eLife

Reviewed Preprint
v2 • July 14, 2025
Revised by authors

Reviewed Preprint
v1 • November 14, 2024

Neuroscience

# Oxytocin restores context-specific hyperaltruistic preference

**Hong Zhang, Yinmei Ni, Jian Li**

School of Psychological and Cognitive Sciences and Beijing Key Laboratory of Behavior and Mental Health, Peking University, Beijing, China • IDG/McGovern Institute for Brain Research, Peking University, Beijing, China



**eLife Assessment**

This revised paper provides **valuable** findings that altruistic tendency during moral decision-making is gain/loss context-dependent and oxytocin can restore the absence of altruistic choices in the loss domain. The methods and analyses are **solid**, yet the study could still benefit from better overall framing and more clarity and precision in the definition of key constructs, as pointed out by reviewers. If these concerns are addressed, this study would be of interest to social scientists and neuroscientists who work on moral decision-making and oxytocin.

https://doi.org/10.7554/eLife.102756.2.sa4

## Abstract

An intriguing advancement in recent moral decision-making research suggests that people are more willing to sacrifice monetary gains to spare others from suffering than to spare themselves, yielding the hyperaltruistic tendency. Other studies, however, indicate an opposite egoistic bias in that subjects are less willing to harm themselves for the benefits of others than for their own benefits. These results highlight the delicate inner workings of moral decision and call for a mechanistic account of hyperaltruistic preference. We investigated the boundary conditions of hyperaltruism by presenting subjects with trade-off choices combing monetary gains and painful electric shocks, or, choices combing monetary losses and shocks. We first showed in study 1 that switching the decision context from gains to losses effectively eliminated the hyperaltruistic preference and the decision context effect was associated with the altered relationship between subjects' instrumental harm (IH) trait attitudes and their relative pain sensitivities. In the pre-registered study 2, we tested whether oxytocin, a neuropeptide linked to parochial altruism, might restore the context-dependent hyperaltruistic preference. We found that oxytocin increased subjects' reported levels of framing the task as harming (vs. helping) others, which mediated the correlation between IH and relative pain sensitivities. Thus, the loss decision context and oxytocin diminished and restored the mediation effect of subjective harm framing, respectively. Our results help to elucidate the psychological processes underpinning the contextual specificity of hyperaltruism and carry implications in promoting prosocial interactions in our society.

## Introduction

The Chinese proverb "A virtuous man acquires wealth in an upright and just way" stresses the universal moral code of refraining from harming others for personal gain[1]. Disregard of the suffering of others is often associated with aggressive and antisocial tendencies in psychopathy[2,3]. Recent studies further developed the moral decision theory and suggested that people are willing to pay more to reduce other's pain than their own pain, yielding a "hyperaltruistic" preference in moral dilemmas[4–7]. For example, it was shown that the hyperaltruistic preference modulated neural representations of the profit gained from harming others via the functional connectivity between the lateral prefrontal cortex, a brain area involved in moral norm violation, and profit sensitive brain regions such as the dorsal striatum[6]. However, moral norm is also context dependent: vandalism is clearly against social and moral norms yet vandalism for self-defense is more likely to be ethically and legally justified (the Doctrine of necessity). Therefore, a crucial step is to understand the boundary conditions for hyperaltruism. We set out to address this question by examining how the moral perception of decision context affects people's hyperaltruistic prefernence[8–12]. More importantly, we also test whether the contextual effect of hyperaltruistic disposition is susceptible to oxytocin, a neuropeptide heavily implicated in social bounding and social cognition[13–15].

Classic moral dilemmas often involve the tradeoff between personal material wellbeing and the adherence of social norm (or moral principle)[16,17]. Previous studies have shown that the moral preference can be highly context specific. For example, studies showed that people were more likely to engage in unethical behavior (lie or cheat) to avoid monetary loss than to secure monetary gain[18–20]. Other research, instead, found that it was more common for people to help others from losing money than to help them to gain money [21,22]. However, the boundary conditions for hyperaltruism remains elusive[4–7]. Hyperaltruistic disposition was typically measured in a money-pain trade-off task where individuals' monetary gain was pitted against the physical pain experienced by others and themselves. In turn, the hyperaltruistic preference was calculated by comparing the amounts of monetary gain subjects were willing to forgo to reduce others' pain relative to the reduction of their own pain. Therefore, it is likely that replacing monetary gain with monetary loss in the money-pain trade-off task might bias subjects' hyperaltruistic preference due to stronger sensitivity toward monetary loss (loss aversion) or highlighted vigilance in the face of potential loss[23–28]. Indeed, elevated attention to losses can alternate subjects' sensitivities to the general reinforcement structure and have distinct effects on their arousal and performance[27]. Additionally, prospective losses can elicit negative emotions[29,30], which may subsequently influence how individuals evaluate their own pain relative to others'. For example, studies have shown that negative mood can both increase pain sensitivities and decrease empathy for others, both of which yield opposing effects on hyperaltruistic preferences[31–35]. Finally, how the moral dilemma is framed can be a critical factor influencing moral decision-making[36–38]. For instance, explicitly manipulating moral frames as "harming others" (harm frame) or "not helping others" (help frame) resulted in a significant social framing effect such that subjects showed stronger prosocial preference in the harm frame compared to the help frame[25,39]. It is thus possible that people may implicitly form moral impressions (help or harm) based on the decision contexts and adjust their behavior accordingly. This way, the contextual effect on hyperaltruistic preference can be associated with individuals' internal framing of moral contexts.

Oxytocin, a neuropeptide synthesized in the hypothalamus, has been shown to modulate a variety of emotional, cognitive and social behaviors through distributed receptors in various brain areas[40,41]. Oxytocin has been shown to play a critical role in social interactions such as maternal attachment, pair bonding, consociate attachment and aggression in a variety of animal models[42,43]. Humans are endowed with higher cognitive and affective capacities and exhibit

far more complex social cognitive patterns[44]. It has been suggested that oxytocin increases prosocial behaviors such as trust and altruism by enabling individuals to overcome proximity avoidance, enhancing empathy for others' suffering, and reducing the processing of negative stimuli (fearful or angry faces) [45–52]. However, the evidence for whether and how oxytocin influences hyperaltruism remains scarce[53]. This may be partly because the prosocial effect of oxytocin seems to be both personality trait and decision context dependent[13,54–57]. For example, recent studies have shown that oxytocin both promotes in-group cooperation and defensive aggression toward competing out-group members[58–60]. It is therefore plausible that oxytocin might also exert a context-dependent effect on hyperaltruistic preference in a moral decision task. Specifically, oxytocin might influence the way subjects internally frame the task as benefiting from others' pain or sacrificing self- interest to avoid others' harm given the widely reported link between oxytocin and empathy[46,61–63]. Also, previous literature on the prosocial effect of oxytocin dovetails nicely with the utilitarian moral literature suggesting that moral behavior is closely related to people's personality traits such as the instrumental harm attitude (IH), the degree to which people are willing to compromise their moral beliefs by inflicting harm on others to achieve better outcomes[1].

We conducted two studies to test the above hypotheses. In study 1, we manipulated the decision context of the monetary valence (from making monetary gains to avoiding monetary losses) in a well-established money-pain trade-off task to examine the contextual specificity of the hyperaltruistic preferences (**Fig. 1A**). We compared subjects' hyperaltruistic preferences in decision scenarios where options were constructed such that higher monetary gains (or smaller monetary loss in the loss context) were associated with more painful electric shocks (more pain option), and lower monetary gains (or bigger monetary loss in the loss context) were associated with less painful shocks (less pain option). We found that in line with results reported in previous studies, subjects showed clear hyperaltruistic preference in the gain context. However, shifting the decision context from gains to losses eliminated such preference. In study 2, we employed a pre-registered, placebo-controlled experimental design to examine how oxytocin might modulate the context effect of hyperaltruism (**Fig. 1B**). We found that oxytocin had no effect on hyperaltruistic preference in the gain context. However, oxytocin restored subjects' hyperaltruistic preferences in the loss context.

Importantly, oxytocin administration prompted subjects to more likely perceive the task structure as harming others, which in turn mediated their hyperaltruistic preferences.

# Results

Prior to the money-pain trade-off task, we individually calibrated each subject's pain threshold using a standard procedure[4–6]. This allowed us to tailor a moderate electric stimulus that corresponded to each subject's subjective pain intensity. Subjects then engaged in 240 decision trials (60 trials per condition), acting as the "decider" and trading off between monetary gains or losses for themselves and the pain experienced by either themselves or an anonymous "pain receiver" (gain-self, gain-other, loss-self and loss-other, see Supplementary Fig. 9 and methods section for details).

## Loss context eliminated hyperaltruistic preference

We first tested whether the decision context (gain or loss) directly affected subjects' hyperaltruistic tendencies by examining the proportion of trials in which subjects chose the less painful option both for themselves and for others (**Fig. 1A**). Since the choice sets for the self- and other-recipient are the same, subjects' hyperaltruistic tendencies can be examined by the choice differences regarding different shock recipients (self vs. other). A larger choice difference between other- and self-recipients indicates a stronger hyperaltruistic tendencies in both the gain and loss

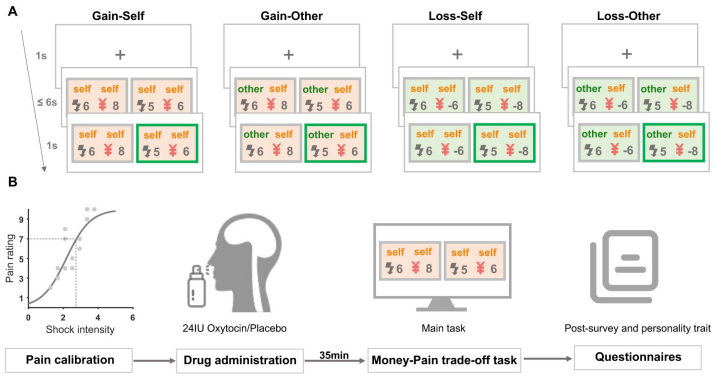



**Fig. 1.**

### Experimental design and task.

(**A**) Experimental Task. Subjects performed a money-pain trade-off task in which they were designated as deciders. Four conditions (Gain-Self, Gain-Other, Loss-Self, Loss-Other) were introduced across decision contexts (gain vs. loss) and shock-recipients (self vs. other). In each trial, subjects were asked to choose between two options with various amounts of monetary and harm consequences. The chosen option was highlighted for 1s after subjects' decisions. (**B**) Procedures of the oxytocin study (study 2). Before the task, a pain calibration procedure was performed on each subject to determine their pain thresholds for electrical shock stimuli. Subjects were then administered with 24IU oxytocin nasal spray or placebo (saline). Thirty-five minutes later, subjects commenced the money-pain trade-off task. Finally, they filled out questionnaires including post-task surveys and assessments of personality traits.

contexts. The main effects of the shock recipient (self vs. other) and decision context (gain vs. loss) were both significant (recipient: $F_{1,79}$ = 6.625, $P$ = 0.012, $\eta^2$ = 0.077 ; valence: $F_{1,79}$ = 8.379, $P$ = 0.005, $\eta^2$ = 0.096), indicating that subjects were in general hyperaltruistic across decision contexts and more likely to choose the more painful option in the loss context (**Fig. 2A**). There was also a significant recipient × valence interaction effect ($F_{1,79}$ = 33.047, $P$ < 0.001, $\eta^2$ = 0.295), suggesting that the loss context (vs. gain context) significantly decreased subjects' hyperaltruistic tendencies (**Fig. 2A**). Further simple effect analysis confirmed that hyperaltruism only existed in the gain context ($F_{1,79}$ = 16.798, $P$ < 0.001, $\eta^2$ = 0.175) but was eliminated in the loss context ($F_{1,79}$ = 0.650, $P$ = 0.423, $\eta^2$ = 0.008). We also modeled subjects' choices using an influential model where subjects' behavior could be characterized by the harm (electric shock) aversion parameter κ, reflecting the relative weights subjects assigned to Δm and Δs, the objective difference in money and shocks between the more and less painful options, respectively (ΔV = (1- κ)Δm - κΔs Eq.1, See Methods for details)[4–6]. Higher κ indicates that higher sensitivity is assigned to Δs than Δm and vice versa. Consistent with previous literature, we found that the harm aversion parameter κ for the other-recipient was significantly greater than that of the self-recipient in the gain context[4–6]; however, such harm aversion asymmetry between self- and other-recipient disappeared in the loss context ($\kappa_{other}$ - $\kappa_{self}$ > 0 in the gain context: $t_{79}$=3.869, $P$ < 0.001 ; $\kappa_{other}$ and $\kappa_{self}$ showed no difference in the loss context: $t_{79}$ = 0.834, $P$ = 0.407; $\kappa_{other}$ - $\kappa_{self}$ showed context difference: $t_{79}$ = 5.591, $P$ < 0.001 **Fig. 2B**).

To further identify the distinct effects of Δm and Δs in biasing subjects' choice behavior, we also ran a mixed-effect logistic regression analysis to examine how the change of decision context (gain vs. loss) affected each individual's sensitivities to money (Δm) and electric shock (Δs). Subjects' choices were regressed against Δm and Δs and this analysis yielded significant effects on relative harm sensitivity (*other* $\beta_{\Delta}$ - *self* $\beta_{\Delta}$: difference of regression coefficients of Δs in the other- and self-conditions) in the gain context but not in the loss context (gain context: $t_{79}$ = 3.298, $P$ = 0.001; loss context: $t_{79}$ = 0.015, $P$ = 0.988) and significant decision context effect ($t_{79}$ = 3.630, $P$ = 0.001; **Fig. 2C**). On the contrary, decision context did not have an effect on subjects' relative sensitivities towards monetary (Δm) difference ($t_{79}$ = 1.480, $P$ = 0.143), though the relative money sensitivities (*other* $\beta_{\Delta+}$- *self* $\beta_{\Delta+}$ : difference of regression coefficients of Δm in the other- and self-condition) were significant in both contexts (gain context: $t_{79}$ = 5.241, $P$ < 0.001; loss context: $t_{79}$ = 4.355, $P$ < 0.001; **Fig. 2D**).

These results suggest that contrary to the prediction of loss aversion, the relative money sensitivity did not change when the task structure switched from monetary gains to losses. Instead, subjects' susceptibilities to harm might underlie the diminishment of hyperaltruism across decision contexts (also see Supplementary Fig. 1).

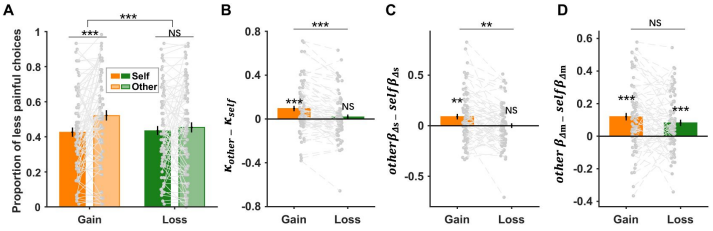
## Individual differences in moral preference

Recent theoretical development in moral decision-making stresses the importance of distinctive dimensions of instrumental harm and impartial beneficence in driving people's moral preference[64,65]. The instrumental harm (IH) and impartial beneficence (IB) components of the Oxford utilitarianism scale[1] of moral psychology were used to measure the extent to which individuals are willing to compromise their moral beliefs by breaking the rules or inflicting harm on others to achieve better outcomes (IH) and the extent of the impartial concern for the well-being of everyone (IB), respectively. We tested how hyperaltruism was related to both IH and IB across decision contexts using an exploratory multiple regression analysis. Moral preference, defined as $\kappa_{other}$ -$\kappa_{self}$, was negatively associated with IH ($\beta$ = -0.031 ± 0.011, $t$156 =-2.784, $P$ = 0.006) but not with IB ($\beta$ = 0.008 ± 0.016, $t$156 =0.475, $P$ = 0.636) across gain and loss contexts, reflecting a general connection between moral preference and IH (**Fig. 3A** & B). Note that the interaction effects of valence × IH and valence × IB were not significant (valence × IH effect: $\beta$ = 0.016 ± 0.022, $t$152 = 0.726, $P$ = 0.469, **Fig. 3A**; valence × IB effect: $\beta$ = -0.004 ± 0.031, $t$152 = -0.115, $P$ = 0.908, **Fig. 3B**).

## Fig. 2.

### Context specific hyperaltruistic preferences.

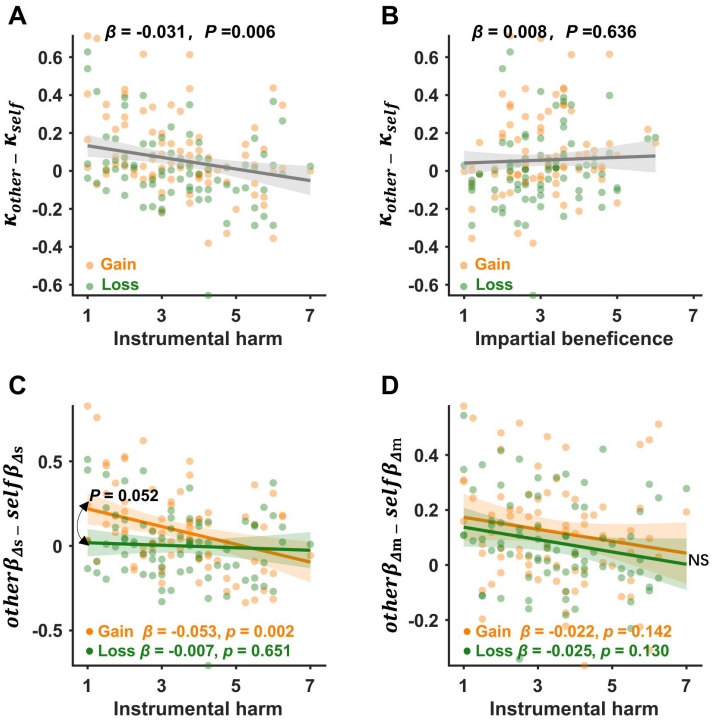
(**A**) In the gain context, subjects chose the less painful option more frequently for others than for themselves, demonstrating a hyperaltruistic preference. However, this tendency was not observed in the loss context. (**B**) The harm aversion parameter κ for others was significantly greater than that of self in the gain but not the loss context. (C) Furthermore, the relative harm sensitivity, calculated as the difference of regression coefficients of Δs in the other- and self-conditions ($other\beta_{\Delta s} - self\beta_{\Delta s}$) was significant in the gain context, but not in the loss context. (**D**) However, the relative money sensitivity, the difference of regression coefficients of Δm in the other- and self-conditions ($other\beta_{\Delta m} - self\beta_{\Delta m}$), did not show contextual specificity. Error bars represent SE across subjects. NS, not significant; ** $P < 0.01$ and *** $P < 0.001$.




## Fig. 3.

### Individual difference in moral preferences.

(**A-B**) Hyperaltruism ($\kappa_{\$\%\&'(} - \kappa_{)'*+}$) was negatively associated with instrumental harm (IH) attitudes but showed no significant relationship with impartial beneficence (IB). (**C**) The correlation between IH and subjects' relative harm sensitivities ($other\beta_{\Delta s} - self\beta_{\Delta s}$) was marginally different between the gain and loss contexts. (D) However, the correlation between IH and subjects' monetary sensitivities ($other\beta_{\Delta m} - self\beta_{\Delta m}$) showed no context difference. The regression coefficients in Figure A, C & D showed the relationship between IH and moral behavior, controlling for empathic concern (EC) and impartial beneficence (IB), while the coefficient in Figure B showed the relationship between IB and moral preference, controlling for EC and IH. NS, not significant.



