$, s.e. = 0.01, $t(4569) = 7.67$, $p < .001$), indicating that, as expected, participants looked more at A (versus B) in the coins condition as opposed to the digits condition, and that the larger A's payoff, the more they looked at A (versus B). The standard estimates of these fixed effects are plotted in Figure 2a (sjplot R package; Lüdecke, 2020). These results suggest that in the coins condition, participants experienced a higher difference in visual saliency when comparing the payoffs of the offender and the victim. This difference might be determined by an automatic tendency of the participants to shift their attention toward the richer stimulus (the offender's payoff).

To further investigate this hypothesis, we ran a first fixation analysis to test whether in the coins condition the participants were more inclined to direct their attention towards the offender A. We calculated the proportion of times in which participants looked at player A first, controlling for the position of the stimulus on the screen (left/right), in both conditions: a paired samples t-test showed that participants' first fixation fell significantly more often on the offender in the coins condition compared to the digits condition ($t(35) = 4.89$, $p < .001$, Cohen's $d = .82$). This finding further supports the idea that participants experienced a higher difference in visual saliency in the coins condition, which automatically induced them to primarily focus on the offender's payoff.

Hypothesis 2: The more participants look at the offender's payoff, as opposed to the victim's, the more likely they are to punish and to spend more on punishment. As expected, models 1.2 showed that the likelihood of punishment increases with the increase of the difference of the percentage of attention towards the offender A (versus victim B) (56% more likely to punish when participant looks at the offender 1% more than the victim, which is significantly higher than the chance level, i.e., 50%), est. = 0.26, s.e. = 0.04, $z = 7.01$, $p <$

.001); the opposite effect is found for compensation, which is less likely to be chosen with the increase of the difference in attention allocation (45% more likely to compensate, therefore significantly less than the 50% chance level, when the participant looks at the offender 1% more than the victim; est. = -0.2, s.e. = 0.04, z = -5.12, p < .001). On the other hand, model 1.3 showed that the gaze had no effect on the amount spent to punish or compensate. A Pearson's correlation confirmed these results and showed that the difference between dwell time on the offender and on the victim positively correlates with the percentage of punishment vs compensation, in both conditions (coins: r = .68, p < .001; digits: r = .60, p < .001), and with the amount spent on punishment, minus compensation, only in the coin condition (coins: r = .45, p = .006; digits: r = .19, p = .262) .

Hypothesis 3: Participants are more likely to punish, and spend more on punishment, in the coins condition compared to the digits condition. Contrary to our expectations, models 1.2 show no effect of condition on the choice to punish (est. = 0.04, s.e. = 0.07, z = 0.64, p = .521) or compensate (est. = 0.05, s.e. = 0.07, z = -0.61, p = .544). The magnitude (probability) and the significance of these fixed effects are plotted in Figure 2b. However, models 1.3 show that participants spend more to punish in the coins condition compared to the digits condition (β = 0.06, s.e. = 0.02, t(1675) = 2.535, p = .01), but no effect of condition is found on compensation amount (β = 0.04, s.e. = 0.03, t(1134) = 1.18, p = .238) (Figure 2c).

The exploratory analysis (SM 1.5) confirmed the results of the mixed model analysis.

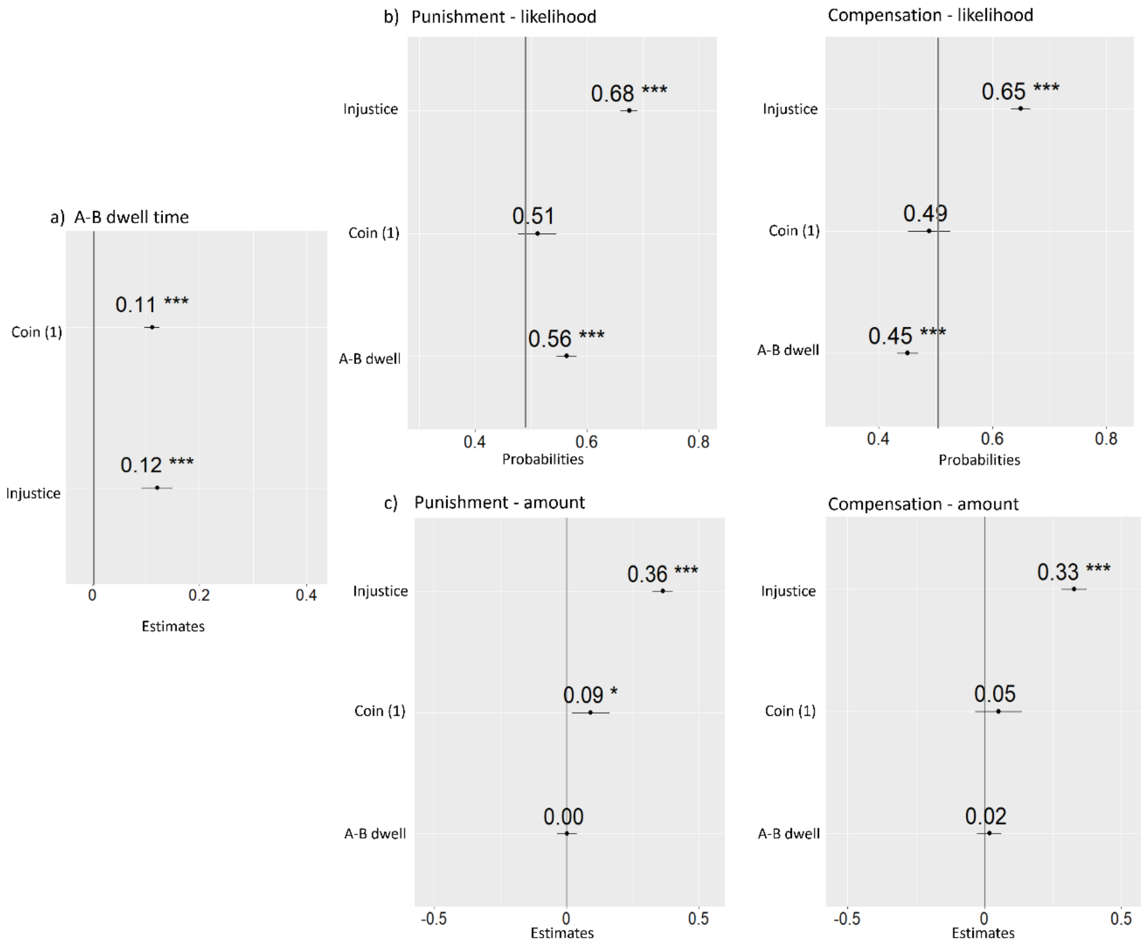

Figure 2. Experiment one **-** lab-based experiment. Magnitudes (standardised estimates or probabilities), error bars (95% CI) and significance (**p* < .05; ****p* < .001) of the fixed effects of the mixed models, with the black vertical line indicating null effect: a) standard estimates of the effects condition and injustice level on A-B dwell time (Model 1.1); b) probabilities of the effects of injustice level, A-B dwell time and condition on the likelihood to punish and compensate (Models 1.2); c) standard estimates of the effects injustice level, A-B dwell time and condition on the amount spent to punish and compensate (Models 1.3).

3. Experiment Two: Attentional Correlates and Online Eye-Tracking Data

To compensate for the limits of the laboratory experiment (small number of participants and gender/age-unbalanced sample) we ran a second online experiment with a larger and more heterogeneous sample of participants. The rationale and hypotheses for this experiment are the same as those for experiment one. The study received ethical approval from the Ethics committee at the corresponding author's institution, with protocol number [MASKED FOR PEER REVIEW].

3.1 Method

3.1.1 Participants

Sample size was determined before data collection. Our goal was to obtain .80 power to detect the smallest effect size of interest obtained in the paired samples t-test of experiment 1 (d=0.19) at the standard .05 alpha error probability. This returned N=177; however, since this was our first online eye-tracking study and we were unsure about the attrition rate, we planned for a final sample of 220 participants. Three-hundred and forty-one participants started the experiment, and 220 completed it (110 identifying as females, 109 identifying as males, and one self-identifying as gender non-conforming; all participants were 18 or older, mean age = 39.2, SD = 13.5); 19 participants were excluded from the analyses because of failed calibration, leaving us with 201 participants.

3.1.2 Materials

The main structure of TPJG was the same as in the lab-based experiment. The presentation of the offender's (player A) and the victim's (player B) payoff on either side of the screen (left or right) was counterbalanced, all trials were randomised, and there was no time limit to respond, but participants were encouraged to answer as quickly as possible. The task was built in Gorilla (www.gorilla.sc; Gorilla Experiment Builder (Anwyl-Irvine et al, 2018)), a cloud-based research platform where it is also possible to run webcam eye-tracking experiments. There were some differences between the lab-based and the online experiment, which are described in detail in the SM (2.1). In summary, since the aim of the study was to understand the relationship between attention and preference for punishment vs compensation, rather than preference to react or not to react to injustice, the option "Leave" was removed; consequently, the "fair" trials were removed, and the levels of injustice were increased, from 0, 25, 50, 75, 100, to 25, 50, 75, 100, 125, 150, 175, 200, for a total of 16 trials. To offset this forced choice between punish and compensate, we added the option to spend 0 to punish or compensate when choosing the amount. The scenarios presented here were hypothetical and the experiment not incentivised.

3.1.3 Apparatus: Online Eye-tracking

To capture eye movement, Gorilla uses Webgazer.js (https://webgazer.cs.brown.edu/; Papoutsaki et al, 2016) to detect a participant's face; the software then uses prediction models to infer people's gaze. What can be detected are estimates of gaze locations and a percentage occupancy of areas of interest. No video of the participant's face is recorded, and therefore anonymity is guaranteed. The experiment was built in such a way that the payoffs were presented in the top left and the top right quadrants, which were therefore considered our areas of interest; the percentage of dwell time while the payoffs were presented was extracted from those locations. These were the only data available from Gorilla, considering the settings chosen when building the experiment based on our hypotheses. Previous research used this same apparatus for online eye-tracking and found that lab-based results on decision-making and choice could be reliably replicated (Yang & Krajbich, 2021).

3.1.4 Procedure

The recruitment was done through Prolific (www.prolific.co), an online recruitment platform (Eyal et al. 2021; Palan and Schitter, 2018). From Prolific, participants were redirected to Gorilla, where they gave consent to participate. Calibration and validation were conducted at the beginning of the task and again after eight trials. Participants had three attempts to succeed in the calibration/validation procedure, otherwise they were allowed to continue with the task, but their eye-movement data were not used. Two practice trials were administered before starting with the actual game. The experiment took around 10 minutes. All participants were paid £7.5/h for their participation. Since the scenario was hypothetical, no task-dependent payment was added.