Interestingly, by examining the separate contributions of Δm and Δs to moral preferences, we found that the decision context modulated the relationship between IH and subjects' relative harm sensitivities (gain context: $\beta$ = -0.053 ± 0.016, $t$152 =-3.224, $P$ = 0.002; loss context: $\beta$ = -0.007 ± 0.016, $t$152 =-0.453, $P$ = 0.651; valence × IH: $\beta$ = 0.045 ±0.023, $t$152= 1.959, $P$ = 0.052 ; **Fig. 3C**), yet it did not modulate the relationship between IB and subjects' relative harm sensitivities (gain context: $\beta$ = 0.019 ± 0.023, $t$152 =0.804, $P$ = 0.423; loss context: $\beta$ = 0.010 ± 0.023, $t$152 = 0.425, $P$ = 0.672; valence × IB: $\beta$=-0.009 ± 0.033, $t$152 = -0.268, $P$ = 0.789 ; also see Supplementary Fig. 3A). However, subjects' monetary sensitivities were not related with IH (gain context: $\beta$ = -0.022 ±0.015, $t$152 =-1.476, $P$ = 0.142; loss context: $\beta$ = -0.023 ±0.015, $t$152= -1.523, $P$ = 0.130; valence × IH: $\beta$ = -0.001 ± 0.021, $t$152 = -0.034, $P$ = 0.973 ; **Fig. 3D**), nor with IB (gain context: $\beta$ = -0.008 ± 0.021, $t$152 = -0.369, $P$ = 0.713 ; loss context: $\beta$ = -0.009 ± 0.021, $t$152 = -0.414, $P$ = 0.679 ; valence × IB: $\beta$ = -0.001 ± 0.030, $t$152 = -0.032, $P$ = 0.974 ; Supplementary Fig. 3B; also see Supplementary Fig. 2 and Table S1). These results suggest that the context dependent moral preference may be specifically due to the context modulation of subjects' relative harm sensitivities, accompanied by the altered correlation between relative harm sensitivities and IH across decision contexts. Furthermore, our results are consistent with the claim that profiting from inflicting pains on another person (IH) is inherently deemed immoral[1]. Hyperaltruistic preference, therefore, is likely to be associated with subjects' IH dispositions.

## Oxytocin restored hyperaltruistic preferences in the loss context

To probe how oxytocin might alter subjects' moral preference, we conducted the preregistered study 2 where a separate cohort of subjects performed the same task while undergoing the placebo/oxytocin administration in a with-subject design (N = 46, see Methods for experimental details). First, in the placebo session of study 2, we replicated the major findings of study 1 (**Fig. 2A**), demonstrating that hyperaltruistic preference was not significant in the loss context compared to the gain context (Placebo: valence × recipient interaction: $F_{1,45}$ = 9.103, $P$ = 0.004, $\eta^2$ = 0.168; simple effect: gain context $F_{1,45}$ = 9.356, $P$ = 0.004, $\eta^2$ = 0.172; loss context $F_{1,45}$ = 0.005, $P$ = 0.946, $\eta^2$ < 0.001; **Fig. 4A**). The main effect of treatment (placebo vs. oxytocin) was significant ($F_{1,45}$ = 41.961, $P$ < 0.001, $\eta^2$ = 0.483), indicating that the oxytocin treatment prompted subjects to select the more painful option more often. Interestingly, the treatment × recipient × valence interaction effect was also significant ($F_{1,45}$ = 6.309, $P$ = 0.016, $\eta^2$ = 0.123), suggesting that the recipient × valence interaction might be different due to the placebo and oxytocin treatments. Indeed, with the administration of oxytocin, the hyperaltruistic preference was restored in the loss context, without affecting the hyperaltruistic pattern in the gain context (oxytocin: valence × recipient: $F_{1,45}$ = 0.480, $P$ = 0.492, $\eta^2$ = 0.011; simple effect: gain context: $F_{1,45}$ = 25.364, $P$ < 0.001, $\eta^2$ = 0.360; loss context: $F_{1,45}$ = 24.408, $P$ < 0.001, $\eta^2$ = 0.352; **Fig. 4A**). These results together highlight that oxytocin might play an important valence-dependent modulatory role in restoring hyperaltruistic moral preference. Our modeling analysis revealed similar results: there was a significant treatment × valence interaction effect ($F_{1,45}$ = 6.349, $P$ = 0.015, $\eta^2$ = 0.124) on subjects' moral preference (defined as $\kappa_{other}$ - $\kappa_{self}$, **Fig. 4B**). With the administration of oxytocin, however, the diminished hyperaltruistic preference in the loss context (placebo: $F_{1,45}$ = 1.295, $P$ = 0.261, $\eta^2$ =0.028) was successfully restored (oxytocin: $F_{1,45}$=17.990, $P$ < 0.001, $\eta^2$ = 0.286, simple effect analysis, **Fig. 4B**). Additionally, the model-based moral preference was correlated with subjects' IH reports (but not IB) across subjects in both the placebo and oxytocin treatments (Supplementary Fig. 4 & 5, also see Table S2), replicating the results we obtained in study 1(**Fig. 3A** & B).

## Hyperaltruistic preference and increased harm sensitivities with Oxytocin administration

In study 1, we showed that the lack of hyperaltruism in the loss context was specifically related to the diminished relative harm sensitivities. In study 2, we also ran a mixed-effect logistic regression of subjects' choices against continuous independent variables including Δm, Δs and categorical

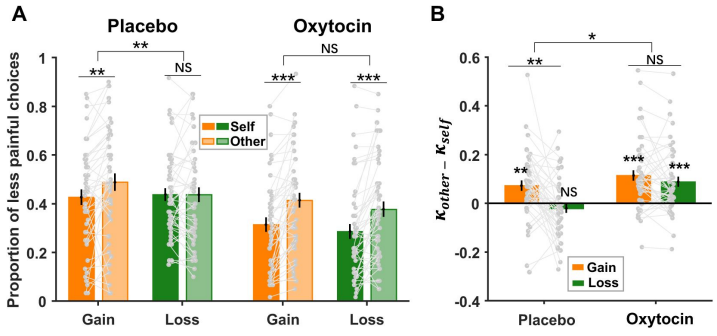



**Fig. 4.**

**Oxytocin significantly promoted hyperaltruistic preference in the loss context.**

(**A**) In contrast to the placebo condition, oxytocin administration elicited hyperaltruistic behavior (a greater tendency of choosing less painful options in other condition than self-condition) in the loss context compared to the gain context. (**B**) Model-based hyperaltruistic parameter ($\kappa_{other} - \kappa_{self}$) showed similar patterns: Hyperaltruistic tendency was reduced by the loss context in the placebo session but restored in the oxytocin session. Error bars represent SE across subjects. NS, not significant; * $P < 0.05$, ** $P < 0.01$ and *** $P < 0.001$.

variables including treatment (placebo vs. oxytocin), harm/shock recipient (self vs. other) and decision context valence (gain vs. loss; see methods for details). This regression analysis yielded a significant Δs × treatment × recipient × valence interaction effect ($\beta$ = 0.125 ± 0.053, $P$ = 0.018, 95% CI = [0.022, 0.228]; Supplementary Fig.6 & 8), indicating the relative harm sensitivities (*other* $\beta_{\Delta}$, − *self* $\beta_{\Delta}$,) was modulated by the specific treatment and decision combinations. Indeed, repeated ANONA analysis confirmed that in the placebo session, as in study 1, the relative harm sensitivities decreased in the loss context (relative to the gain context, simple effect: $F_{1,45}$ =11.521, $P$ = 0.001, $\eta^2$ = 0.204; **Fig. 5A**); however, with the administration of oxytocin, there was no significant difference of the relative harm sensitivities between the gain and loss contexts (simple effect: $F_{1,45}$ = 0.131, $P$ = 0.719, $\eta^2$ = 0.001; **Fig. 5A**). It is worth noting that oxytocin did not alter the gain/loss relative money sensitivity difference (valence × treatment interaction: $F_{1,45}$ = 1.933, $P$ = 0.171, $\eta^2$ = 0.041; **Fig. 5B**), which corresponded to a non-significant Δm × treatment × recipient × valence interaction effect ($\beta$ = -0.052 ± 0.100, $P$ = 0.601, 95% CI = [-0.249, 0.144]; Supplementary Fig. 6). Furthermore, in the placebo session, the decision context significantly modulated the relationship between the relative harm sensitivity and IH (gain context: $\beta$ = -0.071 ± 0.024, $t_{84}$ = -3.005, $P$ = 0.004; loss context: $\beta$ = 0.006 ± 0.024, $t_{84}$ = -0.243, $P$ = 0.809; valence × IH: $\beta$ = 0.077 ± 0.033, $t_{84}$ =2.297, $P$ = 0.024; **Fig. 5C**). However, under the treatment of oxytocin, the valence × IH interaction became non-significant ($\beta$ = -0.006 ± 0.021, $t_{84}$ =-0.272, $P$ = 0.786; **Fig. 5D**), despite the significant negative relationship between the relative harm sensitivities and IH in both decision contexts (gain context: $\beta$ = -0.037 ± 0.015, $t_{84}$ = -2.458, $P$ = 0.016; loss context: $\beta$ = -0.042 ± 0.015, $t_{84}$ = -2.843, $P$ = 0.006; **Fig. 5D**) (also see Supplementary Fig. 7 and Table S2). The effect size (Cohen's $f^2$) for this exploratory analysis was calculated to be 0.491 and 0.379 for the placebo and oxytocin conditions, respectively. The post hoc power analysis with a significance level of α = 0.05, 7 regressors (IH, IB, EC, decision context, IH×context, IB×context, and EC×context), and sample size of N = 46 yielded achieved power of 0.910 (placebo treatment) and 0.808 (oxytocin treatment). These findings further highlighted the influence of decision context on hyperaltruistic moral preference and stressed the importance of oxytocin in restoring the correlation between the relative harm sensitivity and the dispositional personality traits such as IH.

## Oxytocin eliminated the modulation effect of decision context on hyperaltruism

We reasoned that the decision context effect of hyperaltruistic preferences might be rooted in how our subjects internally framed the task as "harming" (inflicting pain on the receiver to increase monetary gain/avoid monetary loss) or "helping" others (sacrificing greater monetary gain/accepting greater monetary loss to alleviate the receiver's pain). To test this hypothesis, in study 2 we asked subjects to subjectively report whether they perceived the task as "harmful" or "helpful" (see Methods section for more details) in both placebo and oxytocin sessions (gain and loss contexts for each session). The non-parametric Friedman tests on subjects' harm framing report showed a significant difference in gain and loss contexts under placebo condition ($\xi^2$ = 2.827, $P$ = 0.028, Bonferroni correction; **Fig. 6A**), suggesting that subjects were more likely to perceive the task as harming others in the gain context. In addition, the administration of oxytocin increased subjects' perception of causing harm to others in the loss context ($\xi^2$ = 2.665, $P$ = 0.046, Bonferroni correction) and removed contextual disparities in harm perception (valence × treatment interaction: $\xi^2$ = 7.410, $P$ = 0.006) (**Fig. 6A**), suggesting that the oxytocin administration successfully erased the framing difference between the gain and loss decision contexts.

We further hypothesized that the correlations between instrumental harm (IH) personality traits and subjects' differential harm sensitivities (*other* $\beta_{,}$ − *self* $\beta_{,}$, **Fig. 5C** & D) might be mediated by how strongly subjects perceived the decision task as harming or helping others. A moderated mediation analysis was thus conducted to directly test this hypothesis. In the moderated mediation model, the effects of decision context can be expressed as the moderation effects on mediation

**Fig. 5.**

### Oxytocin effect on relative harm sensitivities.

(**A**) Oxytocin significantly modulated the context-specificity of relative harm sensitivity ($other\beta_{\Delta s} - self\beta_{\Delta s}$). (**B**) Oxytocin did not influence the contextual differences of the relative monetary sensitivities ($other\beta_{\Delta m} - self\beta_{\Delta m}$). (**C-D**) The decision context modulated the correlation between instrumental harm (IH) and subjects' harm sensitivities in the placebo session (**C**), whereas oxytocin eliminated the contextual modulation (**D**). Error bars represent SE across subjects. The regression coefficients in Figure C & D showed the association between IH and relative harm sensitivities, controlling for empathic concern and impartial beneficence. NS, not significant; * $P < 0.05$, ** $P < 0.01$.

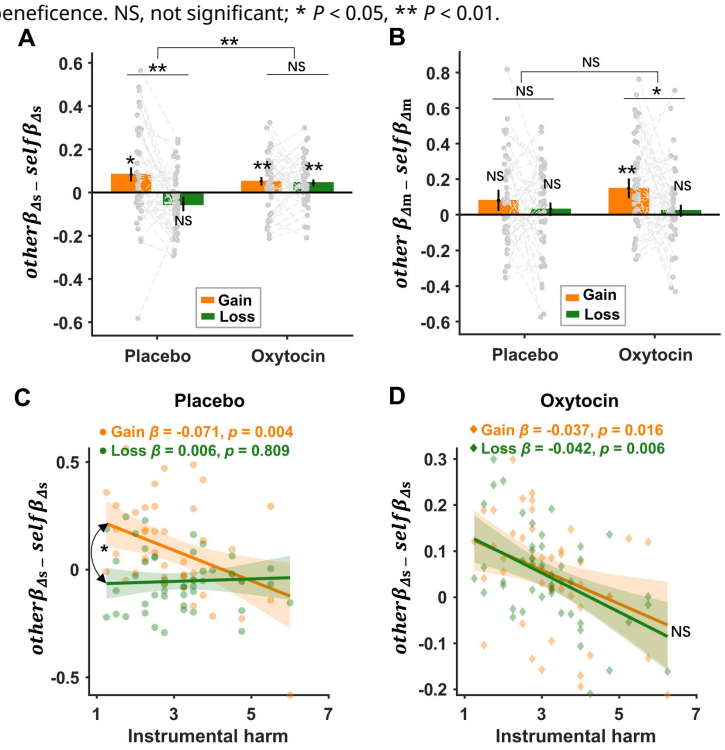
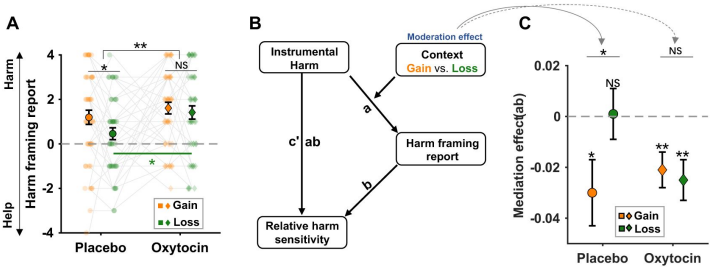



**Fig. 6.**

### Oxytocin modulated the contextual influence on hyperaltruistic behaviors.

(**A**) Monetary loss (relative to gain) significantly reduced subjects' perception of harm framing in the task. Oxytocin augmented harm framing perception, particularly in the loss context, effectively removing the contextual specificity of harm framing perception. (**B**) The conceptual diagram of the moderated mediation model. We assume that the perceived harm framing mediates the relationship between instrumental harm and relative harm sensitivity, with decision context moderating the mediation effect. (**C**) The moderating effect of decision context was significant under placebo condition. However, oxytocin obliterated the contextual moderation effect by reinstating the mediating role of harm perception in the loss context. Error bars represent SE across subjects. NS, not significant; * $P < 0.05$, ** $P < 0.01$.



(**Fig. 6B**). As we expected, in the placebo session, the moderating effect of decision context was significant (placebo: the differential mediation effect in gain and loss contexts: Δab = -0.036, *P* = 0.036, 95%CI = [-0.073, -0.003] ; **Fig. 6C**). Specifically, the mediation effect of perceived harm was significant in the gain context (ab = -0.030, *P* = 0.021, 95%CI = [-0.056, -0.006]) but not in the loss context (ab = 0.001, *P* = 0.915, 95%CI = [-0.019,0.019]). However, oxytocin eliminated the contextual moderation effect (oxytocin: the differential mediation effect in gain and loss contexts: Δab = -0.006, *P* = 0.731, 95%CI = [-0.044, 0.033]; **Fig. 6C**) by reinstating the mediation role of harm perception in the loss context (ab = -0.025, *P* = 0.002, 95%CI = [-0.042, -0.011]) (also see Table S3). These results indicated that subjects' perception of the task structure as helping or harming others mediated the correlation between personality trait IH and subjects' hyperaltruistic preference. Switching the decision context from gains to losses directly modulated subjects' moral perception of the task and nullified hyperaltruism. Finally, oxytocin administration restored participants' hyperaltruistic preference by reinstating the mediation role of perceived harm report on the correlation between IH and hyperaltruism across gain and loss decision contexts.

## Discussion

Utilitarian moral decision-making studies often involve moral dilemmas where utilitarian gains are traded against moral norms such as no deception or no infliction of suffering on others[1,16,66]. Recent studies extended this line of research by comparing subjects' sensitivities toward others' suffering relative to their own suffering and revealed an intriguing hyperaltruistic preference: people were willing to forgo more monetary gain to reduce others' pain than their own pain[4–7]. In two experiments, we replicated findings in the previous literature (Figs.1, 4). More importantly, we further showed that hyperaltruistic preference was susceptible to the corresponding decision context. Replacing the trade-off between monetary gain and electric shocks (gain context) with arbitration between monetary loss and electric shocks (loss context) eliminated subjects' hyperaltruistic preference that would otherwise have been observed. Importantly, oxytocin restored the hyperaltruistic preference in the loss context by biasing subjects to more likely perceive the decision context as harming others for self-interest. Finally, moral framing, or how subjects perceive the decision context as helping or harming others, mediated the association between the instrumental harm (IH) personality trait and subjects' sensitivities to others' suffering relative to their own suffering. Therefore, we demonstrated that both the hyperaltruistic preference and the effects of oxytocin were context-dependent and provided a mechanistic account and the boundary condition for hyperaltruistic disposition.

Recent development in utilitarian moral psychology highlights two independent dimensions that collectively shape people's moral preference: attitudes toward instrumental harm (IH), the suffering of other individuals to achieve greater good, and toward impartial beneficence (IB), an impartial concern for the well-being of everyone[1]. In our experiments, we found that subjects' IH attitude was negatively associated with their hyperaltruistic preference (**Fig. 3A**). However, there was no significant relationship between IB attitudes and hyperaltruism (**Fig. 3B**), suggesting the money- pain trade-off task might be better suited to study the specific relationship between IH attitude and moral preference[16]. Further analyses decomposing subjects' hyperaltruistic preference into their relative sensitivities towards shock difference (Δs) and money difference (Δm) suggested that the decision context (loss vs. gain) altered the relationship between IH and Δs relative sensitivity (**Figs. 3C**, **5C**) but not the association between IH and Δm sensitivity (**Fig. 3D**). These results confirm that how subjects evaluate others' suffering relative to their own suffering underlies the relationship between IH and hyperaltruistic preferences.