3.1.5 Design and Analysis

Design and measures were the same as experiment 1, except that now the participant's choice was limited to punish or compensate, and the amount spent could also be zero.

As stated in the preregistered analysis plan, we calculated the proportion of punishment and compensation and the average amount spent across the levels of injustice for each participant and we performed: a paired samples t-test (one tail) to test whether the average dwell time on the offender and on the victim differs between conditions (coins vs digits) (Hypothesis 1); a Pearson's correlation to test whether the difference between dwell time on the offender and on the victim positively correlates with punishment (Hypothesis 2); a paired samples t-test (one tail) to test whether participants are more likely to punish in the coins condition, and whether they spend more to punish in the coins condition, compared to digits (Hypothesis 3).

To allow for a direct comparison of these results with those from experiment 1, we also run the mixed models. We report these analyses and their results in full in the SM (2.5). It is important to note that the results include all trials, even those where participants selected £0 as the amount to punish or compensate; however, results do not change when these trials are excluded from the analysis. The same observation holds for experiments 3-5. The results excluding trials in which participants spent £0 can be downloaded as an html file created with R markdown https://osf.io/2ndcm?view_only=b8a20e04eeb345e486b19a58cd9ee224

3.2 Results

Descriptive statistics and the results of the manipulation check are reported in the SM (2.3 and 2.4).

The results from the preregistered analyses confirmed the lab-based findings.

Hypothesis 1: The difference in the duration of gaze directed towards the offender versus the victim is larger in the coin condition compared to the digits condition. A paired samples t-test showed that the percentage of dwell time on the offender compared to the victim marginally differs between conditions when considering the two-tailed test ($t(200) = 1.90$, $p = .059$, $d = .13$, 95% CI [-0.04, 2.08]), becoming significant when considering the pre-registered one-tailed test ($p = .030$). This is confirmed by the mixed model results reported in the SM (2.5).

In this online experiment, we did not conduct a first fixation analysis since the available eye-tracking data did not include this information.

Hypothesis 2: The more participants look at the offender's payoff, as opposed to the victim's, the more likely they are to punish and the more they spend on punishment. A Pearson's correlation showed that the difference between dwell time on the offender and on the victim positively correlates with the likelihood of punishment, in both conditions (coins: $r = .26$, $p < .001$; digits: $r = .25$, $p < .001$), and with the amount spent on punishment, minus compensation, in the coin condition (coins: $r = .18$, $p = .009$; digits: $r = .04$, $p = .46$), like in experiment 1. This is confirmed by the mixed model results reported in the SM, as participants were 55% more likely to punish, hence more than 50% chance level, when participant looks at the offender 1% more than the victim. This analysis also showed that when participant looks at the offender 1% more than the victim, they spent significantly more to punish, although the effect is small (SM 2.5).

Hypothesis 3: Participants are more likely to punish, and spend more to punish, in the coins condition compared to the digits condition. A paired samples t-test did not support the hypothesis that participants punish more in the coins condition ($t(200) = 1.46$, $p = .145$, $d = .10$, 95% CI [-0.01, 0.05] (two-tailed); $p = .072$ (one-tailed)), but another paired samples t-test showed that they spend more to punish in the coins condition, compared to digits ($t(200) = 1.97$, $d = .14$, $p = .050$, 95% CI [-0.01, 4.76] (two-tailed); $p = .025$ (one-tailed)). This is confirmed by the mixed model results reported in the SM (2.5).

3.2.1 Exploratory Findings in Experiments One and Two: evidence of offender bias

We further analysed the data from experiment 1 and 2 to obtain a more detailed picture of participants' preferences and understand whether one reaction (e.g., punishment) was preferred over the other (e.g., compensation), and whether one target (e.g., offender) was more attractive than the other (e.g., victim), irrespective of the mode of presentation (coins or digits). Results, reported in SM (3), showed that despite a larger difference in the coins condition, punishment was preferred to compensation in both conditions, and that, despite the difference between the dwell time on A and dwell time on B being larger in the coins condition, participants preferred to look at the offender A rather than the victim B, in both conditions.

4. Experiments One and Two: Discussion

In these two eye-tracking experiments we showed that, as expected, the decision to punish offenders or compensate victims can be predicted by analysing attentional processes.

Specifically, the more people look at the offender's payoff, the more they are inclined to punish, as opposed to compensate. These findings not only support the well-established idea that attention and choice are related, in that people tend to choose the option that they look at the longest (Armel et al., 2008; Krajbich et al, 2010), but they extend it further: participants are not simply choosing with higher probability the item they are paying attention to the most, but they are selecting an action (i.e., punish or compensate) as a consequence of the information they are paying attention to (offender's or victim's payoff).

From our findings, we can also hypothesize that visual appearance of stimuli, albeit task-irrelevant, acts as attentional attractor and affects our exogenous attention: specifically, when the offender's payoff was visually more salient than the victim's (coins condition), people attended to it immediately and for a longer duration[1]. This suggests that the format in which information is presented can influence how attention is directed toward one piece of information over another, thereby affecting the perceived relevance of the information acquired and influencing the subsequent decision.

Manipulating exogenous attention did not clearly lead to a choice manipulation: in fact, even if people looked at the offender more in the coins condition compared to the digits condition, they did not punish significantly more. However, in both experiments, the bottom-up effect seemed to influence the amount spent to punish, since people spent more to punish in the coins condition compared to the digits condition. This finding brings further support to the idea that deciding how to react to injustice and deciding how harshly to react are two distinct processes: previous studies suggest that, whilst deciding how to react is driven by a more rational process of unfairness and injustice detection, the severity of the reaction is

[1] We note that that a random effects mini meta-analysis across the two experiments, which is reported in the SM (7.1), showed a non-significant effect; however, results should be interpreted with caution given the high residual heterogeneity and the limited number of studies included

underpinned by more emotional processes (Stallen et al, 2018; Civai et al, 2019; Gummerum et al, 2022). Therefore, this suggests that, while the decision to react may not be easily influenced by task-irrelevant factors such as the presentation mode, these same factors may influence the emotional processes that determine the severity of the reaction.

The exploratory findings show an overall preference for punishment over compensation, irrespective of the mode of presentation. These results are in line with some previous literature (Stallen et al, 2018; Kühne et al, 2015), but inconsistent with other findings showing that compensation can be the preferred choice (Lotz et al, 2011; Thulin & Bicchieri, 2016; Jordan et al, 2016; Wang et al, 2024). Importantly, most of the studies that find compensation as the preferred choice correlate this behaviour with emotional reactions, such as feelings of stress and moral outrage, which could also be manipulated to trigger compensation (e.g., Thulin & Bicchieri, 2016; Wang et al, 2024), and which the current setting may have failed to elicit. Moreover, reputation effects seem to be important for boosting costly compensation, since this behaviour is a strong signal of trustworthiness (Jordan et al, 2016); in the current setting, given the anonymity of the players, reputation effects can be excluded. The current study also found a stronger attraction towards information about the offender's payoff compared to the victim's payoff, which we will call "offender bias". This suggests that, in the context of this specific game, where the offender is also the main agent that determines the outcome before the participant's decision, this preference is partially driven by top-down motivational factors potentially triggered by an agency bias. As some studies have suggested, punishment seems to be the most rewarding choice (Stallen et al, 2018; Kühne et al, 2015); however, this preference might be boosted by an automatic attentional bias towards the offender's payoff, as the current findings suggest. In other words, there could be an interaction between bottom-up and top-down attentional factors in third-party individual decision-making: it is possible that bottom-up mechanisms amplify the natural top-down propensity to focus more on perpetrators' outcomes than on victims' outcomes when attention is automatically attracted by a salient piece of information.

To account for the specific contribution of bottom-up and top-down factors in determining the observed attentional bias towards punishment, we ran three additional behavioural experiments, whereby, in each trial, we manipulated the information available to the participants in order to limit bottom-up access to, but not knowledge of, either the offender's or the victim's payoff and investigate whether this could influence choice. In experiment 3, participants could choose which information to access (offender's or victim's payoff), whilst in experiments 4 and 5 the information was randomly revealed, and therefore completely exogenously manipulated.

5. Experiment Three: Information Frames and Top-Down Selection

In experiment 3, we wanted to test whether limiting bottom-up processing can affect punishment and compensation by asking participants what information they want to have access to. In this experiment, which will be referred to as top-down (TD), participants could choose whether they wanted to reveal either the offender's or the victim's payoffs, before deciding how to react: in each trial, the payoffs were covered by two squares, clearly labelled "A" and "B", so that participants could choose whether to reveal the payoff of the offender (A) or of the victim (B) (Figure 3a).

Additionally, empathy was included as a potential predictor of information selection and choice, as outlined in the introduction, to assess whether it could account for some of the variance in these behaviours. To capture the different roles of affective and cognitive empathy, we administered the Questionnaire for Cognitive and Affective Empathy (QCAE; Reniers et al, 2011).

We hypothesise that:

1. The more participants reveal the offender's payoff rather than the victim, the more they are likely to punish, and to spend more to do so;
2. Affective empathy positively predicts the likelihood to reveal B's payoff and positively predicts compensation;
3. Cognitive empathy positively predicts the amount spent to react to injustice (punish or compensate), but no specific hypotheses are made on the effect on the type of reaction.

Experiments 3-5 received ethical approval from the Ethics committee at the corresponding author's institution, with protocol number [MASKED FOR PEER REVIEW].

5.1 Method

5.1.1 Participants

The data from a representative sample, with respect to age, sex, and ethnicity, of the UK population (18 or older) were collected. We decided to limit the sample to the residents of the UK to favour a higher heterogeneity in terms of other demographics (i.e., age, sex, and ethnicity). The recruitment was done through the platform Prolific, where the minimum number of participants to recruit for a representative sample is 300; therefore, we tested 300 participants. We note that, whilst the sample is considered representative of the UK population based on the Simplified GB Census, it is still limited to Prolific workers and therefore the term "representative" should be contextualised accordingly. An attentional check was included: a question was added to the QCAE asking participants whether they agreed with the statement "I climb to the top of Mount Everest every day to get to work"; participants who scored anything more than 2 (with one being "strongly disagree") were discarded. This left us with data from 285 participants.

The number of participants included in the analysis in experiments 3-5 afforded 80% power to detect an effect size of d = 0.17 in a paired samples t-test, where the effect of interest is the difference between the proportion of punishing and compensating choices when the offender's or the victim's payoffs were revealed, with a 5% false-positive risk rate.