It might be argued that the decision context effect on hyperaltruistic preference was due to loss aversion, a phenomenon that people weigh prospective loss more prominently than monetary gain[23]. Our results, however, showed that the relative evaluation of Δm did not differ significantly across decision contexts (gain vs. loss, **Figs. 2D** & 5B). Furthermore, there was no

significant interaction effect of IH and decision contexts on the relative evaluation of Δm (**Fig. 3D**, supplementary Fig. 7A), excluding the role of loss aversion in mediating the context effect on hyperaltruistic preference. From a theoretical point of view, loss aversion only proportionally shifts harm aversion parameter κ (representing the relative importance of Δm and Δs) in the self- and other- conditions, respectively, and will not change the direction of hyperaltruistic preference ($\kappa_{other}$ - $\kappa_{self}$).

Instead, we propose that the moral framing of the decision context as helping or harming others may drive subjects' hyperaltruistic preference. Indeed, we showed that subjects were less likely to perceive the task as harming others for monetary benefit when the decision context switched from gain to loss condition (**Fig. 6A**). In addition, subjects' harm framing reports fully mediated the correlation between IH and the relative harm sensitivity in the gain conditions (**Figs. 5C** & 6C). Decision context exerted its effect by modulating the mediation effect of harm framing report such that the loss context eliminated the correlation between the relative harm sensitivity and IH (**Fig. 5C**). Our results are consistent with a recent study where the exogenous moral framing of the task (harming or helping others) significantly biased subjects' prosocial behavior[25,39]. Subjects behaved more prosocially under the harming frame compared to the helping frame. In our experiments, the decision contexts prompted different endogenous moral framing of the task such that more harm framing led to higher hyperaltruism levels. These results taken together suggest that moral framing may be a critical factor influencing subjects' prosocial preference. Another recent study reported that human subjects can be both egoistic and hyperaltruistic in moral decisions[7]. While the hyperaltruistic preference was elicited by comparing how subjects evaluated others' suffering to their own suffering in the standard monetary gain-pain trade-off task, egoistic tendencies emerged when subjects had to decide whether to harm themselves for others' benefit (compared to harming themselves for their own benefit). These ostensibly puzzling results can be reconciled under the framework of internal moral framing of the task. In order for participants to express moral or prosocial choices, they have to identify how salient a moral code is in place. Internally framing the task as benefiting from others' suffering would be more likely to discourage people from actions they would have otherwise taken if the task is perceived as sacrificing self interest in order to help others. The personality trait instrument harm (IH) attitudes tap into subjects' relative harm sensitivities and the correlation between IH and relative harm sensitivity is modulated by the decision context.

In our pre-registered study 2, we directly tested the potential mediating role of moral framing and found that the administration of oxytocin significantly increased subjects' moral framing of the decision context as harming others in both the gain and loss contexts (**Fig. 6A**). It is worth noting that oxytocin administration did not affect subjects' IH disposition (Supplementary Fig. 4B). Therefore, oxytocin effectively nullify the modulation effects of decision context and preserved the correlation between the relative harm sensitivity and IH in both decision contexts (**Figs. 5D**, **6C**). Oxytocin has long been featured in the general approach-avoidance hypothesis[67], which proposes that oxytocin attenuates people's avoidance of negative social or nonsocial stimuli[49,67–69]. Other studies conducted on both animals and humans suggest that oxytocin increases pain tolerance and attenuates acute pain experience [70–72]. However, hyperaltruistic preference is tied to the differential evaluation of other's suffering relative to subjects' own suffering and thus a general approach-avoidance theory of oxytocin might not be capable to account for the decision context specific effect. Recently, oxytocin has been shown to play a significant role in regulating parochial altruism by promoting both in- group cooperation and out-group aggression [58–60]. Our results corroborated with this line of research and highlighted the role of oxytocin in modulating the moral perception of the decision environment. Importantly, we demonstrated that moral perception of the task structure mediated the correlation between subjects' personality attitudes towards harming others (IH) and their relative harm sensitivities which directly contributed to the emergence of hyperaltruistic preference. Through this lens,

decision-context can be viewed as the implicit proxy of the mediator (task moral perception). As oxytocin increases the task perception as more harming others for self-interest (**Fig. 6A**), the correlation between IH and the relative harm sensitivities is restored.

In summary, in two studies, we demonstrate subjects' hyperaltruistic preference is closely associated with the decision context. Subjects' internal moral framing of the decision context mediates the correlation between personality traits such as instrumental harm (IH) attitude and the relative harm sensitivity and the decision context exerts a modulation effect on the mediation effect. Intranasal oxytocin administration drives subjects to be more harm frame oriented and abolishes the modulation effects of the decision contexts. Therefore, our study provides valuable insights into the psychological and cognitive mechanisms underlying hyperaltruistic behavior and highlights the crucial role of oxytocin in shaping moral decision-making. Finally, our findings carry significant implications for future research on moral behavior and its impairment in specific crisis contexts.

# Materials and Methods

## Participants

All subjects were students recruited through online university platform with written informed consent. Subjects were free of historic or current neurological or psychological disorders and with corrected normal vision. This study was approved by the institutional review board of the school of psychological and cognitive sciences at Peking University.

We conducted a power analysis (G*Power 3.1) to determine the number of subjects sufficient to detect a reliable hyperaltruistic effect reported in the previous literatures[4,7,73]. Based on the small to medium effect size (Cohen's d = 0.2), 75 subjects were needed to detect a significant effect ($\alpha = 0.05$, $\beta = 0.9$, two-by-two within ANOVA) for study 1. We ended up recruiting 83 right-handed subjects for study 1. Two subjects were excluded from data analysis due to their exclusive selection of the same option (more or less painful) across all trials, and 1 subject did not complete the experiment, leaving a total of 80 subjects (40 males, mean age = 21.38 ± 2.67 years). Study 2 (the oxytocin study) was specifically designed to test the oxytocin effect on hyperaltruism and was preregistered (*https://osf.io/fhwa9*) via the Open Science Framework. Due to the potential confounds of the oxytocin effects, we only recruited male subjects for Study 2[74,75]. It should be noted that in preregistration we originally planned to recruit 60 male subjects for Study 2 but ended up recruiting 46 male subjects (mean age = 21.74 ± 2.33 years) based on the sample size reported in previous oxytocin studies[57,69]. Additionally, a power analysis suggested that the sample size > 44 should be enough to detect a small to median effect size of oxytocin (Cohen's d = 0.21, $\alpha = 0.05$, $\beta = 0.8$) using a 2 × 2 × 2 within-subject design[76]. All the subjects were instructed to avoid taking caffeine, cigarettes, and alcohol 24 hours before the experiment and to refrain from eating or drinking two hours before the experiment.

## Experimental Procedures

Each time, two subjects (a pair) arrived with a 5-minute interval and entered into separate testing rooms without seeing each other to ensure complete anonymity. After signing the informed consent, subjects completed the pain calibration procedure (described below) in separate rooms. We then followed the protocol developed in the previous literature to assign experimental roles to both subjects before the main task[4]. Briefly, each subject was told that there was a second subject in the next testing room (whom they would not meet in person). Subjects were instructed that a random coin toss would decide their roles as the decider or receiver in the upcoming money-pain tradeoff task. Unbeknownst to the subjects, both of them were assigned to the role of deciders.

For study 2, the experimental procedures were similar to those in study 1. In addition, both subjects were administrated with either oxytocin or placebo in two sessions (session 1 & 2) of 5~7 days apart in a randomized order, yielding 11 pairs (22 subjects) having the oxytocin session first and the other 12 pairs (24 subjects) taking the placebo session first. Also, once the subjects' task roles (deciders) were announced in session 1, their roles remained in session 2.

## Pain calibration

We adopted a standard pain titration paradigm that has been widely used in previous literature [4,77]. Electric shocks were generated with a Digitimer DS5 electric stimulator (Digitimer, UK) and applied to the inner side of the subject's left wrist via two electrodes. By slowly increasing or decreasing the electric shock intensities, we asked subjects to rate their pain experience on a 11-point scale ranging from 0 (no pain at all) to 10 (intolerable) until a rating of 10 was reached (subject's maximum tolerance threshold). Next, we generated shocks ranging from 30% to 90% of the subjects' maximum threshold in 10% increments. Each shock intensity was delivered three times. Subjects in total received 21 shocks in randomized order and gave pain ratings on the 11-point scale. For each subject, we fitted a sigmoid function to subjects' subjective pain ratings and chose the current intensity that corresponded to each subject's subjective rating of level 7 from the derived function (**Fig. 1B**). This individualized intensity was then used in the following Money-Pain tradeoff task.

## Oxytocin/placebo administration

The oxytocin and placebo administration procedure was similar to that used in the previous studies[55,57]. For each treatment, oxytocin or placebo (saline) spray was administered to subjects thrice, consisting of one inhalation of 4 international units (IU) into each nostril resulting a total of 24 IU. The order of oxytocin and placebo sessions was counterbalanced across subjects. After each session, subjects were asked to rate on a 5-point scale about their perception of the administered substance being oxytocin, with 1 indicating it was unlikely to be oxytocin, 5 indicating strong confidence that it was oxytocin. Subjects' rating indicated that they had no bias towards the substance used in each session (placebo session: mean = 3.13 ± 0.89; oxytocin session: mean = 3.11 ± 1.16 ; the difference between the two sessions: $t45 = 0.0965$, $P = 0.924$). No subjects reported any side effect after the experiments.

## Experimental design

Study 1 adopted a 2 (decision context: gain vs. loss) × 2(shock recipient: self vs. other) within-subject design (**Fig.1A**). Each subject had to choose between two options representing the tradeoff between money and shock. In the gain context, subjects decided whether to inflict more pain (more electric shocks) either on themselves (gain- self condition) or on the receivers (gain-other condition) to gain more money. In the loss context, subjects would decide whether to inflict more pain either on themselves (loss-self condition) or on the receivers (loss-other condition) to avoid a bigger monetary loss. The sequence of the gain and loss contexts were block designed and counterbalanced across subjects. In study 2, we employed a 2 (decision context: gain vs. loss) × 2 (shock recipient: self vs. other) × 2 (treatment: placebo vs. oxytocin) within-subject design to examine oxytocin's effect on subjects' hyperaltruistic preference. Each subject performed the same money-pain tradeoff tasks (same as study 1) twice. Approximately 35 minutes before each experiment and immediately after the pain calibration procedure, each subject was intranasally administered 24 IU oxytocin or placebo (saline) (**Fig.1B**).

In the Money-Pain tradeoff task, subjects were asked to choose between options associated with different levels of electric shocks and monetary amounts. Crucially, more painful options were always associated with larger monetary gain (or smaller monetary loss), while less painful options offered smaller monetary gains (or larger monetary losses). The experimental task was coded with Psychtoolbox (version 3.0.17) in Matlab (2018b). We first created 60 gain trials using the procedure

reported in earlier research[4]. Specifically, each trial was determined by a combination of the differences of shocks (Δs, ranging from 1 to 19, with increment of 1) and money (Δm, ranging from ¥0.2 to ¥19.8, with increment of ¥0.2) between the two options, resulting in a total of 19×99=1881 pairs of [Δs, Δm]. To ensure the trials were suitable for most subjects, we evenly distributed the desired ratio Δm / (Δs + Δm) between 0.01 and 0.99 across 60 trials for each condition. For each trial, we selected the closest [Δs, Δm] pair from the [Δs, Δm] pool to the specific Δm / (Δs + Δm) ratio, which was then used to determine the actual money and shock amounts of two options. The shock amount ($S_{less}$) for the less painful option was an integer drawn from the discrete uniform distribution [1∼19], constraint by $S_{less}$ + Δs < 20. Similarly, the money amount ($M_{less}$) for the less painful option was drawn from a discrete uniform distribution [¥0.2 ∼ ¥19.8], with the constraint of $M_{less}$ + Δm < 20. Once the $S_{less}$ and $M_{less}$ were selected, the shock ($S_{more}$) and money ($M_{more}$) magnitudes for the more painful option were calculated as: $S_{more} = S_{less} + \Delta s$, $M_{more} = M_{less} + \Delta m$. For the loss condition, we flipped the signs for the monetary amount and then switched the monetary amounts between the more and less painful options. For example, the two options [¥15 & 10 shocks; ¥10 & 5 shocks] in the gain condition would turn to [- ¥10 & 10 shocks; - ¥15 & 5 shocks] in the loss condition. Subjects completed 60 trials each for all the 4 decision-context and shock-recipient combinations (gain-self, gain-other, loss-self and loss other) thus yielding a total of 240 trials delivered across two blocks (gain and loss blocks). We randomly generated each subject's trial set using the above method. Across trials, Δs and Δ m was uncorrelated ($r$ = -0.0012, $P$ = 0.986). The trial sequences and the more/less painful option location (left vs. right) on the computer screen was randomized across subjects within each block.

For each trial, subjects had a maximum of 6s to choose either the more or less painful option by pressing a keyboard button with their right or left index finger. Button presses resulted in the chosen option being highlighted on the screen for 1s. If no choice was entered during the 6s response window, a message 'Please respond faster!' was displayed for 1s, and this trial was repeated. An inter-trial interval (ITI) of 1s was introduced before the beginning of the next trial (**Fig.1B**). Each subject was endowed with ¥40 at the beginning of the task (the maximum amount subjects could have lost in each loss-self and loss-other trials together). At the end of the experiment, one trial for each experimental condition was randomly selected and subjects were renumerated and shocked according to the combined payoff (four conditions) and electric shocks (only in gain-self and loss-self conditions). Each subjects received a ¥60 show up fee in study1(¥220 in study2) and the average total subject fee is ¥97(¥318 in study2).

## Personality trait measures and questionnaires

The Oxford utilitarianism scale was utilized to evaluate two independent dimensions that drive people's prosocial preference in a moral dilemma[1]: the dimension of instrumental harm (IH) measures individuals' willingness to compromise moral principles by either breaking rules or causing harm to others to achieve favorable outcomes; whereas the impartial beneficence (IB) captures their levels of impartial concern for the well-being of the general population. IH has four items including statements such as "Sometimes it is morally necessary for innocent people to die as collateral damage if more people are saved overall". The IB measure contains 5 items such as "It is morally wrong to keep money that one doesn't really need if one can donate it to causes that provide effective help to those who will benefit a great deal". Subjects responded to each IH and IB item via a 7-point Likert scale. IH and IB scores were derived by calculating the average score of all items on the subscale for each subject. We also measured subjects' empathetic concern (EC), the identification with the whole of humanity and concern for future generation, using the interpersonal reactivity index (IRI) [7,8]. The EC comprises 7 items with a 5-point rating scale and has been shown to be correlated with both IH and IB (see Supplementary Fig. 2). Therefore, we performed multiple regression analysis to examine the relationships between IB, IH and moral behaviors, with EC included as a covariate of no interest.

All subjects completed debriefing questionnaires that assessed their beliefs about the experimental setup, such as whether they believed that the recipient would actually receive the electric shocks. Moreover, for study 2, we also collected another questionnaire by asking subjects how they perceived the task structure (whether they regarded the experiment as helping or harming other subjects) in both decision contexts (gain and loss). Subjects rated their moral perception within gain and loss contexts using a scale ranging from -4 to 4. Positive values indicated harm frame (inflicting pain on the receiver to increase monetary gain/avoid monetary loss), whereas negative values indicated help frame (sacrificing greater monetary gain/accepting greater monetary loss to alleviate others' pain) with 0 indicating neutrality. Therefore, a larger rating indicated subject's moral perception of a more harming frame.

## Data Analysis

### Harm aversion model

We analyzed subjects' choice data using the harm aversion model[6], where choices were driven by the subjective value difference (ΔV) between the less and more painful options (Eq. 1). Parameter κ ($0 \le \kappa \le 1$) quantifies the relative weight deciders attribute to electric shock versus money.

$$\Delta V = (1 - \kappa)\Delta m - \kappa \Delta s \quad \text{(Eq. 1)}$$

We separately estimated κ in the four conditions related to both the shock recipients (self vs. other) and decision contexts (gain vs. loss) and thus yielded $\kappa_{gain\text{-}self}$, $\kappa_{gain\text{-}other}$, $\kappa_{loss\text{-}self}$, and $\kappa_{loss\text{-}other}$ respectively. Δm and Δs denoted the objective differences in money and electric shocks between the more and less painful options. In this model, trial-wise ΔV was transformed into choice probability via a softmax function where γ serves as a subject-specific parameter for choice consistency (Eq. 2) (see Supplementary Fig. 10 for the model comparison results that confirmed a single γ for all conditions fitted subjects' behavior better than other candidate models).

$$P(\text{less painful option}) = \frac{1}{1+e^{-\gamma \Delta V}} \quad \text{(Eq. 2)}$$

Model parameters were estimated for each subject using maximum likelihood estimation (MLE), and the maximization was achieved using the fminsearchbnd function in MATLAB (Mathworks). We repeated the estimation with 300 random initial parameter values to achieve stable parameter estimators.

### Regression models

We applied the mixed-effect logistic regression model to investigate how Δm and Δs independently influenced subjects' choices. Although this regression model seems equivalent to the harm aversion model mentioned above, the regression coefficients obtained from the regression model allowed us to separately examine how the decision contexts and oxytocin treatment affected subjects' sensitivities towards Δm and Δs. In the regression model 1 of Study 1, choices were coded as 1 if subjects chose the less painful option and 0 otherwise. Additionally, decision context (1 for gain and 0 for loss) and shock recipient (1 for self and 0 for other) were treated as categorical variables. Each regressor had a fixed effect across all subjects and a random effect specific to each subject. The regression model 1 was specified as following[79]:

$$\begin{aligned} choice \sim{}& \Delta s*context*recipient + \Delta m*context*recipient + \\ & (1 + \Delta s*context*recipient + \Delta m*context*recipient | subject) \end{aligned}$$

To examine how oxytocin affected subjects' behavior, in Study 2, we expanded regression model 1 by including the variable of treatment (1 for oxytocin treatment and 0 for placebo) as an additional categorical independent variable. The regression model 2 was described as:

$$choice \sim \Delta s*context*recipient*treatment+\Delta m*context*recipient*treatment+ (1+\Delta s*context*recipient*treatment+\Delta m*context*recipient*treatment \mid subject)$$

### Moderated mediation analysis

We set out to examine whether subjects' internal moral framing mediates the association between their personality traits (IH) and moral preferences, and more importantly whether the mediation is context (gain vs. loss) specific. To test the moderated mediation, we constructed a mediation model where the mediator (M):

$$M \sim \beta_0^M+\beta_1^M IH+\beta_2^M DC+\beta_3^M IH\times DC+\beta_4^M IB+\beta_5^M EC \quad \text{(model M)}$$

and the dependent variable model was:

$$Y \sim \beta_0^Y+\beta_1^Y IH+\beta_2^Y M+\beta_3^Y DC+\beta_4^Y IH\times DC+\beta_5^Y M\times DC+\beta_6^Y IB+\beta_7^Y EC \quad \text{(model Y)}$$

Where Y, M respectively referred to the dependent variable (relative harm sensitivity) and the mediator (harm framing report). Variables IH, DC, IB and EC corresponded to independent variable instrumental harm attitude (IH), decision context (gain vs. loss, the moderator variable), impartial beneficence (IB, variable of no interest) and empathic concern (EC, variable of no interest), accordingly. The existence of interaction effect ($\beta_3^M$,$\beta_4^Y$, $\beta_5^Y$) in either path a (X→M), path b (M→Y) or path c' (X→Y)in the model is sufficient to claim moderation of mediation[80].