5.1.2 Materials

The TPJG main structure and rules were the same as previously described in experiment two, coins condition. Unlike the previous task, at the beginning of each trial the payoffs were covered by two grey squares; participants could only select one square, and only one payoff would be revealed. The letters "A" and "B" appeared on the squares, so that participants could always choose whether to reveal the payoff of the offender (A) or of the victim (B). Once the payoff was revealed, participants could indicate whether they wanted to punish ("Take from A") or compensate ("Give to B") (see Figure 3a). The experiment had 8 trials (randomised), which varied in the level of injustice: the offender could take 25, 50, 75, 100, 125, 150, 175, 200 coins from the victim. For half of the trials, the offender's payoff was presented on the right. In this new set of experiments, we used coins to represent the payoffs, as we believed that this representation would facilitate information processing by combining both a visual representation of the magnitude (piles of coins) and numerical values displayed beneath the images. Since participants only saw either one (offender) or the other (victim) payoffs, never both simultaneously, the visual representation of one payoff was never directly compared to the other. Therefore, the previously observed "attentional advantage" of the offender's payoff over the victim's when expressed in coins would not be a factor in this instance.

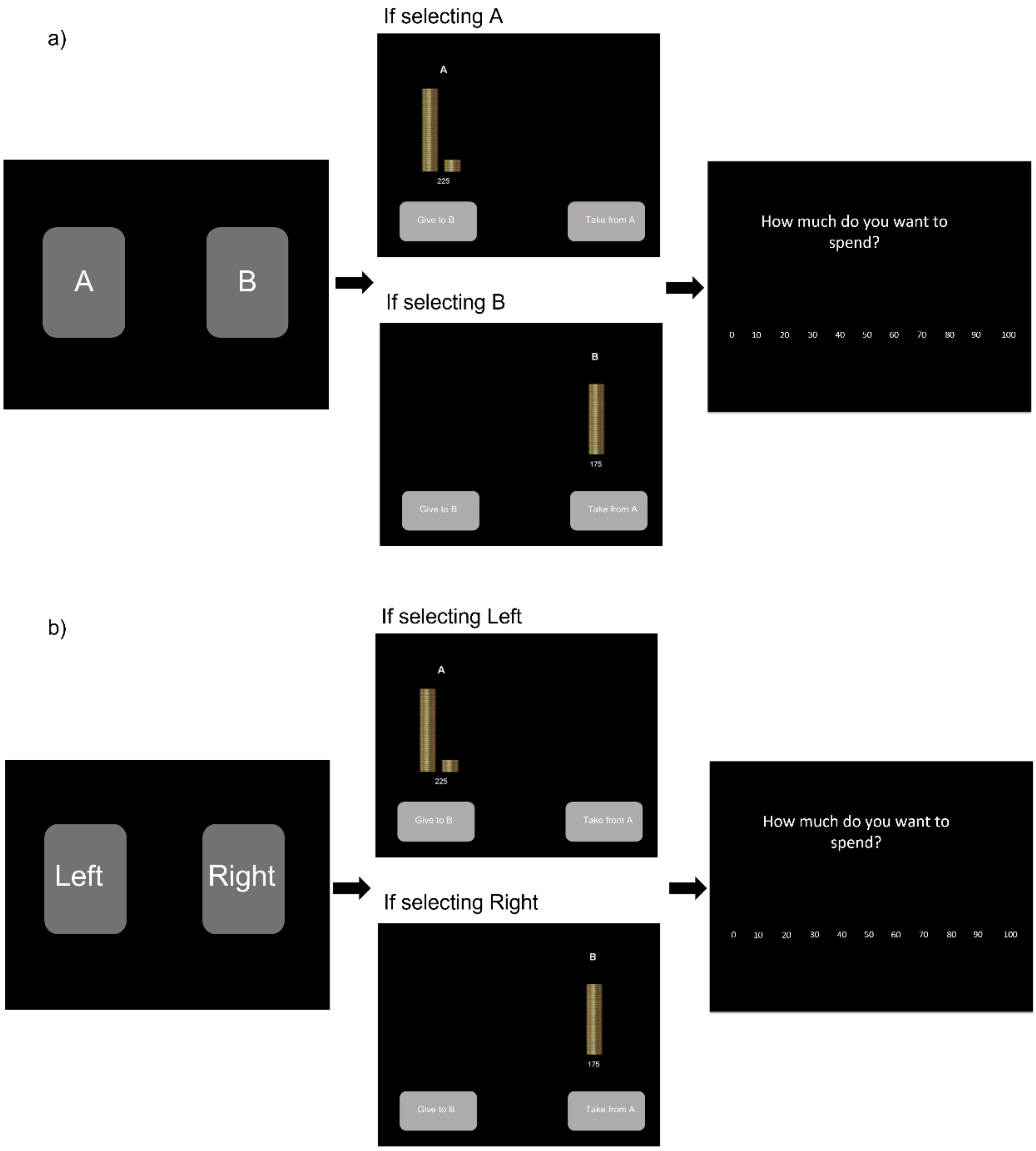

Figure 3. Experiments three (TD– top-down) and four/five (BU– bottom-up) – TPJG structure. First participants choose which payoff to uncover, then they choose one of the options ("Give to B", "Take from A"), and ultimately, they choose how much they want to spend, from 0 to 100 chips. a) In experiment three (TD), participants can choose whether to uncover A's or B's payoff; b) in experiments four and five (BU), they choose a location (left or right) but not the payoff to uncover; there is no relationship between the location and the payoff.

The QCAE contains 31 statements that participants rate on a 4-point Likert scale from "strongly disagree" (1) to "strongly agree" (4); examples of statements are "I am inclined to get nervous when others around me seem to be nervous" or "I can easily tell if someone else wants to enter a conversation". The scores can be categorised in five subscales, two of Cognitive Empathy (perspective taking; online simulation) and three of Affective Empathy (emotional contagion; proximal responsivity; peripheral responsivity). For this experiment, as well as for experiments 4 and 5, only the two scores of Cognitive and Affective empathy were considered for analysis, with higher scores indicating higher empathy.

5.1.3 Procedure

The procedure for experiments 3-5 is similar: participants were redirected to Gorilla from Prolific. Here, they first gave consent to participate, then played the TPJG, and took the QCAE. Finally, they were debriefed. The experiment took no more than 12 minutes. All participants were paid £7.5/h for their participation. In experiments 3 and 4, since the TPJG scenarios were hypothetical, no task-dependent payment was added.

5.1.4 Design and Pre-registered Analysis

We measured:

- participant's choice of payoff to reveal (offender A or victim B);
- participants' decision to punish or compensate;
- the amount participants spend on punishing or compensating.

Pre-registered mixed models were used to account for the fixed effects of predictors and the random effect (intercept) of participants:

- Model 3.1: a generalised linear mixed model was run to see whether the choice of payoff to reveal (A or B) would predict the likelihood of punishment (Hypothesis 1).

Cognitive and affective empathy scales, as well as the level of injustice, were added as predictors (Hypotheses 2 and 3). The full results for this model are reported in Figure 4a.

- Models 3.2: A linear mixed model was run to test for these predictive effects on the amount spent to punish or compensate (Hypothesis 1). The full results for these models are reported in Figure 4b.
- Model 3.3: A second generalised linear mixed model investigated whether empathy could predict the choice of payoff to reveal (Hypothesis 2). The full results for this model are reported in Figure 4c.

5.1.5 Exploratory analysis

To test whether there was an offender bias and a punishment preference, we calculated the proportion of "A" choices for each participant, as well as the proportion of choices to punish, and we performed two one-sample t-tests (test value = 0.5). We also performed two paired samples t-tests on the proportion of choices to punish and compensate for both revealed payoffs (A or B) to see whether there was a punishment preference.

For experiments 3-5, we conducted an exploratory analysis to investigate whether cognitive and affective empathy could moderate the effect of the revealed payoff on 1) the likelihood of punishment/compensation and 2) the amount spent to punish or compensate.

5.2 Results

Descriptive statistics and manipulation check are reported in SM (4.1 and 4.2).

Hypothesis 1: The more participants decide to reveal the offender's payoff rather than the victim, the more they are likely to punish, and spend more to do so. Model 3.1 showed that, as expected, the choice to reveal the offender's payoff positively predicted the likelihood of punishment (78% more likely to punish if participant chose to reveal the offender's payoff, hence higher than 50% chance; est. = 1.25, s.e. = 0.10, z = 11.59, p < .001) (Figure 4a).

Models 3.2 showed that the amount spent to punish is positively predicted by the choice to reveal the offender's payoff (β = 0.31, s.e. = 0.05, $t(1388) = 5.99$, $p < .001$); the amount spent to compensate, on the other hand, was not predicted by the choice to reveal the offender's payoff (β = -0.03, s.e. = 0.05, $t(781) = -0.57$, $p = .566$) (Figure 4b).

Hypothesis 2: affective empathy positively predicts the likelihood to reveal B's payoff and positively predicts compensation. Contrary to our expectations, model 3.3 showed that affective empathy did not predict the choice to reveal the victim's payoff, and neither did cognitive empathy (affective: est. = 0.07, s.e. = 0.07, $z = 0.96$, $p = .339$; cognitive: est. = 0.04, s.e. = 0.07, $z = -0.51$, $p = .611$) (Figure 5c). Model 3.1 showed that affective empathy did not predict compensation either (est. = 0.02, s.e. = 0.08, $z = 0.31$, $p = .755$) (Figure 4a).

Hypothesis 3: cognitive empathy positively predicts the amount of chips spent to either punish or compensate. Models 3.2 showed no effect of empathy on the amount spent to punish (cognitive: β = -0.06, s.e. = 0.05, $t(279) = -1.28$, $p = .201$; affective: β = 0.03, s.e. = 0.05, $t(284) = 0.71$, $p = .481$), or to compensate (cognitive: β = -0.01, s.e. = 0.05, $t(238) = -0.29$, $p = .772$; affective: β = 0.04, s.e. = 0.05, $t(232) = 0.92$, $p = .360$) (Figure 4b). Interestingly, cognitive empathy negatively predicted the likelihood of punishment (43% less likely to punish for an increase of one cognitive empathy unit; est. = -0.3, s.e. = 0.08, $z = -3.69$, $p < .001$) (Figure 4a)).

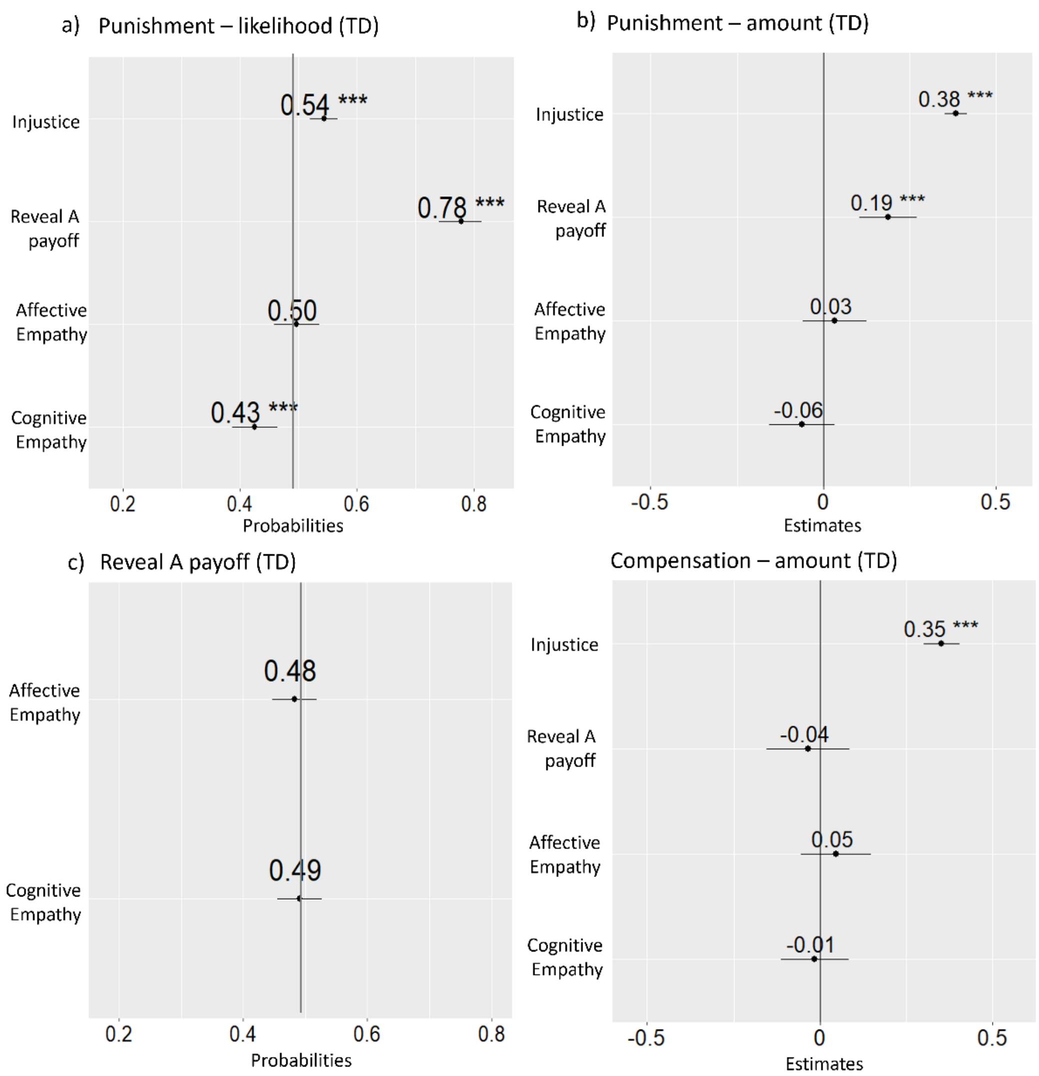

Figure 4. Experiment three. Magnitudes (standardised estimates or probabilities), error bars (95% CI) and significance (***p < .001) of the fixed effects of the mixed models, with the vertical black line indicating null effect: a) probabilities of the effects of revealed payoff, injustice level, affective and cognitive empathy on the likelihood to punish (Model 3.1); b) standard estimates of the effects of revealed payoff, injustice level, affective and cognitive empathy on the amount spent to punish (above) and to compensate (below) (Models 3.2); c) probabilities of the effects cognitive and affective empathy on the choice of revealing the offender’s payoff (Model 3.3).

5.2.1 Results for the exploratory analysis

One-sample t-tests showed that indeed participants uncovered the offender's payoff significantly more often than the victim's, confirming the offender bias (t(284) = 9.21, p < .001, d = .55, 95% CI [0.1, 0.15]), and that they preferred punishment to compensation (t(284) = 10.05, p < .001, d = .59, 95% CI [0.12, 0.17]). Additional paired samples t-tests showed that the proportion of punishment was higher than compensation after choosing to reveal the offender's payoff, which all participants did at least once (t(284) = 13.10, p < .001, d = 0.78, 95% CI [0.4, 0.54]), confirming the previous mixed model analysis; on the other hand, in the trials where participants chose to reveal the victim's payoff, which 52 participants never chose to do, there was no difference between the proportion of punishment and compensation (t(284) = -0.71, p = .475, d = -0.04, 95% CI [-0.1, 0.05]) (Figure 5).

Results for the moderating effects of empathy show that the higher the cognitive empathy, the less likely participants are to choose punishment when player's A payoff is revealed (est. = -0.28, s.e. = 0.12, z = -2.42, p = .016), suggesting that higher cognitive empathy may shift the preference towards compensation when deciding to reveal the offender's payoff.

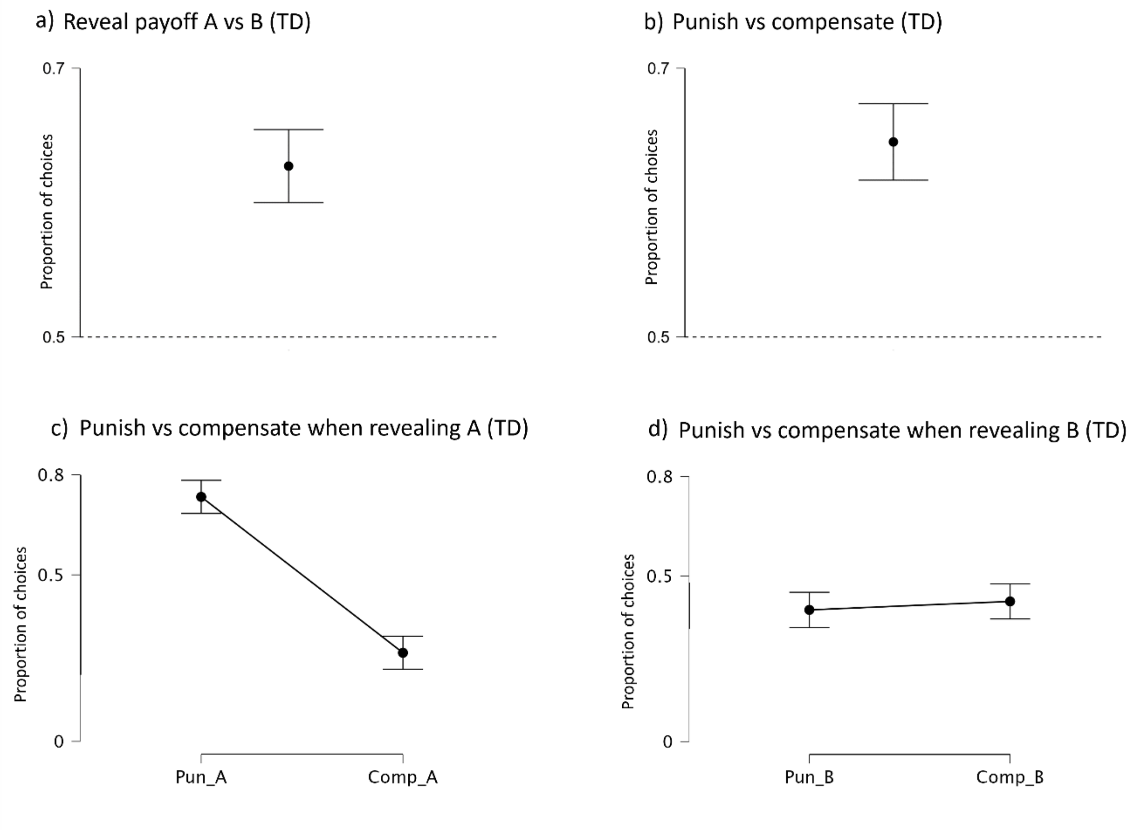

Figure 5. Experiment three. One-sample t-tests (test-value = 0.5) (95% CI) showing an offender bias and a punishment preference on the: a) choice of payoff to reveal (A vs B) and b) proportion of punishment vs compensation choices. Paired samples t-tests (95% CI) showing c) a significant difference between proportion of punishment and compensation choices when choosing to reveal A's payoff; d) a non-significant difference between proportion of punishment and compensation choices when choosing to reveal B's payoff.

6. Experiment Four: Information Frames and Bottom-Up Presentation

In experiment 4, referred to as bottom-up (BU), the information that participants can directly access was exogenously manipulated: participants could reveal only one of the two payoffs, but, unlike experiment 3, they did not know in advance which one they were revealing, since the two squared covering the payoffs were labelled "Left" and "Right", with no indication of the player's identity (Figure 3b). This way, the piece of information participants were exposed to was not necessarily the one they would have chosen.

For experiment 4, we hypothesise that:

1. If availability of information (bottom-up) influences choices, then the revealed payoff positively predicts the likelihood of punishment (if the offender's payoff is revealed) or compensation (if the victim's payoff is revealed) and the amount spent;
2. Affective empathy positively predicts compensation;
3. Cognitive empathy positively predicts the amount of chips spent to react to injustice (either to punish or to compensate).

6.1 Method

6.1.1 Participants

As for experiment three, the data from a representative sample (N=300), with respect to age, gender, and ethnicity, of the UK population (18 or older) were collected through the platform Prolific, and the same attentional check was included. This left us with data from 284 participants.

6.1.2 Materials

The TPJG version employed in this experiment was the same as in experiment three, with one key difference: here, the words "Left" and "Right" appeared on the squares, so that participants could never know in advance whose payoff they were revealing (see Figure 3b). Therefore, irrespective of which square participants selected, they were presented with the offender's payoff 50% of the time. Whilst the action of choosing left or right is not informative of participants' preferences, and we would have achieved a similar result simply by randomly presenting participants with one of the other payoffs, the aim was to keep the structure of the task as close as possible to that of experiment three.

The QCAE was also administered.

6.1.3 Procedure

The procedure was the same as experiment 3.

6.1.4 Design and Pre-registered Analysis

In this experiment, the revealed payoff was experimentally manipulated, and therefore it was considered a within-participant independent variable. We measured:

- participants' choice to punish or compensate;
- the amount participants spend to punish or to compensate.