We utilized the bootstrapping procedure with 5000 bootstrap resamples to analyzed the data[81,82]. This analysis was conducted separately for the placebo and oxytocin sessions. Our results showed that the interaction coefficients (placebo: $\beta_4^Y$= 0.036 ± 0.029, $t_{84}$ = 1.267, $P$ = 0.209; $\beta_5^Y$ = 0.0242 ± 0.017, $t_{84}$ = 1.415, $P$ = 0.161; Oxytocin: $\beta_4^Y$ = 0.006 ± 0.010, $t_{84}$ = 0.319, $P$ = 0.751 ; $\beta_5^Y$= 0.003 ± 0.012, $t_{84}$=0.246, $P$ = 0.806) in model Y were not significant, while $\beta^M$ was significant in the placebo session ($\beta^M$= 0.784 ± 0.343, $t$86 = 2.288, $P$ = 0.025) but not the oxytocin session ($\beta^M$= -0.137 ± 0.291, $t$86 = -0.472, $P$ = 0.638, indicating that the moderation effect present only on the path a (see **Fig. 6B**) in the placebo condition. This aligns with our findings that decision contexts modulated the correlation between IH and relative harm sensitivities (**Fig. 3C** & **Fig. 5C**). Therefore, we report all the results from the reduced version of the moderated mediation model (model Y[5], see supplementary Table S3 for details).

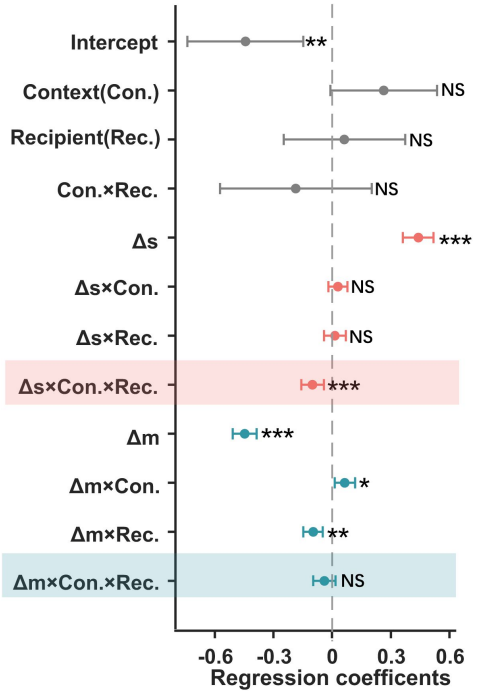
$$Y' \sim \beta_0^{Y'}+\beta_1^{Y'} IH+\beta_2^{Y'} M+\beta_3^{Y'} DC+\beta_6^{Y'} IB+\beta_7^{Y'} EC \quad \text{(model Y')}$$

### Statistics and software

ANOVAs and non-parametric analyses were conducted with SPSS 27.0, whereas regression analyses were performed using "fitglme" and "fitlm" functions in Matlab (2022b). The moderated mediation analysis was performed using "bruceR" and "mediation" packages in R (version 4.2.2). All the reported p-values are two-tailed.

## Data availability

The data and analysis code that support the findings of this study are available at *https://osf.io/tpfeg/*.




## Supplementary Fig. 1.

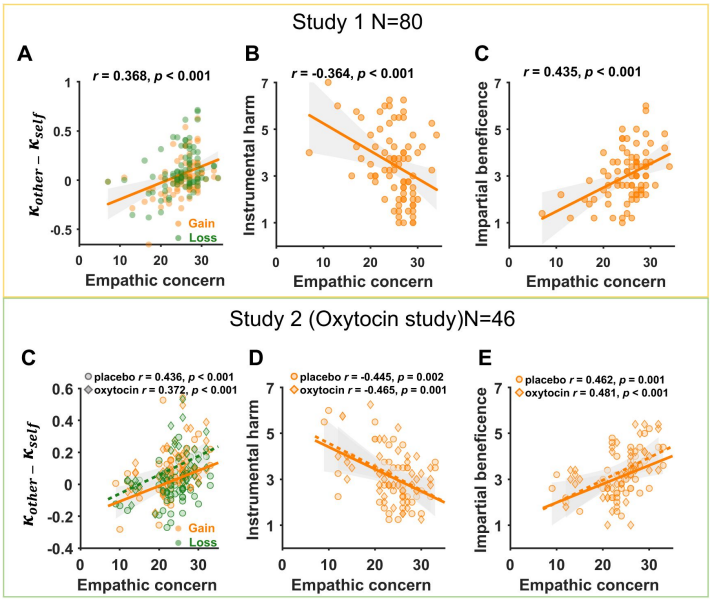
### Mixed-effect logistic regression analysis results for experiment 1(regression model 1).

Δm and Δs represented the objective differences in money and electric shocks between the more and less painful options. Δ*s* and Δ*m* are numerical variables, while the context (gain vs. loss) and recipient (other vs. self) are categorial variables. The Δs × context × recipient interaction and Δm × context × recipient interaction was further elucidated in **Fig. 2C and D**. Con: context; Rec: recipient. Error bars represent 95% confidence interval (CI). NS, not significant, * $P < 0.05$, ** $P < 0.01$ and *** $P < 0.001$.

## Supplementary Fig. 2.

### The relationship between personality trait empathic concern (EC) and utilitarian moral personality traits.

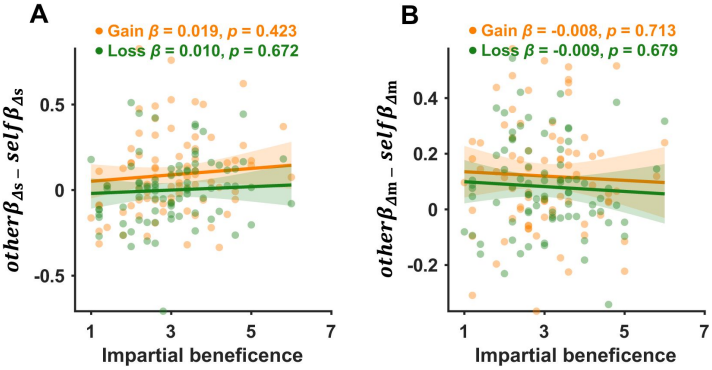
Correlation between EC and hyperaltruism ($\kappa_{other} - \kappa_{self}$), instrumental harm (IH), impartial beneficence (IB) both in the Study1(**A-C**) and Study 2 (**D-E**). All the correlations were performed using Pearson correlations.




## Supplementary Fig. 3.

### Association between impartial beneficence (IB) and subjects' relative harm/money sensitivities in Study 1.

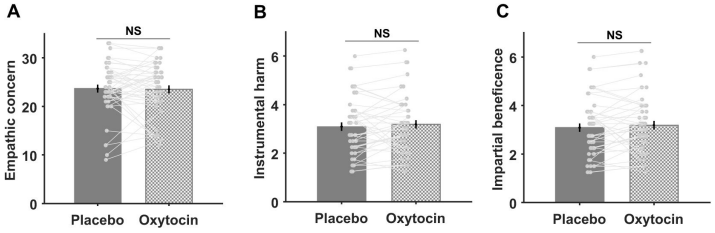
IB showed no significant correlation with subjects' relative harm sensitivities ($other\beta_{\Delta s} - self\beta_{\Delta s}$) **(A)** or relative monetary sensitivities ($other\beta_{\Delta m} - self\beta_{\Delta m}$) **(B)**. The regression coefficients showed the association between IB and relative harm/monetary sensitivities, controlling for empathic concern and instrumental harm.



## Supplementary Fig. 4.

### Oxytocin did not influence personality traits in study 2.

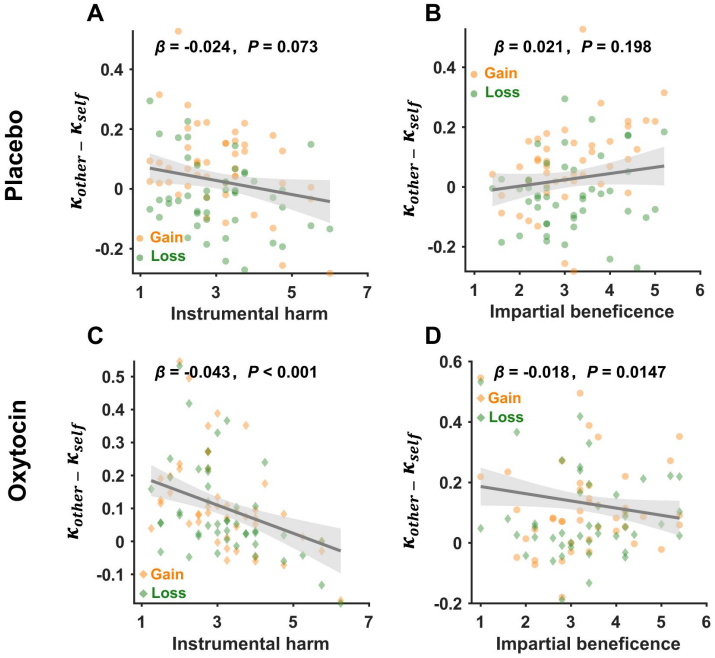
(**A-C**) Oxytocin did not affect subjects' ratings on empathic concern (EC), instrumental harm (IH) or impartial beneficence (IB). Error bars represent SE. NS, not significant.

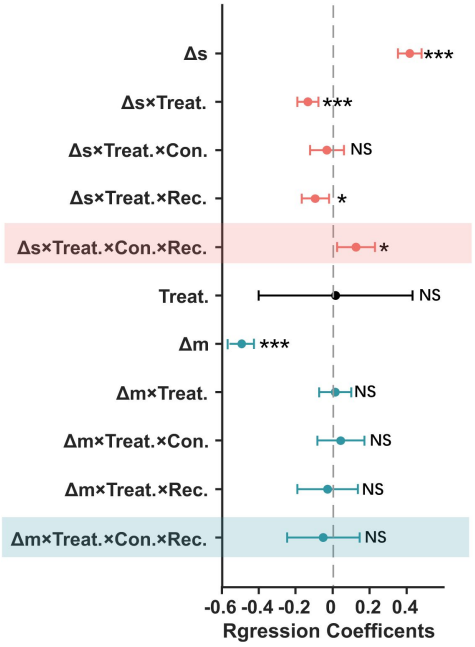



## Supplementary Fig. 5.

### Relationships between hyperaltruism and utilitarian moral personality traits in Study 2.

(**A-B**) In the placebo session, hyperaltruistic preferences showed a marginally negative correlation with instrumental harm (IH) but no significant correlation with impartial beneficence (IB). (**C-D**) In the oxytocin session, hyperaltruistic preferences exhibited significantly negative correlation with IH yet still no significant correlation with IB. The regression coefficients in Figure A, C showed the relationship between IH and moral preference, controlling for empathic concern (EC) and IB, while the coefficient in Figure B, D showed the relationship between IB and moral preference, controlling for EC and IH.






## Supplementary Fig. 6.

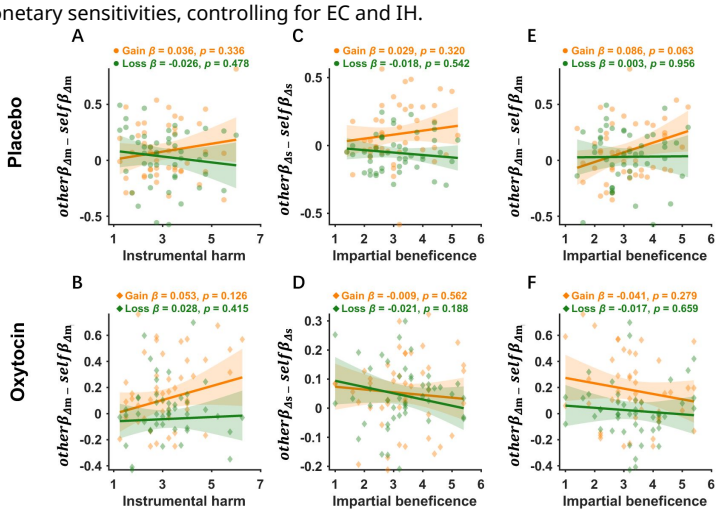
### The main results of mixed-effect logistic regression analysis in study 2(regression model 2).

Δm and Δs represented the objective differences in money and electric shocks between the more and less painful options. The Δs × Treat.× Con.× Rec. interaction was further explained in **Fig. 5A** and Supplementary Fig.8. The Δm × Treat.× Con.× Rec. interaction was further explained in **Fig. 5B**. Treat.: treatment (placebo vs. oxytocin); Con.: decision contexts (gain vs. loss), and Rec.: recipient (self vs. other) are all categorial variables. Error bars represent 95% confidence interval (CI). NS, not significant, * $P < 0.05$, and *** $P < 0.001$.

## Supplementary Fig. 7.

### Associations between relative harm/money sensitivities and utilitarian moral personality traits in Study 2.

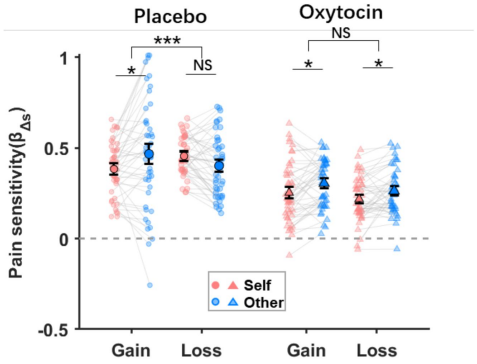
(**A-B**) The relative monetary sensitivities were not significantly related with instrumental harm (IH) in either placebo or oxytocin condition. (**C-D**) In both the placebo and oxytocin conditions, there was no significant correlation between relative harm sensitivity and impartial beneficence (IB). (**E-F**) No significant correlation was found between relative monetary sensitivities and IB in either the placebo or oxytocin conditions. Figure A & B illustrated the relationship between IH and relative monetary sensitivities, with empathic concern (EC) and IB controlled. Figure C ~ F described the relationship between IB and relative harm/monetary sensitivities, controlling for EC and IH.




## Supplementary Fig. 8.

### Pain sensitivity across experimental conditions.

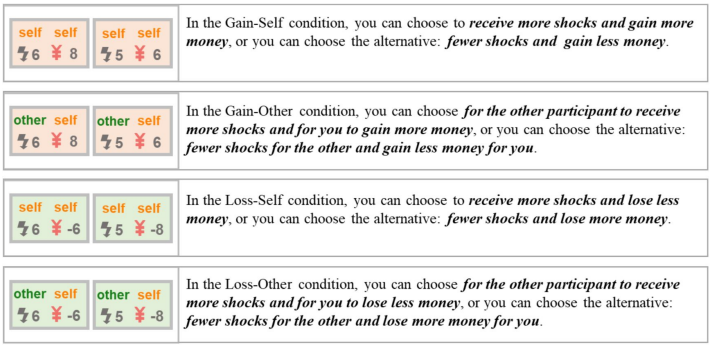
The administration of oxytocin significantly reduced participants' pain sensitivity, yet also restored the harm sensitivity patterns in both the gain and loss conditions relative to the placebo treatment. Error bars represent SE across subjects. NS, not significant, * $P < 0.05$, and *** $P < 0.001$.



## Supplementary Fig. 9.

### The experimental instructions provided to subjects.

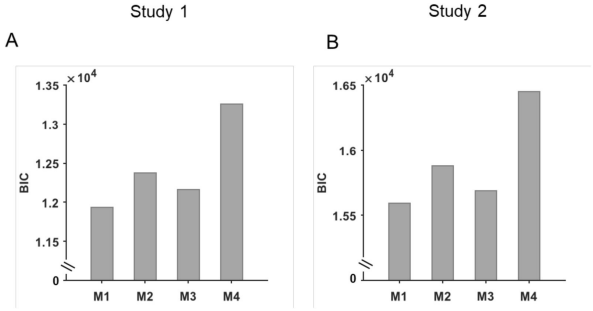
We presented the options to subjects in a neutral and descriptive manner, focusing only on the relevant components (shocks and money) across four conditions.




## Supplementary Fig. 10.

### Model comparison results.

To investigate whether the choice consistency parameter (γ) is condition specific, we compared four candidate models: (M1) γ was constant across all conditions; (M2) γ varied based on the pain recipient (self vs. other); (M3) γ varied between decision contexts (gain vs. loss); and (M4) γ varied across all four conditions (self-gain, other-gain, self-loss, and other-loss). Model 1 (M1) performed the best among all the four candidate models both in Study 1 **(A)** and Study 2 **(B)**.




## Table S1.

### Contribution of EC, IH and IB scores on subjects' behavior in Study 1.

| Predictors | Hyperaltruism | | | | Relative harm sensitivity | | | | Relative monetary sensitivity | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | β | SE | t | p | β | SE | t | p | β | SE | t | p |
| Intercept | -0.086 | 0.153 | -0.563 | 0.574 | -0.086 | 0.162 | -0.532 | 0.596 | 0.121 | 0.147 | 0.825 | 0.411 |
| Empathic concern(EC) | 0.011 | 0.006 | 2.080 | 0.039 | 0.012 | 0.006 | 2.063 | 0.041 | 0.004 | 0.005 | 0.738 | 0.461 |
| Instrumental harm(IH) | -0.039 | 0.015 | -2.501 | 0.013 | -0.052 | 0.016 | -3.224 | 0.002 | -0.022 | 0.015 | -1.476 | 0.142 |
| Impartial beneficence(IB) | 0.009 | 0.022 | 0.421 | 0.675 | 0.019 | 0.023 | 0.804 | 0.423 | -0.008 | 0.021 | -0.369 | 0.713 |
| context | -0.163 | 0.217 | -0.754 | 0.452 | -0.298 | 0.229 | -1.302 | 0.195 | 0.058 | 0.208 | 0.280 | 0.780 |
| EC × Context | 0.002 | 0.008 | 0.216 | 0.829 | 0.003 | 0.008 | 0.383 | 0.702 | -0.004 | 0.007 | -0.484 | 0.629 |
| IH × Context | 0.016 | 0.022 | 0.726 | 0.469 | **0.045** | **0.023** | **1.959** | **0.052** | -0.001 | 0.021 | -0.034 | 0.973 |
| IB × Context | -0.004 | 0.031 | -0.115 | 0.908 | -0.009 | 0.033 | -0.268 | 0.789 | -0.001 | 0.030 | -0.032 | 0.974 |

Note. Hyperaltruism: $\kappa_{other} - \kappa_{self}$, Relative harm sensitivity: $other\beta_{\Delta s} - self\beta_{\Delta s}$, Relative monetary sensitivity: $other\beta_{\Delta m} - self\beta_{\Delta m}$. The decision context (1 for gain and 0 for loss) was coded as a categorical variable.