Pre-registered mixed models were used to account for the fixed effects of predictors and the random effect (intercept) of participants:

- Model 4.1: a generalised linear mixed model was run to see whether the payoff randomly revealed (A or B) would predict the likelihood of punishment (Hypothesis 1). Cognitive and affective empathy scales, as well as the level of injustice were added as predictors (Hypotheses 2 and 3). The full results for this model are reported in Figure 6a.
- Models 4.2: Moreover, a linear mixed model was run to test for these predictive effects on the amount spent to punish and compensate (Hypothesis 1). The full results for these models are reported in Figure 6b.

6.1.5 Exploratory Analysis

To test whether there was a punishment preference in either of the revealed payoff conditions (A or B), we calculated the proportion of choices to punish and compensate for both conditions and performed two paired samples t-tests.

6.2 Results

Descriptive statistics and manipulation check are reported in SM (5.1 and 5.2).

Hypothesis 1: The revealed payoff positively predicts the likelihood of punishment (if the offender's payoff is revealed) or compensation (if the victim's payoff is revealed) and the amount spent. Model 4.1 showed that, as expected, the availability of the offender's payoff positively predicted the likelihood of punishment (78% more likely to punish if the offender's payoff was available; est. = 1.26, s.e. = 0.1, z = 12.91, p < .001). Models 4.2 showed that both the amount spent to punish (β = 0.11, s.e. = 0.04, $t(1170)$ = 2.83, p = .005) and to compensate (β = 0.13, s.e. = 0.05, $t(745)$ = 2.48, p = .013) are positively predicted by the availability of the offender's payoff (Figures 6a and 6b); this result supports our hypothesis as far as punishment is concerned, but goes in the opposite direction when considering compensation.

Hypothesis 2: affective empathy positively predicts compensation. Model 4.1 showed no effect of empathy on the likelihood of compensating (cognitive: est. = 0. 04, s.e. = 0. 07, z = 0.60, p = .550; affective: est. = -0. 03, s.e. = 0. 07, z = -0.43, p = .669) (Figure 6a, for punishment, perfectly collinear with compensation).

Hypothesis 3: cognitive empathy positively predicts the amount spent to either punish or compensate. Model 4.2 showed no effect of empathy on the amount spent to punish (cognitive: β = 0.06, s.e. = 0.05, $t(272)$ = 1.17, p = .244; affective: β = 0.01, s.e. = 0.05, $t(276)$ = 0.15, p = .881) or compensate (cognitive: β = 0.07, s.e. = 0.05, $t(248)$ = 1.47, p = .142; affective: β = 0.02, s.e. = 0.05, $t(249)$ = 0.32, p = .750) (Figure 6b).

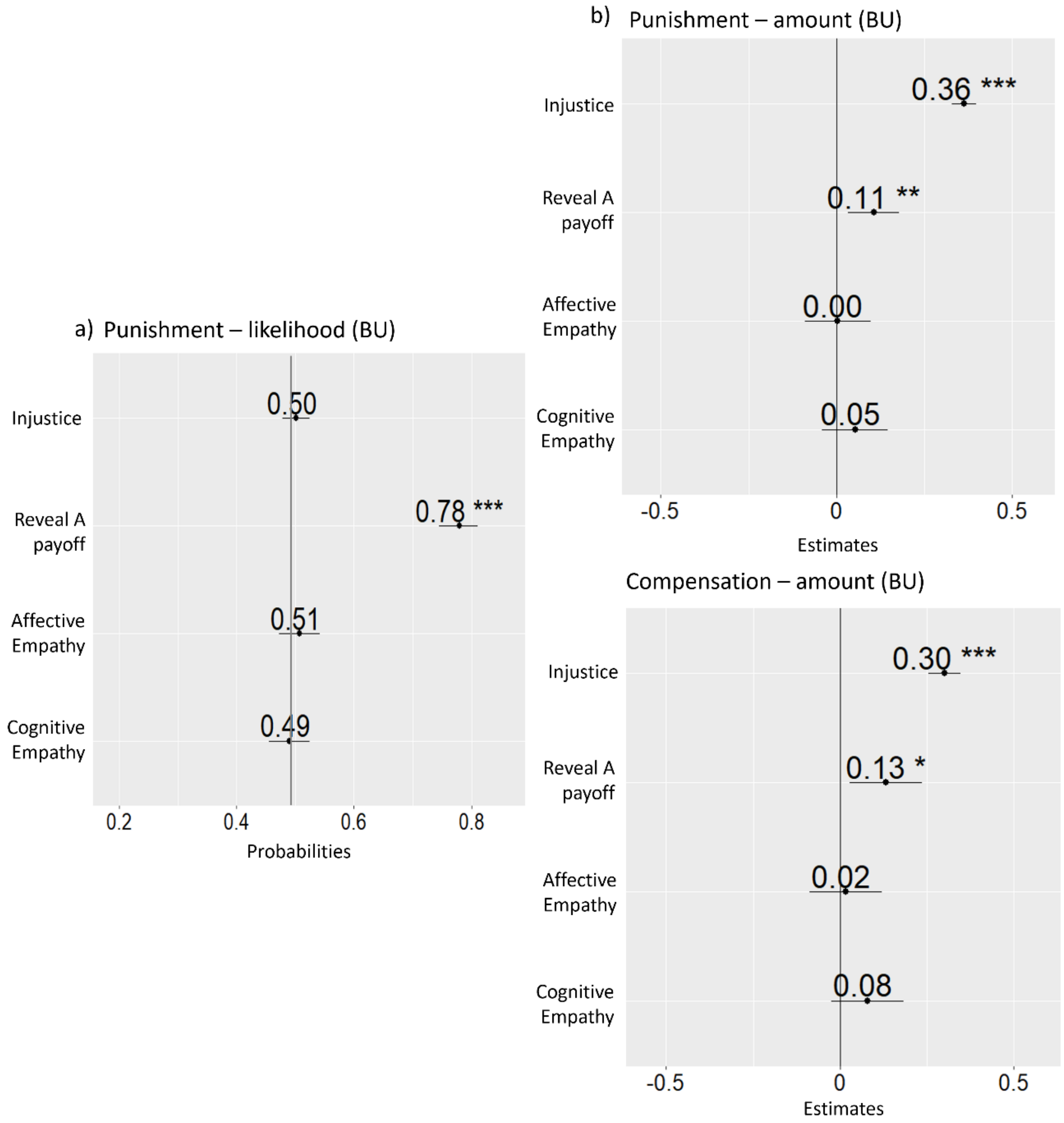

Figure 6. Experiment four. Magnitudes (standardised estimates or probabilities), error bars (95% CI) and significance (*p < .05; **p < .005;*** p < .001) of the fixed effects of the mixed models, with the vertical black line indicating null effect: a) probabilities of the effects of revealed payoff, injustice level, affective and cognitive empathy on the likelihood to punish (Model 4.1); b) standard estimates of the effects of revealed payoff, injustice level, affective and cognitive empathy on the amount spent to punish (above) and compensate (below) (Models 4.2).

6.2.1 Results from the exploratory analysis

Paired samples t-tests showed that, in line with what was found in the TD experiment, while the proportion of punishment is higher than compensation when the offender's payoff is revealed ($t(283) = 13.49$, $p < .001$, $d = .80$, 95% CI [0.4, 0.54]) (Figure 7a), there was no difference when the victim's payoff was revealed ($t(283) = -1.21$, $p = .228$, $d = -.07$, 95% CI [-0.12, 0.03]) (Figure 7b).

Results from the exploratory analysis on the moderating effect of empathy show a significant interaction between revealed payoff and affective empathy (est. = 0.33, s.e. = 0.10, $z = 3.16$, $p = .002$), showing that the higher the affective empathy, the more likely participants are to choose punishment when player's A payoff is revealed, and to choose compensation when player's B payoff is revealed. This may suggest that affective empathy enhances the likelihood of making the congruent choice, or, in other words, being affected by the information frame. No effect is observed on the amount.

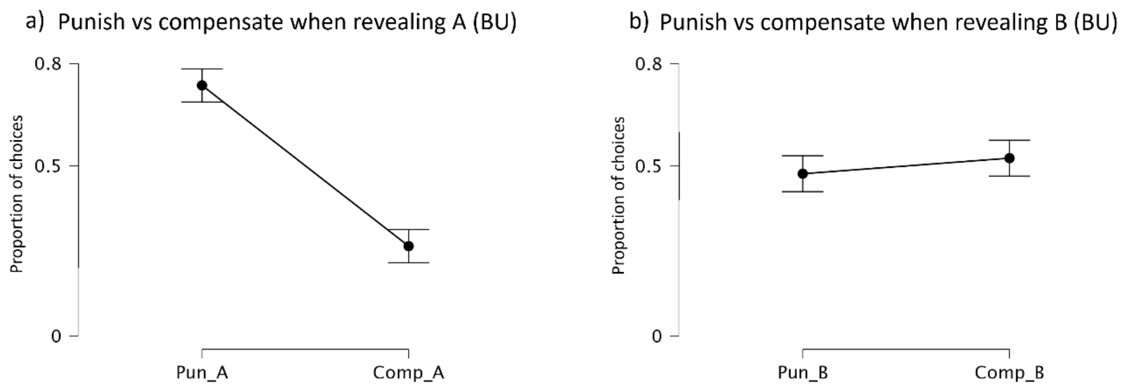

Figure 7. Experiment four. Paired samples t-tests (95% CI) showing a) a significant difference between proportion of punishment and compensation when A's payoff is revealed; b) a non-significant difference when B's payoff is revealed.

We identified two key limitations in Experiment 4. First, although we observed that the percentage of punishing choices varied within participants depending on whether the offender's or the victim's payoffs were presented, this does not necessarily imply that participants would change their preferences when moving from a condition where all payoffs are shown to one where information is selectively framed. To properly assess this, it is necessary to compare choice behaviour in the task where information is framed with choice behaviour in a task where all the payoffs are simultaneously available. Second, in experiment 4, as in experiments 2 and 3, the scenarios are hypothetical: while some evidence suggests that responses to hypothetical and real scenarios do not differ significantly (Gillis & Hettler, 2007), other research shows that findings may change depending on the type of scenario (e.g., Forsythe et al, 1994; Amir et al, 2012). Experiment 5 was run to address these limitations.

7. Experiment Five: Bottom-Up + Baseline

In experiment 5, after the BU task, the standard TPJG, where participants had to choose between compensation or punishment having all information about the payoffs available at once (as in experiment 2), was administered as baseline. This aimed to investigate whether the manipulation in the BU experiment (i.e., random exposure) would effectively change participants' individual choices. Therefore, we analysed whether the preferences for each participant changed between the BU task and the baseline. We hypothesise that:

1. If availability of information (bottom-up, random exposure) influences choice, then a change between choices in the BU task (manipulation) and in the standard TPJG (baseline) will be observed, in that people will punish more (less) in the BU task when the offender (victim) is revealed compared to the baseline;

2. Results from experiment 4 will be replicated: the revealed payoff positively predicts the likelihood of punishment (if the offender's payoff is revealed) and the amount spent.

7.1 Method

7.1.1 Participants

As for experiment 4, the data from a representative sample (N=300), with respect to age, gender and ethnicity, of the UK population (18 or older) were collected through the platform Prolific, and the same attentional check was included. This left us with data from 292 participants. In addition to the show-up fee, participants were paid £1 bonus that was believed to be performance-based.

7.1.2 Materials

The task version employed in this experiment was the same as in experiment 4 (BU task; see Figure 3b). In addition to that, the standard TPJG was administered, like in experiment 2; both the coins and the digits versions were administered, but in counterbalanced blocks rather than as randomised trials. A key difference between the current experiment and the previous online ones (2-4) was that here the game was not hypothetical: in fact, for this experiment, participants were told that players A and B had played before and that at the end of the experiment, one random trial would have been chosen to determine all players' final bonus payoffs. To make up for the deception, all participants were given a £1 bonus, which was the highest amount that they could expect to be paid.

The QCAE was also administered.

7.1.3 Procedure

The main procedure was the same as experiment 3 and 4. Participants first played the BU task, followed by the baseline; this order was established to avoid carry-over effects from the standard TPJG to the BU task.

7.1.4 Design and Analysis

Like in experiment 4, we measured:

- participants' choice to punish or compensate;
- the amount participants spend to punish or compensate.

We considered only the BU task trials in which the victim's payoff (or the offender's payoff) was revealed and compared, through a paired samples t-test, the percentage of punishment choices in the BU task to the percentage of punishment choices in the baseline, averaged across coins and digits (Hypothesis 1).

In addition to this, linear mixed models (Models 5.1 and 5.2) were run on the BU task data to see whether the payoff randomly revealed (A or B) would predict the likelihood of punishment, and the amount spent to punish and compensate, in an attempt to replicate results in experiment 4 (Hypothesis 2).

7.2 Results

Descriptive statistics and manipulation check are reported in SM (61. And 6.2).

Hypothesis 1: People punish more (less) in the BU task when the offender (victim) is revealed compared to the baseline. As predicted, we observe a significant difference in the punishment rate between trials in which only the offender's/victim's payoff was revealed and the baseline: people punished more in the BU task compared to the baseline when the offender's payoff was revealed ($t(291)=5.62$, $p < .001$, $d = .33$, 95% CI [0.21, 0.45]), whilst they punish less when the victim's payoff was revealed ($t(291)= -6.58$, $p < .001$, $d = .38$, 95%

CI [0.27, 0.5]), indicating that the bottom-up frame does change individual preferences (see Figure 8)

Hypothesis 2: Experiment 4 results are replicated. Results from experiment 4 are replicated: model 5.1 showed that, as expected, the availability of the offender's payoff positively predicted the likelihood of punishment (76% more likely to punish if the offender's payoff was available; est. = 1.16, s.e. = 0.1, z = 11.69, p < .001). Models 5.2 showed that both the amount spent to punish (β = 0.08, s.e. = 0.08, t(1289) = 1.97, p = .050) and to compensate (β = 0.15, s.e. = 0.05, t(727) = 2.54, p = .011) are positively predicted by the availability of the offender's payoff. As for experiment 4, these results support our hypothesis as far as punishment is concerned but go in the opposite direction when considering compensation.

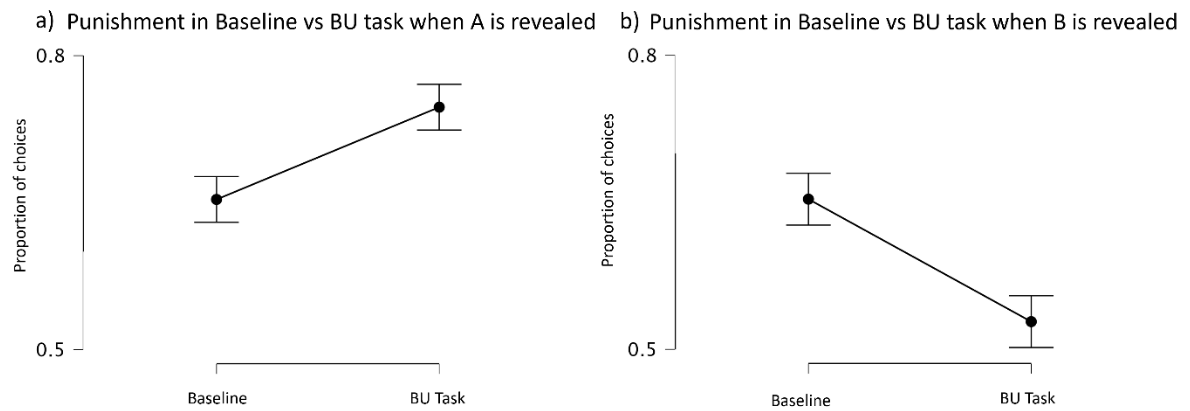

Figure 8. Experiment five. Paired-sample t-tests (95% CI) showing a) a significant higher likelihood of punishment for the BU task compared to the baseline condition when A's payoff is revealed; b) a significant lower likelihood of punishment for the BU task compared to the baseline condition when B's payoff is revealed.

When including cognitive and affective empathy in Models 5.1 and 5.2, results show that punishment is negatively predicted by affective empathy (est. = -0.18, s.e. = 0.08, z = -2.24,

p = .025), and positively predicted by cognitive empathy (est. = 0.17, s.e. = 0.08, z = 2.13, p = .033), whilst no effect of empathy was found when considering the amount. Results from the exploratory analysis on the moderating effect of empathy show a significant interaction between revealed payoff and affective empathy (est. = 0.31, s.e. = 0.10, z = 3.04, p = .002), again showing that the higher the affective empathy, the more likely participants were to choose compensation when player's B payoff is revealed. No effect is observed on the amount.

8. Experiments 3-5: Discussion

Through three behavioural experiments ran online with UK representative samples with respect to age, gender, and ethnicity, we have investigated whether information framing, information selection, attentional mechanisms, and empathy influenced decisions on how to react to injustice. First of all, the findings from the TD experiment, where participants could choose whose payoff to reveal, confirmed what we concluded from the eye-tracking experiments: the existence of an offender bias, as participants were much more likely to reveal the offender's payoff compared to the victim's. Moreover, as hypothesised, participants were more likely to punish than compensate after revealing the offender's payoff. However, despite showing an overall preference for punishment, this preference disappeared when participants revealed the victim's payoff. This means that, when people are more interested in the victim, and choose to get information on their status, they are less likely to punish. Interestingly, this does not clearly translate into a preference reversal, in that compensation is not chosen more often than punishment.

Results from the two BU experiments, where participants were randomly exposed to the offender's or the victim's payoffs with equal frequency, followed the same pattern: as expected, participants were more likely to punish than compensate when exposed to the offender's payoff, but this preference disappeared when they were exposed to the victim's payoff. Importantly, as shown in experiment 5, this manipulation actually changed people's individual choices, in that participants punished more (or less) when exposed to the offender's (or victim's) payoff than they would have otherwise done (baseline condition). This clearly shows that information encoding driven by automatic bottom-up processing can affect moral decisions: whilst in TD it could be argued that those who chose to uncover the victim's payoff might have been more predisposed to compensation to begin with, in BU this confound is removed by the randomness of the exposure, leaving any effect to be explained by the frame. Interestingly, when the amount spent is considered, participants in both experiments 4 and 5 spend more both to punish and to compensate when the offender's payoff is revealed. This suggests that considering the offender's payoff triggers an increased sensitivity to injustice, leading people to spend more to react to injustice regardless of their chosen response.

These findings are in line with other studies, across fields, that show that manipulating people's attentional focus towards the offender or the victim, either explicitly or implicitly, does influence their willingness to punish or compensate (Gromet & Darley, 2009; Kühne et al, 2015). These results show that it is possible to affect people's decisions by redirecting their automatic coding of salient information, supporting and extending the idea that basic information acquisition processes play an important role in shaping moral decision-making (Pärnamets et al, 2015; Ghaffari & Fiedler, 2018) by influencing consequential choices that go beyond the immediate information acquired.

In addition to this, we also looked at cognitive and affective empathy as potential predictors of these effects. We hypothesised that affective empathy would positively predict a focus on the victim, while no directional hypothesis was made for cognitive empathy, except for it being a positive predictor of the overall amount spent to react to injustice. Our results do not clearly support the view that empathy plays a predictive role in choosing between punishment and compensation. In the TD experiment, participants who scored higher in the cognitive empathy subscale were less likely to punish, but this result was not replicated in the BU samples: in fact, in experiment 5 we found the opposite, with punishment being positively predicted by cognitive empathy and negatively predicted by affective empathy. Because of these inconsistencies in our results, we do not believe it would be cautious to consider them as reliable. Several studies suggest that empathy, particularly empathic concern or affective empathy, plays a role in altruistic decision-making (e.g., FeldmanHall et al., 2015; Lim & DeSteno, 2016), and more specifically, in decisions to help or compensate victims of injustice (Leliveld et al., 2012; Hu et al., 2015). However, some studies report different findings, especially when considering the cognitive aspect of empathy, or perspective-taking (Lu & McKeown, 2018). Others suggest that empathy influences both punishment and compensation (Will et al., 2013). One possible explanation for these inconsistencies is that empathy does not distinctly separate preferences for punishment versus compensation but rather differentiates between costly altruistic behaviour (whether punishment or compensation) and self-interest. Our exploratory results, which examine empathy as a moderator of the revealed payoff effect, may partially support this idea: affective empathy increases the likelihood of punishment when the offender's payoff is revealed and the likelihood of compensation when the victim's payoff is revealed. This suggests that rather than driving a preference for one response over the other, affective empathy heightens sensitivity to context and the framing of information. This interpretation is also in line with previous results from Hu and colleagues, who found that higher empathic individuals punished more when instructed to focus on the offender (Hu et al, 2020).

9. General discussion and conclusions

We conducted five experiments aimed to investigate third-party costly reaction to an injustice, i.e., costly punishment of the offender or compensation of the victim, by analysing how attentional processes, the presence of attractors, frame effects, and differences in empathy may explain and affect people's choices, operationalised as behaviour in a third-party game.