## Table S2.

### Contribution of EC, IH and IB scores on subjects' behavior in Study 2.

**a. Placebo condition**

| Predictors | Hyperaltruism | | | | Relative harm sensitivity | | | | Relative monetary sensitivity | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | β | SE | t | p | β | SE | t | p | β | SE | t | p |
| Intercept | -0.085 | 0.123 | -0.688 | 0.493 | 0.065 | 0.170 | 0.382 | 0.704 | -0.205 | 0.266 | -0.771 | 0.443 |
| Empathic concern(EC) | 0.005 | 0.004 | 1.162 | 0.248 | 0.006 | 0.006 | 1.086 | 0.281 | -0.004 | 0.009 | -0.444 | 0.658 |
| Instrumental harm(IH) | -0.032 | 0.017 | -1.904 | 0.060 | -0.071 | 0.024 | -3.005 | 0.003 | 0.035 | 0.037 | 0.967 | 0.336 |
| Impartial beneficence(IB) | 0.047 | 0.021 | 2.200 | 0.031 | 0.029 | 0.029 | 1.000 | 0.320 | 0.086 | 0.046 | 1.887 | 0.063 |
| context | -0.034 | 0.174 | -0.197 | 0.844 | -0.340 | 0.241 | -1.412 | 0.162 | 0.323 | 0.375 | 0.862 | 0.391 |
| EC × Context | 0.002 | 0.006 | 0.332 | 0.740 | 0.005 | 0.008 | 0.601 | 0.549 | 0.003 | 0.012 | 0.264 | 0.792 |
| IH × Context | 0.018 | 0.024 | 0.737 | 0.463 | **0.077** | **0.033** | **2.297** | **0.024** | -0.062 | 0.052 | -1.187 | 0.238 |
| IB × Context | -0.051 | 0.030 | -1.711 | 0.091 | -0.047 | 0.042 | -1.140 | 0.257 | -0.084 | 0.065 | -1.295 | 0.199 |

**b. Oxytocin condition**

| Predictors | Hyperaltruism | | | | Relative harm sensitivity | | | | Relative monetary sensitivity | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | β | SE | t | p | β | SE | t | p | β | SE | t | p |
| Intercept | 0.138 | 0.125 | 1.100 | 0.275 | 0.056 | 0.109 | 0.511 | 0.611 | 0.049 | 0.253 | 0.195 | 0.846 |
| Empathic concern(EC) | 0.010 | 0.004 | 2.397 | 0.019 | 0.006 | 0.004 | 1.675 | 0.098 | 0.001 | 0.009 | 0.157 | 0.876 |
| Instrumental harm(IH) | -0.047 | 0.017 | -2.721 | 0.008 | -0.037 | 0.015 | -2.458 | 0.016 | 0.053 | 0.035 | 1.541 | 0.126 |
| Impartial beneficence(IB) | -0.034 | 0.019 | -1.813 | 0.073 | -0.009 | 0.016 | -0.582 | 0.562 | -0.041 | 0.038 | -1.091 | 0.279 |
| Context(gain vs. loss) | -0.071 | 0.177 | -0.401 | 0.690 | 0.092 | 0.154 | 0.596 | 0.553 | -0.372 | 0.358 | -1.040 | 0.301 |
| EC × Context | 0.002 | 0.006 | 0.385 | 0.701 | -0.002 | 0.005 | -0.336 | 0.737 | 0.012 | 0.012 | 0.985 | 0.327 |
| IH × Context | 0.008 | 0.024 | 0.323 | 0.747 | **-0.006** | **0.021** | **-0.272** | **0.786** | -0.045 | 0.049 | -0.918 | 0.361 |
| IB × Context | -0.010 | 0.026 | -0.384 | 0.702 | -0.012 | 0.023 | -0.526 | 0.600 | 0.024 | 0.053 | 0.457 | 0.649 |

Note. Hyperaltruism: $\kappa_{other} - \kappa_{self}$, Relative harm sensitivity: $other\beta_{\Delta s} - self\beta_{\Delta s}$, Relative monetary sensitivity: $other\beta_{\Delta m} - self\beta_{\Delta m}$. The decision context (1 for gain and 0 for loss) was coded as a categorical variable.

## Table S3.

### Moderated mediation analysis results for Study 2

| Predictors | Placebo | | Oxytocin | |
|---|---|---|---|---|
| | Mediator | Dependent variable | Mediator | Dependent variable |
| | Harm framing report β(SE) | Relative harm sensitivity β(SE) | Harm framing report β(SE) | Relative harm sensitivity β(SE) |
| Intercept | 1.196(0.279)*** | 0.069(0.023)** | 1.609(0.239)*** | 0.049(0.014)*** |
| Context | -0.739(0.396) | -0108(0.032)** | -0.196(0.338) | -0.002(0.019) |
| Empathic concern(EC) | 0.020(0.046) | 0.008(0.004)* | 0.023(0.040) | 0.005(0.002)* |
| Impartial beneficence(IB) | -0.116(0.239) | 0.010(0.019) | -0.270(0.179) | -0.008(0.010) |
| Instrumental harm(IH) | **-0.758(0.257)*** a** | **-0.018(0.016) c'** | **-0.753(0.220)*** a** | **-0.017(0.011) c'** |
| Context × IH | **0.784(0.343)*** | — | **-0.137(0.291)** | — |
| Harm framing report | — | **0.039(0.008)*** b** | — | **0.028(0.006)*** b** |

Note: In the moderated mediation model, IH serves as the independent variable, harm framing report as the mediator, relative harm sensitivity as the dependent variable, with decision context (gain vs. loss) as the moderator. **a** represents the influence of the independent variable on the mediator, while **b** represents the effect of the mediator on the dependent variable, and **c'** represents the direct effect (insignificant indicating a full mediation effect). The coefficients highlighted in red indicate the moderation effect of the context (gain vs. loss). * $P < 0.05$, ** $P < 0.01$ and *** $P < 0.001$.

## Acknowledgements

This work was supported by the National Science and Technology Innovation 2030 Major Program (2021ZD0203702), National Natural Science Foundation of China Grants (32071090) to J.L.

## Additional information

### Funding

**National Science and Technology Innovation 2030 Major Program (2021ZD0203702)**

**National Natural Science Foundation of China (32071090)**

# References


1. Kahane G., Everett J. A. C., Earp B. D., Caviola L., Faber N. S., Crockett M. J., Savulescu J (2018) **Beyond sacrificial harm: A two-dimensional model of utilitarian psychology** *Psychological Review* **125**:131–164 Google Scholar

2. Bartels D. M., Pizarro D. A (2011) **The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas** *Cognition* **121**:154–161 Google Scholar

3. Glenn A. L., Koleva S., Iyer R., Graham J., Ditto P. H (2010) **Moral identity in psychopathy** *Judgment and Decision Making* **5**:497–505 Google Scholar

4. Crockett M. J., Kurth-Nelson Z., Siegel J. Z., Dayan P., Dolan R. J (2014) **Harm to others outweighs harm to self in moral decision making** *Proceedings of the National Academy of Sciences of the United States of America* **111**:17320–17325 Google Scholar

5. Crockett M. J., Siegel J. Z., Kurth-Nelson Z., Ousdal O. T., Story G., Frieband C., Grosse-Rueskamp J. M., Dayan P., Dolan R. J (2015) **Dissociable Effects of Serotonin and Dopamine on the Valuation of Harm in Moral Decision Making** *Current Biology* **25**:1852–1859 Google Scholar

6. Crockett M. J., Siegel J. Z., Kurth-Nelson Z., Dayan P., Dolan R. J (2017) **Moral transgressions corrupt neural representations of value** *Nature Neuroscience* **20**:879–885 Google Scholar

7. Volz L. J., Welborn B. L., Gobel M. S., Gazzaniga M. S., Grafton S. T (2017) **Harm to self outweighs benefit to others in moral decision making** *Proceedings of the National Academy of Sciences of the United States of America* **114**:7963–7968 Google Scholar

8. Bustan S., Gonzalez-Roldan A. M., Schommer C., Kamping S., Löffler M., Brunner M., Flor H., Anton F (2018) **Psychological, cognitive factors and contextual influences in pain and pain-related suffering as revealed by a combined qualitative and quantitative assessment approach** *PLoS One* **13**:e0199814 Google Scholar

9. Dreber A., Ellingsen T., Johannesson M., Rand D. G (2013) **Do people care about social context? Framing effects in dictator games** *Experimental Economics* **16**:349–371 Google Scholar

10. Dungan J. A., Chakroff A., Young L (2017) **The relevance of moral norms in distinct relational contexts: Purity versus harm norms regulate self-directed actions** *PLoS One* **12**:e0173405 Google Scholar

11. Li Z., Xu M (2024) **Oxytocin enhances group-based guilt in high moral disengagement individuals through increased moral responsibility** *Psychoneuroendocrinology* **168**:107131 Google Scholar

12. Linde J., Sonnemans J (2015) **Decisions under risk in a social and individual context: The limits of social preferences?** *Journal of Behavioral and Experimental Economics* **56**:62–71 Google Scholar

13. Bartz J. A., Zaki J., Bolger N., Ochsner K. N (2011) **Social effects of oxytocin in humans: Context and person matter** *Trends in Cognitive Sciences* **15**:301–309 Google Scholar

14. Crespi B. J (2016) **Oxytocin, testosterone, and human social cognition** *Biological Reviews* **91**:390–408 Google Scholar

15. Young L. J (2015) **Oxytocin, Social Cognition and Psychiatry** *Neuropsychopharmacology* **40**:243–244 Google Scholar

16. Greene J. D., Sommerville R. B., Nystrom L. E., Darley J. M., Cohen J. D (2001) **An fMRI investigation of emotional engagement in moral judgment** *Science* **293**:2105–2108 Google Scholar

17. Greene J. D., Nystrom L. E., Engell A. D., Darley J. M., Cohen J. D (2004) **The Neural Bases of Cognitive Conflict and Control in Moral Judgment** *Neuron* **44**:389–400 Google Scholar

18. Kim K. H., Guinote A (2022) **Cheating to win or not to lose: Power and situational framing affect unethical behavior** *Journal of Applied Social Psychology* **52**:137–144 Google Scholar

19. Zhang Y., Zhai Y., Zhou X., Zhang Z., Gu R., Luo Y., Feng C (2023) **Loss context enhances preferences for generosity but reduces preferences for honesty: Evidence from a combined behavioural-computational approach** *European Journal of Social Psychology* **53**:183–194 Google Scholar

20. Steinel W., Valtcheva K., Gross J., Celse J., Max S., Shalvi S (2022) **(Dis)honesty in the face of uncertain gains or losses** *Journal of Economic Psychology* **90**:102487 Google Scholar

21. Everett J. A., Faber N. S., Crockett M. J (2015) **The influence of social preferences and reputational concerns on intergroup prosocial behaviour in gains and losses contexts** *Royal Society Open Science* **2**:150546 Google Scholar

22. Li J., Sun Y., Li M., Li H., Fan W., Zhong Y (2020) **Social distance modulates prosocial behaviors in the gain and loss contexts: An event-related potential (ERP) study** *International Journal of Psychophysiology* **150**:83–91 Google Scholar

23. Kahneman D., Tversky A (1979) **Prospect Theory: An Analysis of Decision under Risk** *Econometrica* **47**:263 Google Scholar

24. Tom S. M., Fox C. R., Trepel C., Poldrack R. A (2007) **The neural basis of loss aversion in decision-making under risk** *Science* **315**:515–518 Google Scholar

25. Liu J., Gu R., Liao C., Lu J., Fang Y., Xu P., Luo Y., Cui F (2020) **The Neural Mechanism of the Social Framing Effect: Evidence from fMRI and tDCS Studies** *The Journal of Neuroscience* **40**:3646–3656 Google Scholar

26. Usher M., McClelland J. L (2004) **Loss Aversion and Inhibition in Dynamical Models of Multialternative Choice** *Psychological Review* **111**:757–769 Google Scholar

27. Yechiam E., Hochman G (2013) **Losses as modulators of attention: Review and analysis of the unique effects of losses over gains** *Psychological Bulletin* **139**:497–518 Google Scholar

28. Pachur T., Schulte-Mecklenbeck M., Murphy R. O., Hertwig R (2018) **Prospect theory reflects selective allocation of attention** *Journal of Experimental Psychology: General* **147**:147–169 Google Scholar

29. De Martino B., Camerer C. F., Adolphs R. (2010) **Amygdala damage eliminates monetary loss aversion** *Proceedings of the National Academy of Sciences* **107**:3788–3792 Google Scholar

30. Charpentier C. J., Martino B. D., Sim A. L., Sharot T., Roiser J. P (2016) **Emotion-induced loss aversion and striatal-amygdala coupling in low anxious individuals** *Social Cognitive and Affective Neuroscience* **11**:569–579 Google Scholar

31. Becker S., Gandhi W., Pomares F., Wager T. D., Schweinhardt P (2017) **Orbitofrontal cortex mediates pain inhibition by monetary reward** *Social Cognitive and Affective Neuroscience* **12**:651–661 Google Scholar

32. Carlino E., Frisaldi E., Benedetti F (2014) **Pain and the context** *Nature Reviews Rheumatology* **10**:6 Google Scholar

33. Jiang D., Tang J., Guan Q., Cui F., Luo Y. J (2021) **Money gained through suffering is less valuable: Pain reduces the sensitivity to outcome magnitude in monetary decision making** *Social Neuroscience* :1–9 Google Scholar

34. Pozo M., Milà-Guasch M., Haddad-Tóvolli R., Boudjadja M. B., Chivite I., Toledo M., Gómez-Valadés A. G., Eyre E., Ramírez S., Obri A., Ben-Ami Bartal I., D'Agostino G., Costa-Font J., Claret M (2023) **Negative energy balance hinders prosocial helping behavior** *Proceedings of the National Academy of Sciences* **120**:e2218142120 Google Scholar

35. Ma N., Li N., He X.-S., Sun D.-L., Zhang X., Zhang D.-R (2012) **Rejection of Unfair Offers Can Be Driven by Negative Emotions, Evidence from Modified Ultimatum Games with Anonymity** *PLoS One* **7**:e39619 Google Scholar

36. McDonald K., Graves R., Yin S., Weese T., Sinnott-Armstrong W (2021) **Valence framing effects on moral judgments: A meta-analysis** *Cognition* **212**:104703 Google Scholar

37. Cao F., Zhang J., Song L., Wang S., Miao D., Peng J (2017) **Framing Effect in the Trolley Problem and Footbridge Dilemma** *Psychological Reports* **120**:88–101 Google Scholar

38. Evans A. M., Van Beest I. (2017) **Gain-loss framing effects in dilemmas of trust and reciprocity** *Journal of Experimental Social Psychology* **73**:151–163 Google Scholar

39. Yang J., Gu R., Liu J., Deng K., Huang X., Luo Y.-J., Cui F (2022) **To Blame or Not? Modulating Third-Party Punishment with the Framing Effect** *Neuroscience Bulletin* **38**:533–547 Google Scholar

40. Petrovic P., Kalisch R., Singer T., Dolan R. J (2008) **Oxytocin Attenuates Affective Evaluations of Conditioned Faces and Amygdala Activity** *Journal of Neuroscience* **28**:6607–6615 Google Scholar

41. Meyer-Lindenberg A., Domes G., Kirsch P., Heinrichs M (2011) **Oxytocin and vasopressin in the human brain: Social neuropeptides for translational medicine** *Nature Reviews Neuroscience* **12**:524–538 Google Scholar

42. Kendrick K. M. (2000) **Oxytocin, motherhood and bonding** *Experimental Physiology* **85**:111–124 Google Scholar

43. Young L. J., Wang Z (2004) **The neurobiology of pair bonding** *Nature Neuroscience* **7**:1048–1054 Google Scholar

44. Dunbar R. I. M., Shultz S (2007) **Evolution in the Social Brain** *Science* **317**:1344–1347 Google Scholar

45. Zak P. J., Stanton A. A., Ahmadi S (2007) **Oxytocin Increases Generosity in Humans** *PLoS One* **2**:e1128 Google Scholar

46. Barchi-Ferreira A., Osório F (2021) **Associations between oxytocin and empathy in humans: A systematic literature review** *Psychoneuroendocrinology* **129**:105268 Google Scholar

47. Abu-Akel A., Palgi S., Klein E., Decety J., Shamay-Tsoory S (2015) **Oxytocin increases empathy to pain when adopting the other- but not the self- perspective** *Social Neuroscience* **10**:7–15 Google Scholar

48. Radke S., Roelofs K., De Bruijn E. R. A. (2013) **Acting on Anger: Social Anxiety Modulates Approach-Avoidance Tendencies After Oxytocin Administration** *Psychological Science* **24**:1573–1578 Google Scholar

49. Evans S., Shergill S. S., Averbeck B. B (2010) **Oxytocin Decreases Aversion to Angry Faces in an Associative Learning Task** *Neuropsychopharmacology* **35**:13 Google Scholar

50. Kirsch P., Esslinger C., Chen Q., Mier D., Lis S., Siddhanti S., Gruppe H., Mattay V. S., Gallhofer B., Meyer-Lindenberg A (2005) **Oxytocin Modulates Neural Circuitry for Social Cognition and Fear in Humans** *The Journal of Neuroscience* **25**:11489–11493 Google Scholar

51. Wu Q., Mao J., Li J (2020) **Oxytocin alters the effect of payoff but not base rate in emotion perception** *Psychoneuroendocrinology* **114**:104608 Google Scholar

52. Kapetaniou G. E., Reinhard M. A., Christian P., Jobst A., Tobler P. N., Padberg F., Soutschek A (2021) **The role of oxytocin in delay of gratification and flexibility in non-social decision making** *eLife* **10**:e61844 https://doi.org/10.7554/eLife.61844 | Google Scholar

53. Zheng X., Wang J., Yang X., Xu L., Becker B., Sahakian B. J., Robbins T. W., Kendrick K. M (2024) **Oxytocin, but not vasopressin, decreases willingness to harm others by promoting moral emotions of guilt and shame** *Molecular Psychiatry* **29**:3475–3482 Google Scholar

54. Han S., Ma Y (2015) **A Culture–Behavior–Brain Loop Model of Human Development** *Trends in Cognitive Sciences* **19**:666–676 Google Scholar

55. Ma Y., Li S., Wang C., Liu Y., Li W., Yan X., Chen Q., Han S (2016) **Distinct oxytocin effects on belief updating in response to desirable and undesirable feedback** *Proceedings of the National Academy of Sciences* **113**:9256–9261 Google Scholar

56. Ma Y., Liu Y., Rand D. G., Heatherton T. F., Han S (2015) **Opposing Oxytocin Effects on Intergroup Cooperative Behavior in Intuitive and Reflective Minds** *Neuropsychopharmacology* **40**:2379–2387 Google Scholar

57. Liu Y., Li S., Lin W., Li W., Yan X., Wang X., Pan X., Rutledge R. B., Ma Y (2019) **Oxytocin modulates social value representations in the amygdala** *Nature Neuroscience* **22**:633–641 Google Scholar

58. Zhang H., Gross J., De Dreu C., Ma Y. (2019) **Oxytocin promotes coordinated out-group attack during intergroup conflict in humans** *eLife* **8**:e40698 https://doi.org/10.7554/eLife.40698 | Google Scholar

59. De Dreu C. K. W., Greer L. L., Handgraaf M. J. J., Shalvi S., Van Kleef G. A., Baas M., Ten Velden F. S., Van Dijk E., Feith S. W. W. (2010) **The Neuropeptide Oxytocin Regulates Parochial Altruism in Intergroup Conflict Among Humans** *Science* **328**:1408–1411 Google Scholar

60. De Dreu C. K. W., Gross J., Fariña A., Ma Y. (2020) **Group Cooperation, Carrying-Capacity Stress, and Intergroup Conflict** *Trends in Cognitive Sciences* **24**:760–776 Google Scholar

61. Patin A., Scheele D., Hurlemann R (2018) **Oxytocin and Interpersonal Relationships** In: Hurlemann R., Grinevich V., editors. *Behavioral Pharmacology of Neuropeptides: Oxytocin* Springer International Publishing pp. 389–420 Google Scholar

62. Barraza J. A., Zak P. J (2009) **Empathy toward Strangers Triggers Oxytocin Release and Subsequent Generosity** *Annals of the New York Academy of Sciences* **1167**:182–189 Google Scholar

63. Stevens F., Taber K (2021) **The neuroscience of empathy and compassion in pro-social behavior** *Neuropsychologia* **159**:107925 Google Scholar