First of all, we found that a preference for punishment was predicted by an intrinsic, top-down, focus on the offender's payoff. For the first time using eye-tracking we show that the more people look at the offender, the more they are likely to punish; importantly, these findings extend our understanding of the relationship between attention and choice, as they show that participants not only choose the items they are paying attention to the most (in this case, offender's or victim's payoffs), but also select related actions (decision to punish or compensate) that are a consequence of the items they are attending. We also showed that people display an offender bias, meaning that they are more attracted by the offender's payoff. Both these effects were confirmed in the behavioural experiments 3-5 (information frame): people who choose to reveal the offender's payoff are more likely to punish, and people are more likely to reveal the offender's payoff to begin with. These findings are in line with the idea that top-down attentional mechanisms and intrinsic motivation can explain information selection and choice in a decisional process (e.g., Coricelli et al. 2020).

The current findings also support the hypothesis that exogenously forcing the attentional focus to switch from one payoff to the other by manipulating task-irrelevant factors can influence decision-making. Whilst we did not find clear evidence that visual representation affects the choice to punish or compensate in the eye-tracking experiments (although we do clearly find this effect in the information frame paradigms in experiment 3-5), we found an effect on the amount spent, suggesting that a bottom-up influence might be more effective on the more emotional aspect of the decision, which is not the reaction per se, but the severity of the reaction (Stallen et al, 2018; Civai et al, 2019).The results of two mini meta-analyses on the effects of choice and severity across the five experiments, which are reported in the SM (7.2 and 7.3), confirm this interpretation. In the behavioural experiments, this bottom-up effect is clearly demonstrated by the directional effect of automatically encode information on the observed choices. In fact, when participants are not able to decide which piece of information to reveal (as in the BU experiments), and their focus of attention automatically lands on either the offender or the victim, their decisions are heavily influenced by the available information. Specifically, they are less likely to punish when the victim's payoff is randomly displayed compared to when the offender's payoff is revealed. Interestingly, there is no preference reversal, meaning that, when the victim's payoff is revealed, people do not show a preference for compensation, not even when they choose to reveal the victim's payoff themselves (TD experiment). This might suggest that punishment is indeed driven by very powerful forces, potentially being associated with a much higher emotional activation, as suggested by previous studies (e.g., Stallen et al, 2018; Hallsson et al, 2018; Capraro, 2024). Nevertheless, we show that being exposed to an alternative piece of information does have a significant effect in reducing the observed preference for punishment. Further investigation is needed in order to shed light on the cognitive and emotional mechanisms of this shift: for example, being exposed to the victim's information may trigger compassion or some form of identification that leads people to act more prosocially by helping the person in need (e.g., Wang et al, 2024).

These findings show that people do not approach these situations neutrally: they tend to be more attracted by the active player, or agent (offender), rather than by the powerless and passive player, as victims often are. This bias towards the offender may be an agency bias and have an evolutionary explanation: active players are the ones who initiate changes in the environment to which we need to respond or react. Therefore, paying attention to active players can be crucial for survival. This interpretation is supported by evidence from social cognition development, showing that we are able to pay attention to and imitate bodily acts from early infancy (Meltzoff & Brooks, 2001). This suggests that, in order to increase the engagement with the victim's condition, this needs to be of an order of magnitude more attractive than the offender's. Future research can be carried out to calculate the precise relationship between these magnitudes of attraction, and the factors that might mediate them.

Another potential explanation for the preference for punishment and the focus on the offender's payoff may be competitiveness: in the current set-up, participants may have decided to punish the offender more often because, every time the offender takes money from the victim, they become the richer player in the game, and therefore participants may react against their loss in status rather than the unfairness per se. However, competitiveness does not explain the current findings in full: in fact, when participants are forced to focus on the victim, the preference for punishment disappears, even though the loss of status is still factual. This suggests that competitiveness may enter the equation only when the improved financial status of one player is salient, i.e., visible. Nevertheless, future studies may explore this aspect to explain this kind of preferences.

Regarding individual differences in empathy, our results were unclear and inconsistent across experiments. This may be partially due to the minimal social cues presented in the task, with participants facing only digits or coins without any additional detail (e.g., names, faces, or more elaborated scenarios). As a result, participants may have made their decision with limited emotional perspective-taking. However, our exploratory findings suggest that affective empathy may act as a moderator of the framing effect, enhancing the likelihood of punishment/compensation when the offender's/the victim's payoffs were presented; this is in line with an interpretation of empathy that sees shared affect as a key factor in sharing the other's perspective and orienting a person's attention to aspects of the environment that are important for the other (Kiverstein, 2015).

Despite having no direct evidence that this manipulation would work outside of a controlled experimental environment, findings on news framing described in the introduction (paragraph 1.4.2) show that these attentional and framing effects are relevant in more natural setting, where people's moral judgment and attitude around different issues like immigration or gun violence are influenced by journalistic choices (see Lecheler & De Vreese, 2019 for a review). The strength of the current findings is to show that these effects emerge even when using a paradigm where stimuli are presented in a rapid sequence and require minimal depth of information processing and emotional involvement, mirroring the condition in which we often consume news.

All five experiments have limitations, both methodological and theoretical. For example, in experiments 1 and 2, the two conditions coins and digits are administered using a within-participants design, and therefore a carryover effect from one condition to the other cannot be excluded. As reported in experiment one's pre-registration, two previous studies employing a between-participants design had been conducted to test the materials (Civai & Johns, 2018): the main results showed an increased attentional bias towards the offender when the payoff was represented as coins, as well as an increased preference towards punishment in the coin condition, suggesting that main findings are minimally affected by the design. From a more theoretical perspective, we adopted a behavioural economics approach, which allowed us to operationalize reactions to justice violations in terms of value-based choice; whilst this approach enabled us to investigate situations where both options may be reasonable reactions to unfairness, it also sacrificed the complexity of real-life situations, where, at times, one option is simply not a feasible substitute for the other (e.g., murder). Additional studies may continue exploring these processes in more naturalistic settings and include a broader range of situations, such as different types of moral violations. Finally, we also limited participation to residents in the UK since, wherever possible (experiments 3-5), we aimed to capture a representative sample of the population in terms of age, gender, and ethnicity to increase generalisability. We recognise that for any behaviour, and in particular social and moral behaviour, culture is a key factor of influence; for this reason, future studies should aim to incorporate data from other countries and other cultures (e.g., collectivistic cultures), and focus on the cultural influence of these preferences, to increase the generalizability of the current claims.

To conclude, the way in which we experience news and information nowadays is increasingly more personalised, since the information we are exposed to is based on our interests and previous choices, a phenomenon that has been defined as an information bubble (Pariser, 2011), that creates echo-chambers (Sunstein, 2018), and that is becoming even likelier with advances in generative artificial intelligence (Capraro et al., 2024). The current findings support the idea that exposing people to information they would not have chosen in the first place can indeed change their decisions, offering a way forward for whoever is interested in building algorithms that burst information bubbles.

References

1. Alós-Ferrer, C., Jaudas, A., & Ritschel, A. (2021). Attentional shifts and preference reversals: An eye-tracking study. *Judgment and Decision Making, 16*(1), 57-93.
2. Alós-Ferrer, C., & Ritschel, A. (2022). Attention and salience in preference reversals. *Experimental Economics, 25*(3), 1024-1051.
3. Amir, O., Rand, D. G., & Gal, Y. A. K. (2012). Economic games on the internet: The effect of $1 stakes. *PloS one, 7*(2), e31461.

4. Anwyl-Irvine, A., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. (2018). Gorillas in our midst: Gorilla. sc. *Behavior Research Methods*.
5. Arieli, A., Ben-Ami, Y., & Rubinstein, A. (2011). Tracking decision makers under uncertainty. *American Economic Journal: Microeconomics, 3*(4), 68-76.
6. Armel, K. C., Beaumel, A., & Rangel, A. (2008). Biasing simple choices by manipulating relative visual attention. *Judgment and Decision making*, *3*(5), 396-403.
7. Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.
8. Beratšová, A., Krchová, K., Gažová, N., & Jirásek, M. (2016). Framing and bias: a literature review of recent findings. *Central European journal of management*, *3*(2).
9. Capraro, V. (2024). The dual-process approach to human sociality: Meta-analytic evidence for a theory of internalized heuristics for self-preservation. *Journal of Personality and Social Psychology*.
10. Capraro, V., Lentsch, A., Acemoglu, D., Akgun, S., Akhmedova, A., Bilancini, E., ... & Viale, R. (2024). The impact of generative artificial intelligence on socioeconomic inequalities and policy making. *PNAS Nexus*, *3*(6).
11. Chavez, A. K., & Bicchieri, C. (2013). Third-party sanctioning and compensation behavior: Findings from the ultimatum game. *Journal of Economic Psychology*, *39*, 268-277.
12. Civai, C., & Johns, P. (2018, July 19-22). *Attentional correlates of third-party punishment and compensation* [Conference presentation]. Society for the Advancement of Behavioral Economics/International Association for Research in Economic Psychology (SABE / IAREP), Middlesex University, London.
13. Civai, C., Huijsmans, I., & Sanfey, A. G. (2019). Neurocognitive mechanisms of reactions to second-and third-party justice violations. *Scientific Reports*, *9*(1), 1-11.

14. Civai, C., Teodorini, R., & Carrus, E. (2020). Does unfairness sound wrong? A cross-domain investigation of expectations in music and social decision-making. *Royal Society open science, 7*(9), 190048.

15. Coricelli, G., Polonio, L., & Vostroknutov, A. (2020). The process of choice in games. In *Handbook of experimental game theory*. Edward Elgar Publishing.

16. David, B., Hu, Y., Krüger, F., & Weber, B. (2017). Other-regarding attention focus modulates third-party altruistic choice: an fMRI study. *Scientific Reports, 7*(1), 43024.

17. Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, *44*, 113–126. doi:10.1037/0022-3514.44.1.113

18. Decety, J., & Yoder, K. (2015). Empathy and motivation for justice: Cognitive empathy and concern, but not emotional empathy, predict sensitivity to injustice for others. *Social Neuroscience*, *11*(1), 1–14.

19. Devetag, G., Di Guida, S., & Polonio, L. (2016). An eye-tracking study of feature-based choice in one-shot games. *Experimental Economics*, *19*, 177-201.

20. Eyal, P., David, R., Andrew, G., Zak, E., & Ekaterina, D. (2021). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, 1-20.

21. Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, *39*(2), 175-191.

22. Fehr, E., & Fischbacher, U. (2004). Social norms and human cooperation. *Trends in cognitive sciences, 8*(4), 185-190.