64. Awad E., Dsouza S., Shariff A., Rahwan I., Bonnefon J.-F (2020) **Universals and variations in moral decisions made in 42 countries by 70,000 participants** *Proceedings of the National Academy of Sciences* **117**:2332–2337 Google Scholar

65. Everett J. A. C., Colombatto C., Awad E., Boggio P., Bos B., Brady W. J., Chawla M., Chituc V., Chung D., Drupp M. A., Goel S., Grosskopf B., Hjorth F., Ji A., Kealoha C., Kim J. S., Lin Y., Ma Y., Maréchal M. A., ..., Crockett M. J (2021) **Moral dilemmas and trust in leaders during a global health crisis** *Nature Human Behaviour* **5**:1074–1088 Google Scholar

66. Zhu L., Jenkins A. C., Set E., Scabini D., Knight R. T., Chiu P. H., King-Casas B., Hsu M (2014) **Damage to dorsolateral prefrontal cortex affects tradeoffs between honesty and self-interest** *Nature Neuroscience* **17**:1319–1321 Google Scholar

67. Harari-Dahan O., Bernstein A (2014) **A general approach-avoidance hypothesis of Oxytocin: Accounting for social and non-social effects of oxytocin** *Neuroscience & Biobehavioral Reviews* **47**:506–519 Google Scholar

68. Harari-Dahan O., Bernstein A (2017) **Oxytocin attenuates social and non- social avoidance: Re-thinking the social specificity of Oxytocin** *Psychoneuroendocrinology* **81**:105–112 Google Scholar

69. Wang D., Ma Y (2020) **Oxytocin facilitates valence-dependent valuation of social evaluation of the self** *Communications Biology* **3**:433 Google Scholar

70. Rash J. A., Aguirre-Camacho A., Campbell T. S (2014) **Oxytocin and Pain: A Systematic Review and Synthesis of Findings** *The Clinical Journal of Pain* **30**:453–462 Google Scholar

71. Boll S., Almeida De Minas A. C., Raftogianni A., Herpertz S. C., Grinevich V (2018) **Oxytocin and Pain Perception: From Animal Models to Human Research** *Neuroscience* **387**:149–161 Google Scholar

72. Lopes S., Osório F. D. L (2023) **Effects of intranasal oxytocin on pain perception among human subjects: A systematic literature review and meta-analysis** *Hormones and Behavior* **147**:105282 Google Scholar

73. Faul F., Erdfelder E., Buchner A., Lang A.-G (2009) **Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses** *Behavior Research Methods* **41**:1149–1160 Google Scholar

74. Lynn S. K., Hoge E. A., Fischer L. E., Barrett L. F., Simon N. M (2014) **Gender differences in oxytocin-associated disruption of decision bias during emotion perception** *Psychiatry Research* **219**:198–203 Google Scholar

75. Hoge E. A., Anderson E., Lawson E. A., Bui E., Fischer L. E., Khadge S. D., Barrett L. F., Simon N. M (2014) **Gender moderates the effect of oxytocin on social judgments** *Human Psychopharmacology: Clinical and Experimental* **29**:299–304 Google Scholar

76. Walum H., Waldman I. D., Young L. J (2016) **Statistical and Methodological Considerations for the Interpretation of Intranasal Oxytocin Studies** *Biological Psychiatry* **79**:251–257 Google Scholar

77. Vlaev I., Seymour B., Dolan R. J., Chater N (2009) **The Price of Pain and the Value of Suffering** *Psychological Science* **20**:309–317 Google Scholar

78. Davis M. H (2011) **Interpersonal Reactivity Index [Dataset]**

79. Wilkinson G. N., Rogers C. E (1973) **Symbolic Description of Factorial Models for Analysis of Variance** *Applied Statistics* **22**:392 Google Scholar

80. Preacher K. J., Hayes A. F (2008) **Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models** *Behavior Research Methods* **40**:879–891 Google Scholar

81. Preacher K. J., Rucker D. D., Hayes A. F (2007) **Addressing Moderated Mediation Hypotheses: Theory, Methods, and Prescriptions** *Multivariate Behavioral Research* **42**:185–227 Google Scholar

82. Hayes A. F (2015) **An Index and Test of Linear Moderated Mediation** *Multivariate Behavioral Research* **50**:1–22 Google Scholar

## Author information

### Hong Zhang*

School of Psychological and Cognitive Sciences and Beijing Key Laboratory of Behavior and Mental Health, Peking University, Beijing, China

*Equal Contribution

**Yinmei Ni***
School of Psychological and Cognitive Sciences and Beijing Key Laboratory of Behavior and Mental Health, Peking University, Beijing, China
ORCID iD: 0000-0003-3614-1828

**For correspondence:** niyinmei@pku.edu.cn
*Equal Contribution

**Jian Li**
School of Psychological and Cognitive Sciences and Beijing Key Laboratory of Behavior and Mental Health, Peking University, Beijing, China, IDG/McGovern Institute for Brain Research, Peking University, Beijing, China
ORCID iD: 0000-0002-3941-2622

**For correspondence:** leekin@gmail.com

## Editors



**Reviewer #1 (Public review):**

Summary:

Zhang et al. addressed the question of whether hyperaltruistic preference is modulated by decision context and tested how oxytocin (OXT) may modulate this process. Using an adapted version of a previously well-established moral decision-making task, healthy human participants in this study undergo decisions that gain more (or lose less, termed as context) meanwhile inducing more painful shocks to either themselves or another person (recipient). The alternative choice is always less gain (or more loss) meanwhile less pain. Through a series of regression analyses, the authors reported that hyperaltruistic preference can only be found in the gain context but not in the loss context, however, OXT reestablished the hyperaltruistic preference in the loss context similar to that in the gain context.

Strengths:

This is a solid study that directly adapted a previously well-established task and the analytical pipeline to assess hyperaltruistic preference in separate decision contexts. Context-dependent decisions have gained more and more attention in literature in recent years, hence this study is timely. It also links individual traits (via questionnaires) with task performance, to test potential individual differences. The OXT study is done with great methodological rigor, including pre-registration. Both studies have proper power analysis to determine the sample size.

Weaknesses:

Despite the strengths, multiple analytical decisions have to be explained, justified, or clarified. Also, there is scope to enhance the clarity and coherence of the writing - as it stands, readers will have to go back and forth to search for information. Last, it would be helpful to add line numbers in the manuscript during the revision, as this will help all reviewers to locate the parts we are talking about.

Introduction:
(1) The introduction is somewhat unmotivated, with key terms/concepts left unexplained until relatively late in the manuscript. One of the main focuses in this work is "hyperaltruistic", but how is this defined? It seems that the authors take the meaning of "willing to pay more to reduce other's pain than their own pain", but is this what the task is measuring? Did participants ever need to PAY something to reduce the other's pain? Note that some previous studies indeed allow participants to pay something to reduce other's pain. And what makes it "HYPER-altruistic" rather than simply "altruistic"? Plus, in the intro, the authors mentioned that the "boundary conditions" remain unexplored, but this idea is never touched again. What do boundary conditions mean here in this task? How do the results/data help with finding out the boundary conditions? Can this be discussed within wider literature in the Discussion section? Last, what motivated the authors to examine decision context? It comes somewhat out of the blue that the opening paragraph states that "We set out to [...] decision context", but why? Are there other important factors? Why decision context is more important than studying those others?

Experimental design:
(2) The experiment per se is largely solid, as it followed a previously well-established protocol. But I am curious about how the participants got instructed? Did the experimenter ever mention the word "help" or "harm" to the participants? It would be helpful to include the exact instructions in the SI.

(3) Relatedly, the experimental details were not quite comprehensive in the main text. Indeed, Methods come after the main text, but to be able to guide readers to understand what was going on, it would be very helpful if the authors could include some necessary experimental details at the beginning of the Results section.

Statistical analysis
(3) One of the main analyses uses the harm aversion model (Eq1) and the results section keeps referring to one of the key parameters of it (ie, k). However, it is difficult to understand the text without going to the Methods section below. Hence it would be very helpful to repeat the equation also in the main text. A similar idea goes to the delta_m and delta_s terms - it will be very helpful to give a clear meaning of them, as nearly all analyses rely on knowing what they mean.

(4) There is one additional parameter gamma (choice consistency) in the model. Did the authors also examine the task-related difference of gamma? This might be important as some studies have shown that the other-oriented choice consistency may differ in different prosocial contexts.

(5) I am not fully convinced that the authors included two types of models: the harm aversion model and logistic regression models. Indeed, the models look similar, and the authors have acknowledged that. But I wonder if there is a way to combine them? For example:
Choice ~ delta_V * context * recipient (*Oxt_v._placebo)
The calculation of delta_V follows Equation 1.
Or the conceptual question is, if the authors were interested in the specific and independent contribution of dalta_m and dalta_s to behavior, as their logistic model did, why the authors examine the harm aversion first, where a parameter k is controlling for the trade-off? One

way to find it out is to properly run different models and run model comparison. In the end, it would be beneficial to only focus on the "winning" model to draw inferences.

(6) The interpretation of the main OXT results needs to be more cautious. According to the operationalization, "hyperaltruistic" is the reduction of pain of others (higher % of choosing the less painful option) relative to the self. But relative to the placebo (as baseline), OXT did not increase the % of choosing the less painful option for others, rather, it decreased the % of choosing the less painful option for themselves. In other words, the degree of reducing other's pain is the same under OXT and placebo, but the degree of benefiting self-interest is reduced under OXT. I think this needs to be unpacked, and some of the wording needs to be changed. I am not very familiar with the OXT literature, but I believe it is very important to differentiate whether OXT is doing something on self-oriented actions vs other-oriented actions. Relatedly, for results such as that in Fig5A, it would be helpful to not only look at the difference, but also the actual magnitude of the sensitivity to the shocks, for self and others, under OXT and placebo.

Comments on revisions:

I did not change my original public review, as I think it can still be helpful for the field to see the reasoning and argument.

For the revision, the authors have done a thorough job of addressing my previous comments and questions.

The only aspect I would like to ask is that, it would still be great to have a clear definition of hyperaltruism. As it stands, hyperaltruism refers to "people's willingness to pay more to reduce other's pain than
their own pain", ie, this means the "hyper" bit is considered with respect to "self". But shouldn't hyperaltruism be classified contrasting "normal" altruism?

It is fine that it follows a previously published work (Crockett et al., 2014), but it would still be necessary to explain/define the construct being tested in a standalone fashion rather than letting readers to go back to the original work.

https://doi.org/10.7554/eLife.102756.2.sa3

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

Summary:

In this manuscript, the authors reported two studies where they investigated the context effect of hyperaltruistic tendency in moral decision-making. They replicated the hyperaltruistic moral preference in the gain domain, where participants inflicted electric shocks to themselves or another person in exchange for monetary profits for themselves. In the loss domain, such hyperaltruistic tendency abolished. Interestingly, oxytocin administration reinstated the hyperaltruistic tendency in the loss domain. The authors also examined the correlation between individual differences in utilitarian psychology and the context effect of hyperaltruistic tendency.

Strengths:

(1) The research question - the boundary condition of hyperaltruistic tendency in moral decision-making and its neural basis - is theoretically important.
(2) Manipulating the brain via pharmacological means offers causal understanding of the neurobiological basis of the psychological phenomenon in question.
(3) Individual difference analysis reveals interesting moderators of the behavioral tendency.

Weaknesses:

(1) The theoretical hypothesis needs to be better justified. There are studies addressing the neurobiological mechanism of hyperaltruistic tendency, which the authors unfortunately skipped entirely.
(2) There are some important inconsistencies between the preregistration and the actual data collection/analysis, which the authors did not justify.
(3) Some of the exploratory analysis seems underpowered (e.g., large multiple regression models with only about 40 participants).
(4) Inaccurate conceptualization of utilitarian psychology and the questionnaire used to measure it.

Comments on revisions:

The authors have addressed the weakness in the second round of revision

https://doi.org/10.7554/eLife.102756.2.sa2

**Reviewer #3 (Public review):**

Summary:

In this study, the authors aimed to index individual variation in decision-making when decisions pit the interests of the self (gains in money, potential for electric shock) against the interests of an unknown stranger in another room (potential for unknown shock). In addition, the authors conducted an additional study in which male participants were either administered intranasal oxytocin or placebo before completing the task to identify the role of oxytocin in moderating task responses. Participants' choice data was analyzed using a harm aversion model in which choices were driven by the subjective value difference between the less and more painful options.

Strengths:

Overall, I think this is a well-conducted, interesting, and novel set of research studies exploring decision-making that balances outcomes for the self versus a stranger, and the potential role of the hormone oxytocin (OT) in shaping these decisions. The pain component of the paradigm is well designed, as is the decision-making task, and overall the analyses were well suited to evaluating and interpreting the data. Advantages of the task design include the absence of deception, e.g., the use of a real study partner and real stakes, as a trial from the task was selected at random after the study and the choice the participant made were actually executed.

Weaknesses:

The primary weakness of the paper concerns its framing. Although it purports to be measuring "hyper-altruism," which is the same term used in prior similar (although not identical) designs, I do not believe the task constitutes altruism, but rather the decision to engage, or not engage, in instrumental aggression.

I continue to believe that when in the "other" trials the only outcome possible for the study partner is pain, and the only outcome possible for the participant is monetary gain, these trials measure decisions about instrumental aggression. That is the exact definition of instrumental aggression is: causing others harm for personal gain. Altruism is not equivalent to refraining from engaging in instrumental aggression, although some similar mechanisms may support both. True altruism would be to accept shocks to the self for the other's benefit (e.g., money). The interpretation of this task as assessing instrumental aggression is

supported by the fact that only the Instrumental Harm subscale of the OUS was associated with outcomes in the task, but not the Impartial Benevolence subscale. By contrast, the IB subscale is the one more consistently associated with altruism (e.g,. Kahane et al 2018; Amormino at al, 2022) I believe it is important for scientific accuracy for the paper, including the title, to be rewritten to reflect what it is testing.

Although I recognize similar tasks have been previously characterized as "hyper-altruism" I do not believe that is sufficient justification for continuing to promulgate this descriptor without any caveats. I hope the authors will engage more seriously with the idea that this is what the task is measuring.

Relatedly, in the introduction, I believe it would be important to discuss the non-symmetry of moral obligations related to help/harm--we have obligations not to harm strangers but no obligation to help strangers. This is another reason I do not think the term "hyper altruism" is a good description for this task--given it is typically viewed as morally obligatory not to harm strangers, choosing not to harm them is not "hyper" altruistic (and again, I do not view it as obviously altruism at all).

https://doi.org/10.7554/eLife.102756.2.sa1

**Author response:**

The following is the authors' response to the original reviews

> ***Reviewer #1:***
>
> *Despite the strengths, multiple analytical decisions have to be explained, justified, or clarified. Also, there is scope to enhance the clarity and coherence of the writing - as it stands, readers will have to go back and forth to search for information. Last, it would be helpful to add line numbers in the manuscript during the revision, as this will help all reviewers to locate the parts we are talking about.*

We thank the reviewer's suggestions have added the line numbers to the revised manuscript.

> *(1) Introduction:*
>
> *The introduction is somewhat unmotivated, with key terms/concepts left unexplained until relatively late in the manuscript. One of the main focuses in this work is "hyperaltruistic", but how is this defined? It seems that the authors take the meaning of "willing to pay more to reduce other's pain than their own pain", but is this what the task is measuring? Did participants ever need to PAY something to reduce the other's pain? Note that some previous studies indeed allow participants to pay something to reduce other's pain. And what makes it "HYPER-altruistic" rather than simply "altruistic"?*

As the reviewer noted, we adopted a well-established experimental paradigm to study the context-dependent effect on hyper-altruism. Altruism refers to the fact that people take others' welfare into account when making decisions that concern both parties. Research paradigms investigating altruistic behavior typically use a social decision task that requires participants to choose between options where their own financial interests are pitted against the welfare of others (FeldmanHall et al., 2015; Hu et al., 2021; Hutcherson et al., 2015; Teoh et al., 2020; Xiong et al., 2020). On the other hand, the hyperaltruistic tendency emphasizes subjects' higher valuation to other's pain than their own pain (Crockett et al., 2014, 2015, 2017; Volz et al., 2017). One example for the manifestation of hyperaltruism would be the following scenario: the subject is willing to forgo $2 to reduce others' pain by 1 unit (social-decision task) and only willing to forgo $1 to reduce the same amount of his/her own pain

(self-decision task) (Crockett et al., 2014). On the contrary, if the subjects are willing to forgo less money to reduce others' suffering in the social decision task than in the self-decision task, then it can be claimed that no hyperaltruism is observed. Therefore, hyperaltruistic preference can only be measured by collecting subjects' choices in both the self and social decision tasks and comparing the choices in both tasks.

In our task, as in the studies before ours (Crockett et al., 2014, 2015, 2017; Volz et al., 2017), subjects in each trial were faced with two options with different levels of pain on others and monetary payoffs on themselves. Based on subjects' choice data, we can infer how much subjects were willing to trade 1 unit of monetary payoff in exchange of reducing others' pain through the regression analysis (see Figure 1 and methods for the experimental details). We have rewritten the introduction and methods sections to make this point clearer to the audience.

> *Plus, in the intro, the authors mentioned that the "boundary conditions" remain unexplored, but this idea is never touched again. What do boundary conditions mean here in this task? How do the results/data help with finding out the boundary conditions? Can this be discussed within wider literature in the Discussion section?*

Boundary conditions here specifically refer to the variables or decision contexts that determine whether hyperaltruistic behavior can be elicited. Individual personality trait, motivation and social relationship may all be boundary conditions affecting the emergence of hyperaltruistic behavior. In our task, we specifically focused on the valence of the decision context (gain vs. loss) since previous studies only tested the hyperaltruistic preference in the gain context and the introduction of the loss context might bias subjects' hyperaltruistic behavior through implicit moral framing.

We have explained the boundary conditions in the revised introduction (Lines 45 ~ 49).

"However, moral norm is also context dependent: vandalism is clearly against social and moral norms yet vandalism for self-defense is more likely to be ethically and legally justified (the Doctrine of necessity). Therefore, a crucial step is to understand the boundary conditions for hyperaltruism."

> *Last, what motivated the authors to examine the decision context? It comes somewhat out of the blue that the opening paragraph states that "We set out to [...] decision context", but why? Are there other important factors? Why decision context is more important than studying those others?*

We thank the reviewer for the comment. The hyperaltruistic preference was originally demonstrated between conditions where subjects' personal monetary gain was pitted against others' pain (social-condition) or against subjects' own suffering (self-condition) (Crockett et al., 2014). Follow up studies found that subjects also exhibited strong egoistic tendencies if instead subjects needed to harm themselves for other's benefit in the social condition (by flipping the recipients of monetary gain and electric shocks) (Volz et al., 2017). However, these studies have primarily focused on the gain contexts, neglecting the fact that valence could also be an influential factor in biasing subjects' behavior (difference between gain and loss processing in humans). It is likely that replacing monetary gains with losses in the money-pain trade-off task might bias subjects' hyperaltruistic preference due to heightened vigilance or negative emotions in the face of potential loss (such as loss aversion) (Kahneman & Tversky, 1979; Liu et al., 2020; Pachur et al., 2018; Tom et al., 2007; Usher & McClelland, 2004; Yechiam & Hochman, 2013). Another possibility is that gain and loss contexts may elicit different subjective moral perceptions (or internal moral framings) in participants, affecting their hyperaltruistic preferences (Liu et al., 2017; Losecaat Vermeer et al., 2020; Markiewicz & Czupryna, 2018; Wu et al., 2018). In our manuscript, we did not strive to compare which

factors might be more important in eliciting hyperaltruistic behavior, but rather to demonstrate the crucial role played by the decision context and to show that the internal moral framing could be the mediating factor in driving subjects' hyperaltruistic behavior. In fact, we speculate that the egoistic tendencies found in the Volz et al. 2017 study was partly driven by the subjects' failure to engage the proper internal moral framing in the social condition (harm for self, see Volz et al., 2017 for details).

> *(2) Experimental Design:*
>
> *(2a) The experiment per se is largely solid, as it followed a previously well-established protocol. But I am curious about how the participants got instructed? Did the experimenter ever mention the word "help" or "harm" to the participants? It would be helpful to include the exact instructions in the SI.*

In the instructions, we avoided words such as "harm", "help", or other terms reminding subjects about the moral judgement of the decisions they were about to make. Instead, we presented the options in a neutral and descriptive manner, focusing only on the relevant components (shocks and money). The instructions for all four conditions are shown in supplementary Fig. 9.

> *(2b) Relatedly, the experimental details were not quite comprehensive in the main text. Indeed, the Methods come after the main text, but to be able to guide readers to understand what was going on, it would be very helpful if the authors could include some necessary experimental details at the beginning of the Results section.*

We thank the reviewer's suggestion. We have now provided a brief introduction of the experimental details in the revised results section (Lines 125 ~132).

"Prior to the money-pain trade-off task, we individually calibrated each subject's pain threshold using a standard procedure[4–6]. This allowed us to tailor a moderate electric stimulus that corresponded to each subject's subjective pain intensity. Subjects then engaged in 240 decision trials (60 trials per condition), acting as the "decider" and trading off between monetary gains or losses for themselves and the pain experienced by either themselves or an anonymous "pain receiver" (gain-self, gain-other, loss-self and loss-other, see Supplementary Fig. 8 for the instructions and also see methods for details)."

> *(3) Statistical Analysis*
> *(3a) One of the main analyses uses the harm aversion model (Eq1) and the results section keeps referring to one of the key parameters of it (ie, k). However, it is difficult to understand the text without going to the Methods section below. Hence it would be very helpful to repeat the equation also in the main text. A similar idea goes to the delta_m and delta_s terms - it will be very helpful to give a clear meaning of them, as nearly all analyses rely on knowing what they mean.*

We thank the reviewer's suggestion. We have now added the equation of the harm aversion model and provided more detailed description to the equations in the main text (Lines 150 ~155).

"We also modeled subjects' choices using an influential model where subjects' behavior could be characterized by the harm (electric shock) aversion parameter κ, reflecting the relative weights subjects assigned to Δm and Δs, the objective difference in money and shocks between the more and less painful options, respectively ($\Delta V=(1-\kappa)\Delta m - \kappa\Delta s$ Eq.1, See Methods for details)[4–6]. Higher κ indicates that higher sensitivity is assigned to Δs than Δm and vice versa."

> *(3b) There is one additional parameter gamma (choice consistency) in the model. Did the authors also examine the task-related difference of gamma? This might be important as some studies have shown that the other-oriented choice consistency may differ in different prosocial contexts.*

To examine the task-related difference of choice consistency (γ), we compared the performance of 4 candidate models:

Model 1 (M1): The choice consistency parameter γ remains constant across shock recipients (self vs. other) and decision contexts (gain vs. loss).

Model 2 (M2): γ differs between the self- and other-recipient conditions, with $\gamma_{self}$ and $\gamma_{other}$ representing the choice consistency when pain is inflicted on him/her-self or the other-recipient.

Model 3 (M3): γ differs between the gain and loss conditions, with $\gamma_{gain}$ and $\gamma_{loss}$ representing the choice consistencies in the gain and loss contexts, respectively.

Model 4 (M4): γ varies across four conditions, with $\gamma_{self\text{-}gain}$, $\gamma_{other\text{-}gain}$, $\gamma_{self\text{-}loss}$ and $\gamma_{other\text{-}loss}$ capturing the choice consistency in each condition.

Supplementary Fig. 10 shows, after fitting all the models to subjects' choice behavioral data, model 1 (M1) performed the best among all the four candidate models in both studies (1 & 2) with the lowest Bayesian Information Criterion (BIC). Therefore, we conclude that factors such as the shock recipients (self vs. other) and decision contexts (gain vs. loss) did not significantly influence subjects' choice consistency and report model results using the single choice consistency parameter.

> *(3c) I am not fully convinced that the authors included two types of models: the harm aversion model and the logistic regression models. Indeed, the models look similar, and the authors have acknowledged that. But I wonder if there is a way to combine them? For example:*
>
> *Choice ~ delta_V * context * recipient (*Oxt_v._placebo)*
>
> *The calculation of delta_V follows Equation 1.*
>
> *Or the conceptual question is, if the authors were interested in the specific and independent contribution of dalta_m and dalta_s to behavior, as their logistic model did, why did the authors examine the harm aversion first, where a parameter k is controlling for the trade-off? One way to find it out is to properly run different models and run model comparisons. In the end, it would be beneficial to only focus on the "winning" model to draw inferences.*

The reviewer raised an excellent point here. According to the logistic regression model, we have:

$$log(\frac{P}{1-P}) = \beta_0 + \beta_{\Delta m}\Delta m + \beta_{\Delta s}\Delta s$$

Where $P$ is the probability of selecting the less harmful option. Similarly, if we combine Eq.1 ($\Delta V = 1-\kappa)\Delta m - \kappa\Delta s$) and Eq.2 $P(\text{less painful option}) = \frac{1}{1+e^{-\gamma\Delta V}}$) of the harm aversion model, we

have:

$$log(\frac{p}{1-p}) = \gamma(1- \kappa)\Delta m - \gamma\kappa\Delta s$$

If we ignore the constant term $\beta_0$ from the logistic regression model, the harm aversion model is simply a reparameterization of the logistic regression model. The harm aversion model was implemented first to derive the harm aversion parameter (κ), which is an parameter in the range of [0 1] to quantify how subjects value the relative contribution of Δm and Δs between options in their decision processes. Since previous studies used the term $\kappa_{other}$-$\kappa_{self}$ to define the magnitude of hyperaltruistic preference, we adopted similar approach to compare our results with previous research under the same theoretical framework. However, in order to investigate the independent contribution of Δm and Δs, we will have to take γ into account (we can see that the $\beta_{\Delta m}$ and $\beta_{\Delta s}$ in the logistic regression model are not necessarily correlated by nature; however, in the harm aversion model the coefficients (1-κ) and κ is always strictly negatively correlated (see Eq. 1). Only after multiplying γ, the correlation between γ(1-κ) and γκ will vary depending on the specific distribution of γ and κ). In summary, we followed the approach of previous research to estimate harm aversion parameter κ to compare our results with previous studies and to capture the relative influence between Δm and Δs. When we studied the contextual effects (gain vs. loss or placebo vs. control) on subjects' behavior, we further investigated the contextual effect on how subjects evaluated Δm and Δs, respectively. The two models (logistic regression model and harm aversion model) in our study are mathematically the same and are not competitive candidate models. Instead, they represent different aspects from which our data can be examined.

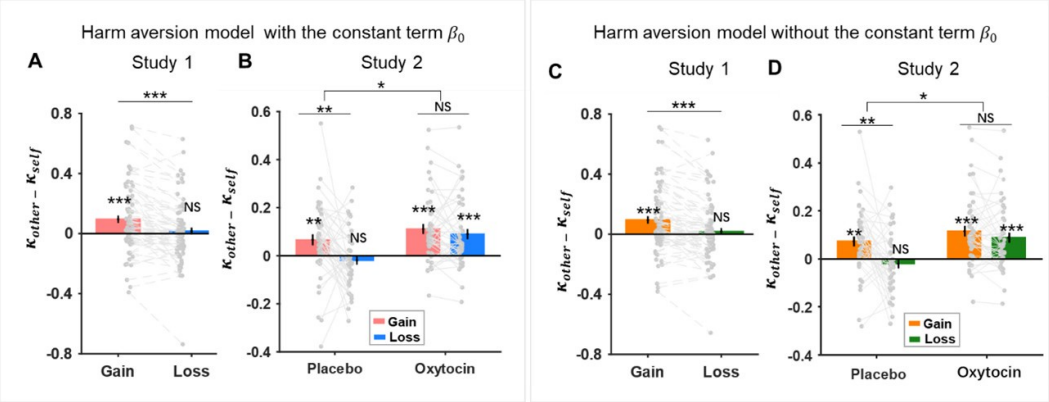
We also compared the harm aversion model with and without the constant term $\beta_0$ in the choice function. Adding a constant term $\beta_0$ the above Equation 2 becomes:

$$P(\text{less painful option}) = \frac{1}{1 + e^{-\gamma\Delta V + \beta_0}}$$

As the following figure shows, the hyperaltruistic parameters ($\kappa_{other}$-$\kappa_{self}$) calculated from the harm aversion model with the constant term (panels A & B) have almost identical patterns as the model without the constant term (panels C & D, i.e. Figs. 2B & 4B in the original manuscript) in both studies.

**Author response image 1.**

Figs. 2B & 4B in the original manuscript) in both studies.




*(3d) The interpretation of the main OXT results needs to be more cautious. According to the operationalization, "hyperaltruistic" is the reduction of pain of others (higher % of choosing the less painful option) relative to the self. But relative to the placebo (as baseline), OXT did not increase the % of choosing the less painful option for others, rather, it decreased the % of choosing the less painful option for themselves. In other words, the degree of reducing other's pain is the same under OXT and placebo, but the degree of benefiting self-interest is reduced under OXT. I think this needs to be unpacked, and some of the wording needs to be changed. I am not very familiar with the OXT literature, but I believe it is very important to differentiate whether OXT is doing something on self-oriented actions vs other-oriented actions. Relatedly, for results such as that in Figure 5A, it would be helpful to not only look at the difference but also the actual magnitude of the sensitivity to the shocks, for self and others, under OXT and placebo.*

We thank the reviewer for this thoughtful comment. As the reviewer correctly pointed out, “hyperaltruism” can be defined as “higher % of choosing the less painful option to the others relative to the self”. Closer examination of the results showed that both the degrees of reducing other’s pain as well as reducing their own pain decreased under OXT (Figure 4A). More specifically, our results do not support the claim that “In other words, the degree of reducing others’ pain is the same under OXT and placebo, but the degree of benefiting self-interest is reduced under OXT.” Instead, the results show a significant reduction in the choice of less painful option under OXT treatment for both the self and other conditions (the interaction effect of OXT vs. placebo and self vs. other: $F_{1.45}$= 16.812, $P$ < 0.001, $\eta^2$ = 0.272, simple effect OXT vs. placebo in the self- condition: $F_{1.45}$=59.332, $P$ < 0.001, $\eta^2$ = 0.569, OXT vs. placebo in the other-condition: $F_{1.45}$= 14.626, $P$ < 0.001, $\eta^2$ = 0.245, repeated ANOVA, see Figure 4A).

We also performed mixed-effect logistic regression analyses where subjects’ choices were regressed against and in different valences (gain vs. loss) and recipients (self vs. other) conditions in both studies 1 & 2 (Supplementary Figs. 1 & 6). As we replot supplementary Fig. 6 and panel B (included as Supplementary Fig. 8 in the supplementary materials) in the above figure, we found a significant treatment × $\Delta_s$ (differences in shock magnitude between the more and less painful options) interaction effect β=0.136±0.029$P$ < =0.001, 95% CI=[-0.192, -0.079]), indicating that subject’s sensitivities towards pain were indeed different between the placebo and OXT treatments for both self and other conditions. Furthermore, the significant four-way $\Delta_s$ × treatment (OXT vs. Placebo) × context (gain vs. loss) × recipient (self vs. other) interaction effect (β=0.125±0.053, $P$=0.018 95% CI=[0.022, 0.228]) in the regression analysis, followed by significant simple effects (In the OXT treatment: $\Delta_s$ × recipient effect in the gain context: $F_{1.45}$= 7.622, $P$ < 0.008, $\eta^2$ = 0.145; $\Delta_s$ × recipient effect in the loss context: $F_{1.45}$= 7.966, $P$ 0.007, $\eta^2$ = 0.150, suggested that under OXT treatment, participants showed a greater

sensitivity toward $\Delta_s$ (see asterisks in the OXT condition in panel B) in the other condition than the self-condition, thus restoring the hyperaltruistic behavior in loss context.

As the reviewer suggested, OXT's effect on hyperaltruism does manifest separately on subjects' harm sensitivities on self- and other-oriented actions. We followed the reviewer's suggestions and examined the actual magnitude of the sensitivities to shocks for both the self and other treatments (panel B in the figure above). It's clear that the administration of OXT (compared to the Placebo treatment, panel B in the figure above) significantly reduced participants' pain sensitivity (treatment × $\Delta_s$: β=-0.136±0.029, $P < 0.001$, 95% CI= [-0.192,-0.079]), yet also restored the harm sensitivity patterns in both the gain and loss conditions. These results are included in the supplementary figures (6 & 8) as well as in the main texts.

> *Recommendations:*
>
> *(1) For Figures 2A-B, it would be great to calculate the correlation separately for gain and loss, as in other figures.*

We speculate that the reviewer is referring to Figures 3A & B. Sorry that we did not present the correlations separately for the gain and loss contexts because the correlation between an individual's IH (instrumental harm), IB (impartial beneficence) and hyperaltruistic preferences was not significantly modulated by the contextual factors. The interaction effects in both Figs. 3A & B and Supplementary Fig.5 (also see Table S1& S2) are as following: Study1 valence × IH effect: β=0.016±0.022, $t_{152}$=0.726, $P$=0.469; valence × IB effect: β=0.004±0.031, $t_{152}$=0.115, $P$=0.908; Study2 placebo condition: valence × IH effect: β=0.018±0.024, $t_{84}$=0.030 $P$=0.463; valence × IB effect: β=0.051±0.030, $t_{84}$=1.711, $P$=0.702. We have added these statistics to the main text following the reviewer's suggestions.

> *(2) "by randomly drawing a shock increment integer Δs (from 1 to 19) such that [...] did not exceed 20 ($S$+ {less than or equal to} 20)." I am not sure if a random drawing following a uniform distribution can guarantee S is smaller than 20. More details are needed. Same for the monetary magnitude.*

We are sorry for the lack of clarity in the method description. As for the task design, we followed adopted the original design from previous literature (Crockett et al., 2014, 2017). More specifically:

"Specifically, each trial was determined by a combination of the differences of shocks (Δs, ranging from 1 to 19, with increment of 1) and money (Δm, ranging from ¥0.2 to ¥19.8, with increment of ¥0.2) between the two options, resulting in a total of 19×99=1881 pairs of [Δs, Δm]. for each trial. To ensure the trials were suitable for most subjects, we evenly distributed the desired ratio Δm / (Δs + Δm) between 0.01 and 0.99 across 60 trials for each condition. For each trial, we selected the closest [Δs, Δm] pair from the [Δs, Δm] pool to the specific Δm / (Δs + Δm) ratio, which was then used to determine the actual money and shock amounts of two options. The shock amount ($S_{less}$) for the less painful option was an integer drawn from the discrete uniform distribution [1-19], constraint by $S_{less}$ + Δs < 20. Similarly, the money amount ($M_{less}$) for the less painful option was drawn from a discrete uniform distribution [¥0.2 - ¥19.8], with the constraint of $M_{less}$ + Δm < 20. Once the $S_{less}$and $M_{less}$ were selected, the shock ($S_{more}$) and money ($M_{more}$) magnitudes for the more painful option were calculated as: $S_{more}$ = $S_{less}$ + Δs, $M_{more}$ = $M_{less}$ + Δm"

We have added these details to the methods section (Lines 520-533).

> ***Reviewer #2:***
>
> *(1) The theoretical hypothesis needs to be better justified. There are studies addressing the neurobiological mechanism of hyperaltruistic tendency, which the authors unfortunately skipped entirely.*

Also in recommendation #1:

> *(1) In the Introduction, the authors claim that "the mechanistic account of the hyperaltruistic phenomenon remains unknown". I think this is too broad of a criticism and does not do justice to prior work that does provide some mechanistic account of this phenomenon. In particular, I was surprised that the authors did not mention at all a relevant fMRI study that investigates the neural mechanism underlying hyperaltruistic tendency (Crockett et al., 2017, Nature Neuroscience). There, the researchers found that individual differences in hyperaltruistic tendency in the same type of moral decision-making task is better explained by reduced neural responses to ill-gotten money (Δm in the Other condition) in the brain reward system, rather than heightened neural responses to others' harm. Moreover, such neural response pattern is related to how an immoral choice would be judged (i.e., blamed) by the community. Since the brain reward system is consistently involved in Oxytocin's role in social cognition and decision-making (e.g., Dolen & Malenka, 2014, Biological Psychiatry), it is important to discuss the hypothesis and results of the present research in the context of this literature.*

We totally agree with the reviewer that the expression “mechanistic account of the hyperaltruistic phenomenon remains unknown” in our original manuscript can be misleading to the audience. Indeed, we were aware of the major findings in the field and cited all the seminal work of hyperaltruism and its related neural mechanism (Crockett et al., 2014, 2015, 2017). We have changed the texts in the introduction to better reflect this point and added further discussion as to how oxytocin might play a role:

“For example, it was shown that the hyperaltruistic preference modulated neural representations of the profit gained from harming others via the functional connectivity between the lateral prefrontal cortex, a brain area involved in moral norm violation, and profit sensitive brain regions such as the dorsal striatum6.” (Lines 41~45)

“Oxytocin has been shown to play a critical role in social interactions such as maternal attachment, pair bonding, consociate attachment and aggression in a variety of animal models[42,43]. Humans are endowed with higher cognitive and affective capacities and exhibit far more complex social cognitive patterns[44]. ” (Lines 86~90)

> *(2) There are some important inconsistencies between the preregistration and the actual data collection/analysis, which the authors did not justify.*

Also in recommendations:

(4) It is laudable that the authors pre-registered the procedure and key analysis of the Oxytocin study and determined the sample size beforehand. However, in the preregistration, the authors claimed that they would recruit 30 participants for Experiment 1 and 60 for Experiment 2, without justification. In the paper, they described a "prior power analysis", which deviated from their preregistration. It is OK to deviate from preregistration, but this needs to be explicitly mentioned and addressed (why the deviation occurred, why the reported approach was justifiable, etc.).