23. FeldmanHall, O., Sokol-Hessner, P., Van Bavel, J. J., & Phelps, E. A. (2014). Fairness violations elicit greater punishment on behalf of another than for oneself. *Nature communications*, *5*(1), 1-6.

24. FeldmanHall, O., Dalgleish, T., Evans, D., & Mobbs, D. (2015). Empathic concern drives costly altruism. *Neuroimage*, *105*, 347-356.
25. Fiedler, S., Glöckner, A., Nicklisch, A., & Dickert, S. (2013). Social value orientation and information search in social dilemmas: An eye-tracking analysis. *Organizational behavior and human decision processes*, *120*(2), 272-284.
26. Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior*, *6*(3), 347-369.
27. Ghaffari, M., & Fiedler, S. (2018). The power of attention: Using eye gaze to predict other-regarding and moral choices. *Psychological science*, *29*(11), 1878-1889.
28. Gillis, M. T., & Hettler, P. L. (2007). Hypothetical and real incentives in the ultimatum game and Andreoni's public goods game: an experimental study. *Eastern Economic Journal*, *33*(4), 491-510.
29. Gromet, D. M., & Darley, J. M. (2009). Punishment and beyond: Achieving justice through the satisfaction of multiple goals. *Law & Society Review*, *43*(1), 1-38
30. Gummerum, M., López-Pérez, B., Van Dijk, E., & Van Dillen, L. F. (2022). Ire and punishment: incidental anger and costly punishment in children, adolescents, and adults. *Journal of Experimental Child Psychology*, *218*, 105376.
31. Hallsson, B. G., Siebner, H. R., & Hulme, O. J. (2018). Fairness, fast and slow: A review of dual process models of fairness. *Neuroscience & Biobehavioral Reviews*, *89*, 49-60.
32. Hu, Y., Strang, S., & Weber, B. (2015). Helping or punishing strangers: neural correlates of altruistic decisions as third-party and of its relation to empathic concern. *Frontiers in behavioral neuroscience*, *9*, 24.
33. Hu, Y., Fiedler, S., & Weber, B. (2020). What drives the (un) empathic bystander to intervene? Insights from eye tracking. *British journal of social psychology*, *59*(3), 733-751.

34. Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research, 40*(10-12), 1489-1506.
35. Jarvenpaa, S. L. (1990). Graphic displays in decision making — the visual salience effect. *Journal of Behavioral Decision Making, 3*(4), 247-262.
36. JASP Team (2022). JASP (Version 0.16.3)[Computer software].
37. Jiang, T., Potters, J., & Funaki, Y. (2016). Eye-tracking social preferences. *Journal of Behavioral Decision Making*, *29*(2-3), 157-168.
38. Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature, 530*(7591), 473-476.
39. Kim, H. J., & Cameron, G. T. (2011). Emotions matter in crisis: The role of anger and sadness in the publics' response to crisis news framing and corporate crisis response. *Communication Research, 38*(6), 826-855.
40. Kiverstein, J. (2015). Empathy and the responsiveness to social affordances. *Consciousness and Cognition*, *36*, 532-542.
41. Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature neuroscience*, *13*(10), 1292-1298.
42. Krupka, E. L., & Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary?. *Journal of the European Economic Association*, *11*(3), 495-524.
43. Kühne, R., & Schemer, C. (2015). The emotional effects of news frames on information processing and opinion formation. *Communication Research*, *42*(3), 387-407.
44. Lecheler, S., Bos, L., & Vliegenthart, R. (2015). The mediating role of emotions: News framing effects on opinions about immigration. *Journalism & Mass Communication Quarterly*, *92*(4), 812-838.

45. Lecheler, S., & De Vreese, C. H. (2019). *News framing effects: Theory and practice.* Routledge.

46. Leliveld, M. C., van Dijk, E., & van Beest, I. (2012). Punishing and compensating others at your own expense: The role of empathic concern on reactions to distributive injustice. *European Journal of Social Psychology*, *42*(2), 135-140.

47. Li, X., & Camerer, C. F. (2022). Predictable effects of visual salience in experimental decisions and games. *The Quarterly Journal of Economics*, *137*(3), 1849-1900.

48. Lim, D., & DeSteno, D. (2016). Suffering and compassion: The links among adverse life experiences, empathy, compassion, and prosocial behavior. *Emotion*, *16*(2), 175.

49. Liu, S., Guo, L., Mays, K., Betke, M., & Wijaya, D. T. (2019, January). Detecting frames in news headlines and its application to analyzing news framing trends surrounding US gun violence. In *Proceedings of the 23rd conference on computational natural language learning (CoNLL)*.

50. Lotz, S., Okimoto, T. G., Schlösser, T., & Fetchenhauer, D. (2011). Punitive versus compensatory reactions to injustice: Emotional antecedents to third-party interventions. *Journal of experimental social psychology*, *47*(2), 477-480.

51. Lu, T., & McKeown, S. (2018). The effects of empathy, perceived injustice and group identity on altruistic preferences: Towards compensation or punishment. *Journal of Applied Social Psychology*, *48*(12), 683-691.

52. Lüdecke, D. (2020). sjPlot: Data visualization for statistics in social science (R Package Version, 2.1)[Computer software].

53. Marchiori, D., Di Guida, S., Polonio, L. (2021) Plasticity of strategic sophistication in interactive decision-making. *Journal of Economic Theory*, *196*, 105291.

54. Meltzoff, A. N., & Brooks, R. (2001). “Like me” as a building block for understanding other minds: Bodily acts, attention, and intention. In B. F. Malle, L. J. Moses, & D. A.

Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 171–191). Cambridge, MA: MIT Press.

55. Milosavljevic, M., Navalpakkam, V., Koch, C., & Rangel, A. (2012). Relative visual saliency differences induce sizable bias in consumer choice. *Journal of Consumer Psychology*, *22*(1), 67-74.
56. Nelissen, R. M. (2008). The price you pay: Cost-dependent reputation effects of altruistic punishment. *Evolution and Human Behavior, 29*(4), 242-248.
57. Palan, S., & Schitter, C. (2018). Prolific. ac—A subject pool for online experiments. Journal of Behavioral and Experimental Finance, 17, 22-27.
58. Papoutsaki, A., Laskey, J., & Huang, J. (2017, March). Searchgazer: Webcam eye tracking for remote studies of web search. In *Proceedings of the 2017 conference on conference human information interaction and retrieval* (pp. 17-26).
59. Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin UK.
60. Pärnamets, P., Johansson, P., Hall, L., Balkenius, C., Spivey, M. J., & Richardson, D. C. (2015). Biasing moral decisions by exploiting the dynamics of eye gaze. *Proceedings of the National Academy of Sciences*, *112*(13), 4170-4175.
61. Pittarello, A., Motro, D., Rubaltelli, E., & Pluchino, P. (2016). The relationship between attention allocation and cheating. *Psychonomic Bulletin & Review*, *23*(2), 609-616.
62. Polonio, L., Di Guida, S., & Coricelli, G. (2015). Strategic sophistication and attention in games: An eye-tracking study. *Games and Economic Behavior, 94*, 80-96.
63. Polonio, L., & Coricelli, G. (2019). Testing the level of consistency between choices and beliefs in games using eye-tracking. *Games and Economic Behavior*, *113*, 566-586.
64. Rahal, R. M., & Fiedler, S. (2019). Understanding cognitive and affective mechanisms in social psychology through eye-tracking. *Journal of Experimental Social Psychology*, *85*, 103842.

65. Reniers, R. L., Corcoran, R., Drake, R., Shryane, N. M., & Völlm, B. A. (2011). The QCAE: A questionnaire of cognitive and affective empathy. *Journal of personality assessment*, *93*(1), 84-95.
66. Stallen, M., Rossi, F., Heijne, A., Smidts, A., De Dreu, C. K., & Sanfey, A. G. (2018). Neurobiological mechanisms of responding to injustice. *Journal of Neuroscience*, *38*(12), 2944-2954.
67. Stewart, N., Gächter, S., Noguchi, T., & Mullett, T. L. (2016). Eye movements in strategic choice. *Journal of behavioral decision making*, *29*(2-3), 137-156.
68. Stroud, N. J. (2017). Attention as a valuable resource. *Political Communication*, *34*(3), 479-489.
69. Sunstein, C. R. (2018). # Republic. In *# Republic*. Princeton university press.
70. Teoh, Y. Y., Yao, Z., Cunningham, W. A., & Hutcherson, C. A. (2020). Attentional priorities drive effects of time pressure on altruistic choice. *Nature communications*, *11*(1), 1-13.
71. Thulin, E. W., & Bicchieri, C. (2016). I'm so angry I could help you: Moral outrage as a driver of victim compensation. *Social Philosophy and Policy*, *32*(2), 146-160
72. Van Doorn, J., & Brouwers, L. (2017). Third-party responses to injustice: a review on the preference for compensation. *Crime Psychology Review*, *3*(1), 59-77.
73. Van Doorn, J., Zeelenberg, M., & Breugelmans, S. M. (2018). An exploration of third parties' preference for compensation over punishment: six experimental demonstrations. *Theory and decision*, *85*(3), 333-351.
74. Wang, H., Wu, X., Xu, J., Zhu, R., Zhang, S., Xu, Z., ... & Liu, C. (2024). Acute stress during witnessing injustice shifts third-party interventions from punishing the perpetrator to helping the victim. *PLoS biology*, *22*(5), e3002195.

75. Weeks, B. E., & Lane, D. S. (2020). The ecology of incidental exposure to news in digital media environments. *Journalism*, *21*(8), 1119-1135.

76. Will, G. J., Crone, E. A., van den Bos, W., & Güroğlu, B. (2013). Acting on observed social exclusion: Developmental perspectives on punishment of excluders and compensation of victims. *Developmental psychology*, *49*(12), 2236.

77. Yang, X., & Krajbich, I. (2021). Webcam-based online eye-tracking for behavioral research. *Judgment and Decision making*, *16*(6), 1485-1505.

78. Zonca, J., Coricelli, G. & Polonio, L. (2019). Does exposure to alternative decision rules change gaze patterns and behavioral strategies in games? *Journal of the Economic Science Association*, *5*(1), 14-25.

79. Zonca, J., Coricelli, G., & Polonio, L. (2020a). Gaze data reveal individual differences in relational representation processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(2), 257–279.

80. Zonca, J., Coricelli, G., & Polonio, L. (2020b). Gaze patterns disclose the link between cognitive reflection and sophistication in strategic interaction. *Judgment and Decision Making*, *15*(2), 230 – 245.