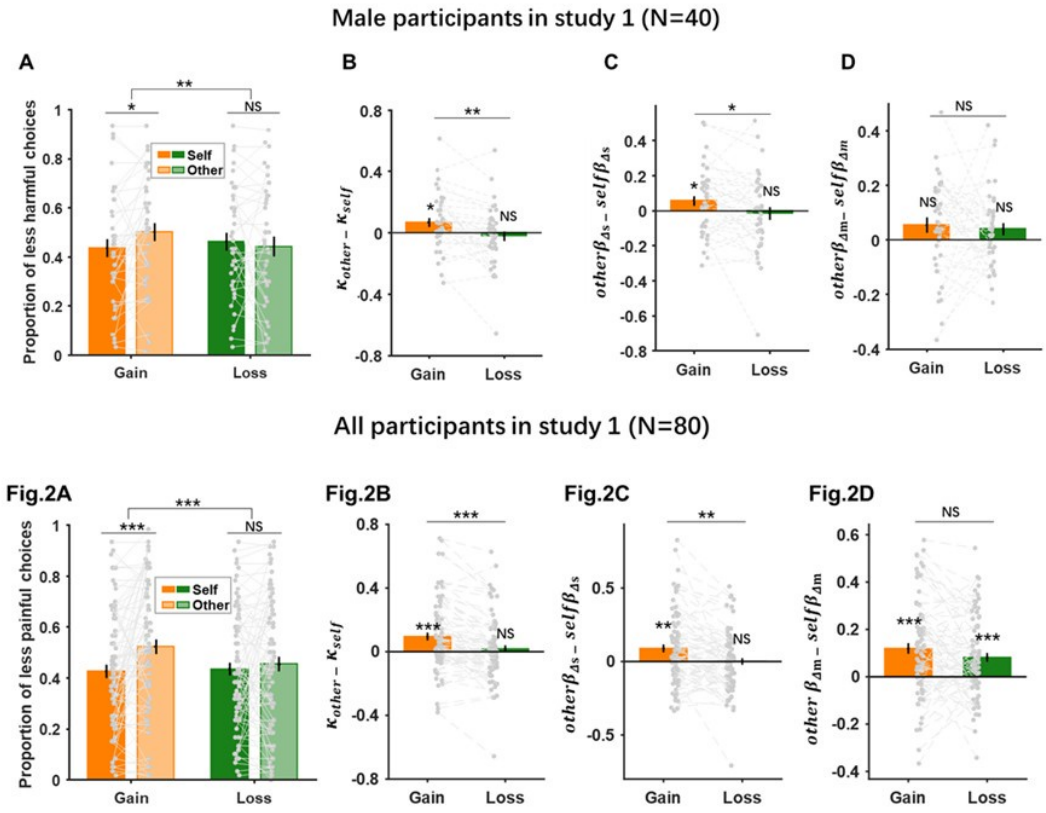
We sincerely appreciate the reviewer’s thorough assessment of our manuscript. In the more exploratory study 1, we found that the loss decision context effectively diminished subjects’

hyperaltruistic preference. Based on this finding, we pre-registered study 2 and hypothesized that: 1) The administration of OXT may salvage subject's hyperaltruistic preference in the loss context; 2) The administration of OXT may reduce subjects' sensitivities towards electric shocks (but not necessarily their moral preference), due to the well-established results relating OXT to enhanced empathy for others (Barchi-Ferreira & Osório, 2021; Radke et al., 2013) and the processing of negative stimuli(Evans et al., 2010; Kirsch et al., 2005; Wu et al., 2020); and 3) The OXT effect might be context specific, depending on the particular combination of valence (gain vs. loss) and shock recipient (self vs. other) (Abu-Akel et al., 2015; Kapetaniou et al., 2021; Ma et al., 2015).

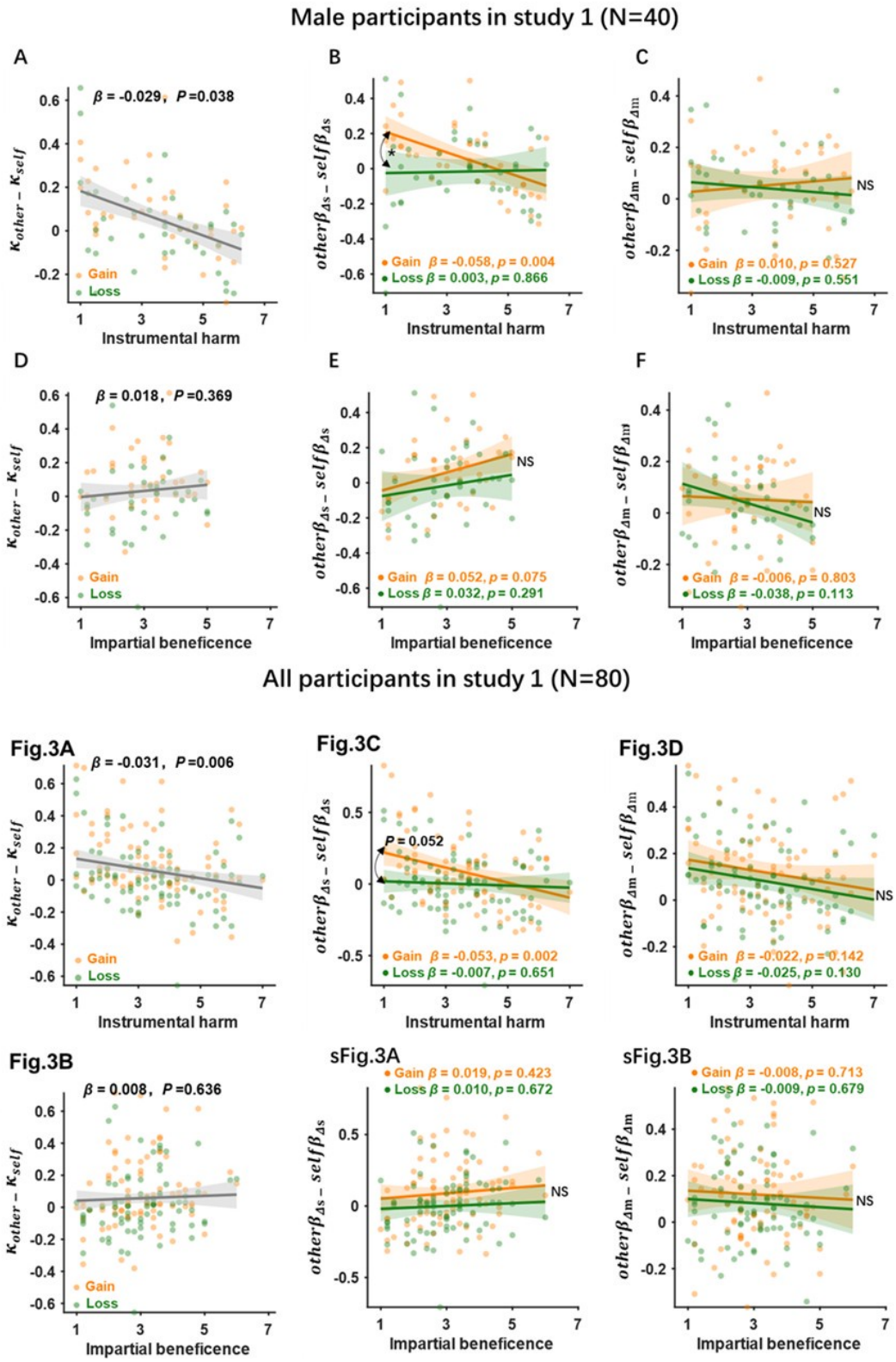
As our results suggested, the administration of OXT indeed restored subjects' hyperaltruistic preference (confirming hypothesis 1, Figure 4A). Also, OXT decreased subjects' sensitivities towards electric shocks in both the gain and loss conditions (supplementary Fig. 6 and supplementary Fig. 8), consistent with our second hypothesis. We must admit that our hypothesis 3 was rather vague, since a seminal study clearly demonstrated the context-dependent effect of OXT in human cooperation and conflict depending on the group membership of the subjects (De Dreu et al., 2010, 2020). Although our results partially validated our hypothesis 3 (supplementary Fig. 6), we did not make specific predictions as to the direction and the magnitude of the OXT effect.

The main inconsistency is related to the sample size. When we carried out study 1, we recruited both male and female subjects. After we identified the context effect on the hyperaltruistic preference, we decided to pre-register and perform study 2 (the OXT study). We originally made a rough estimate of 60 male subjects for study 2. While conducting study 2, we also went through the literature of OXT effect on social behavior and realized that the actual subject number around 45 might be enough to detect the main effect of OXT. Therefore, we settled on the number of 46 (study 2) reported in the manuscript. Correspondingly, we increased the subject number in study 1 to the final number of 80 (40 males) to make sure the subject number is enough to detect a small-to-medium effect, as well as to have a fair comparison between study 1 and 2 (roughly equal number of male subjects). It should be noted that although we only reported all the subjects (male & female) results of study 1 in the manuscript, the main results remain very similar if we only focus on the results of male subjects in study 1 (see the figure below). We believe that these results, together with the placebo treatment group results in study 2 (male only), confirmed the validity of our original finding.

**Author response image 2.**



**Author response image 3.**




We have included additional texts (Lines 447 ~ 452) in the Methods section for the discrepancy between the preregistered and actual sample sizes in the revised manuscript:

"It should be noted that in preregistration we originally planned to recruit 60 male subjects for Study 2 but ended up recruiting 46 male subjects (mean age = years) based on the sample size reported in previous oxytocin studies[57,69]. Additionally, a power analysis suggested that the sample size > 44 should be enough to detect a small to median effect size of oxytocin (Cohen's d=0.24, α=0.05, β=0.8) using a 2 × 2 × 2 within-subject design[76]."

> *(3) Some of the exploratory analysis seems underpowered (e.g., large multiple regression models with only about 40 participants).*

We thank the reviewer's comments and appreciate the concern that the sample size would be an issue affecting the results reliability in multiple regression analysis.

In Fig. 2, the multiple regression analyses were conducted after we observed a valence-dependent effect on hyperaltruism (Fig. 2A) and the regression was constructed accordingly:

*Choice ~ Δs *context*recipient + Δm *context*recipient+(1+ Δs *context*recipient + Δs*context*recipient | subject)*

Where Δs and Δm indicate the shock level and monetary reward difference between the more and loss painful options, context as the monetary valence (gain vs. loss) and recipient as the identity of the shock recipient (self vs. other).

Since we have 240 trials for each subject and a total of 80 subjects in Study 1, we believe that this is a reasonable regression analysis to perform.

In Fig. 3, the multiple regression analyses were indeed exploratory. More specifically, we ran 3 multiple linear regressions:

*hyperaltruism~EC*context+IH*context+IB*context*

*Relative harm sensitivity~ EC*context+IH*context+IB*context*

*Relative money sensitivity~ EC*context+IH*context+IB*context*

Where Hyperaltruism is defined as $\kappa_{other}$ - $\kappa_{self}$, Relative harm sensitivity as $other\beta_{\Delta s}$ - $self\beta_{\Delta s}$ and Relative monetary sensitivity as $other\beta_{\Delta m}$ - $self\beta_{\Delta m}$. EC (empathic concern), IH (instrumental harm) and IB (impartial beneficence) were subjects' scores from corresponding questionnaires.

For the first regression, we tested whether EC, IH and IB scores were related to hyperaltruism and it should be noted that this was tested on 80 subjects (Study 1). After we identified the effect of IH on hyperaltruism, we ran the following two regressions. The reason we still included IB and EC as predictors in these two regression analyses was to remove potential confounds caused by EC and IB since previous research indicated that IB, IH and EC could be correlated (Kahane et al., 2018).

In study 2, we performed the following regression analyses again to validate our results (Placebo treatment in study 2 should have similar results as found in study 1).

*Relative harm sensitivity~ EC*context+IH*context+IB*context*

*Relative money sensitivity~ EC*context+IH*context+IB*context*

Again, we added IB and EC only to control for the nuance effects by the covariates. As indicated in Fig. 5 C-D, the placebo condition in study 2 replicated our previous findings in study 1 and OXT administration effectively removed the interaction effect between IH and valence (gain vs. loss) on subjects' relative harm sensitivity.

To more objectively present our data and results, we have changed the texts in the results section and pointed out that the regression analysis:

*hyperaltruism~EC*context+IH*context+IB*context*

was exploratory (Lines 186-192).

"We tested how hyperaltruism was related to both IH and IB across decision contexts using an exploratory multiple regression analysis. Moral preference, defined as $\kappa_{other}$ - $\kappa_{self}$, was negatively associated with IH (β=-0.031±0.011, $t_{156}$=-2.784, $P$ =0.006) but not with IB (β=0.008±0.016, $t_{156}$=0.475, $P$=0.636) across gain and loss contexts, reflecting a general connection between moral preference and IH (Fig. 3A & B)."

> *(4) Inaccurate conceptualization of utilitarian psychology and the questionnaire used to measure it.*

Also in recommendations:

> *(2) Throughout the paper, the authors placed lots of weight on individual differences in utilitarian psychology and the Oxford Utilitarianism Scale (OUS). I am not sure this is the best individual difference measure in this context. I don't see a conceptual fit between the psychological construct that OUS reflects, and the key psychological processes underlying the behaviors in the present study. As far as I understand it, the conceptual core of utilitarian psychology that OUS captures is the maximization of greater goods. Neither the Instrumental Harm (IH) component nor the Impartial Beneficence (IB) component reflects a tradeoff between the personal interests of the decision-making agent and a moral principle. The IH component is about the endorsement of harming a smaller number of individuals for the benefit of a larger number of individuals. The IB component is about treating self, close others, and distant others equally. However, the behavioral task used in this study is neither about distributing harm between a smaller number of others and a larger number of others nor about benefiting close or distant others. The fact that IH showed some statistical association with the behavioral tendency in the present data set could be due to the conceptual overlap between IH and an individual's tendency to inflict harm (e.g., psychopathy; Table 7 in Kahane et al., 2018, which the authors cited). I urge the authors to justify more why they believe that conceptually OUS is an appropriate individual difference measure in the present study, and if so, interpret their results in a clearer and justifiable manner (taking into account the potential confound of harm tendency/psychopathy).*

We thank the reviewer for the thoughtful comment and agree that "IH component is about the endorsement of harming a smaller number of individuals for the benefit of a larger number of individuals. The IB component is about treating self, close others, and distant others equally". As we mentioned in the previous response to the reviewer, we first ran an exploratory multiple linear regression analysis of hyperaltruistic preference ($\kappa_{other}$ - $\kappa_{self}$) against IB and IH in study 1 based on the hypothesis that the reduction of hyperaltruistic preference in the loss condition might be due to 1) subjects' altered altitudes between IB and hyperaltruistic preference between the gain and loss conditions, and/or 2) the loss condition changed how the moral norm was perceived and therefore affected the correlation between IH and hyperaltruistic preference. As Fig. 3 shows, we did not find a significant IB effect on hyperaltruistic preference ($\kappa_{other}$ - $\kappa_{self}$), nor on the relative harm or money sensitivity (supplementary Fig. 3). These results excluded the possibility that subjects with higher IB might treat self and others more equally and therefore show less hyperaltruistic preference. On the other hand, we found a strong correlation between hyperaltruistic preference and IH (Fig. 3A): subjects with higher IH scores showed less hyperaltruistic preference. Since the

hyperaltruistic preference ($\kappa_{other}$ - $\kappa_{self}$) is a compound variable and we further broke it down to subjects' relative sensitivity to harm and money (*other* $\beta_{\Delta s}$ - *self* $\beta_{\Delta s}$ and *other* $\beta_{\Delta m}$ - *self* $\beta_{\Delta m}$, respectively). The follow up regression analyses revealed that the correlation between subjects' relative harm sensitivity and IH was altered by the decision contexts (gain vs. loss, Fig. 3C-D). These results are consistent with our hypothesis that for subjects to engage in the utilitarian calculation, they should first realize that there is a moral dilemma (harming others to make monetary gain in the gain condition). When there is less perceived moral conflict (due to the framing of decision context as avoiding loss in the loss condition), the correlation between subjects' relative harm sensitivity and IH became insignificant (Fig. 3C). It is worth noting that these results were further replicated in the placebo condition of study 2, further indicating the role of OXT is to affect how the decision context is morally framed.

The reviewer also raised an interesting possibility that the correlation between subject's behavioral tendency and IH may be confounded by the fact that IH is also correlated with other traits such as psychopathy. Indeed, in the Kahane et al., 2018 paper, the authors showed that IH was associated with subclinical psychopathy in a lay population. Although we only collected and included IB and Empathic concern (EC) scores as control variables and in principle could not rule out the influence of psychopathy, we argue it is unlikely the case. First, psychopaths by definition "only care about their own good" (Kahane et al., 2018). However, subjects in our studies, as well as in previous research, showed greater aversion to harming others (compared to harming themselves) in the gain conditions. This is opposite to the prediction of psychopathy. Even in the loss condition, subjects showed similar levels of aversion to harming others (vs. harming themselves), indicating that our subjects valuated their own and others' well-being similarly. Second, although there appears to be an association between utilitarian judgement and psychopathy(Glenn et al., 2010; Kahane et al., 2015), the fact that people also possess a form of universal or impartial beneficence in their utilitarian judgements suggest psychopathy alone is not a sufficient variable explaining subjects' hyperaltruistic behavior.

We have thus rewritten part of the results to clarify our rationale for using the Oxford Utilitarianism Scale (especially the IH and IB) to establish the relationship between moral traits and subjects' decision preference (Lines 212-215):

"Furthermore, our results are consistent with the claim that profiting from inflicting pains on another person (IH) is inherently deemed immoral1. Hyperaltruistic preference, therefore, is likely to be associated with subjects' IH dispositions."

> *(3) Relatedly, in the Discussion, the authors mentioned "the money-pain trade-off task, similar to the well-known trolley dilemma". I am not sure if this statement is factually accurate because the "well-known trolley dilemma" is about a disinterested third-party weighing between two moral requirements - "greatest good for the greatest number" (utilitarianism) and "do no harm" (Kantian/deontology), not between a moral requirement and one's own monetary interest (which is the focus of the present study). The analogy would be more appropriate if the task required the participants to trade off between, for example, harming one person in exchange for a charitable donation, as a recent study employed (Siegel et al., 2022, A computational account of how individuals resolve the dilemma of dirty money. Scientific reports). I urge the authors to go through their use of "utilitarian/utilitarianism" in the paper and make sure their usage aligns with the definition of the concept and the philosophical implications.*

We thank the reviewer for prompting us to think over the difference between our task and the trolley dilemma. Indeed, the trolley dilemma refers to a disinterested third-party's decision between two moral requirements, namely, the utilitarianism and deontology. In our study, when the shock recipient was "other", our task could be interpreted as either the decision between "moral norm of no harm (deontology) and one's self-interest maximization

(utilitarian)", or a decision between "greatest good for both parties (utilitarian) vs. do no harm (deontology)", though the latter interpretation typically requires differential weighing of own benefits versus the benefits of others(Fehr & Schmidt, 1999; Saez et al., 2015). In fact, it could be argued that the utilitarianism account applies not only to the third party's well-being, but also to our own well-being, or to "that of those near or dear to us" (Kahane et al., 2018).

We acknowledge that there may lack a direct analogy between our task and the trolley dilemma and therefore have deleted the trolley example in the discussion.

> *(5) Related to the above point, the sample size of Study 2 was calculated based on the main effect of oxytocin. However, the authors also reported several regression models that seem to me more like exploratory analyses. Their sample size may not be sufficient for these analyses. The authors should: a) explicitly distinguish between their hypothesis-driven analysis and exploratory analysis; b) report achieved power of their analysis.*

We appreciate the reviewer's thorough reading of our manuscript. Following the reviewer's suggestions, we have explicitly stated in the revised manuscript which analyses were exploratory, and which were hypothesis driven. Following the reviewer's request, we added the achieved power into the main texts (Lines 274-279):

"The effect size (Cohen's $f^2$) for this exploratory analysis was calculated to be 0.491 and 0.379 for the placebo and oxytocin conditions, respectively. The post hoc power analysis with a significance level of α = 0.05, 7 regressors (IH, IB, EC, decision context, IH×context, IB×context, and EC×context), and sample size of N = 46 yielded achieved power of 0.910 (placebo treatment) and 0.808 (oxytocin treatment)."

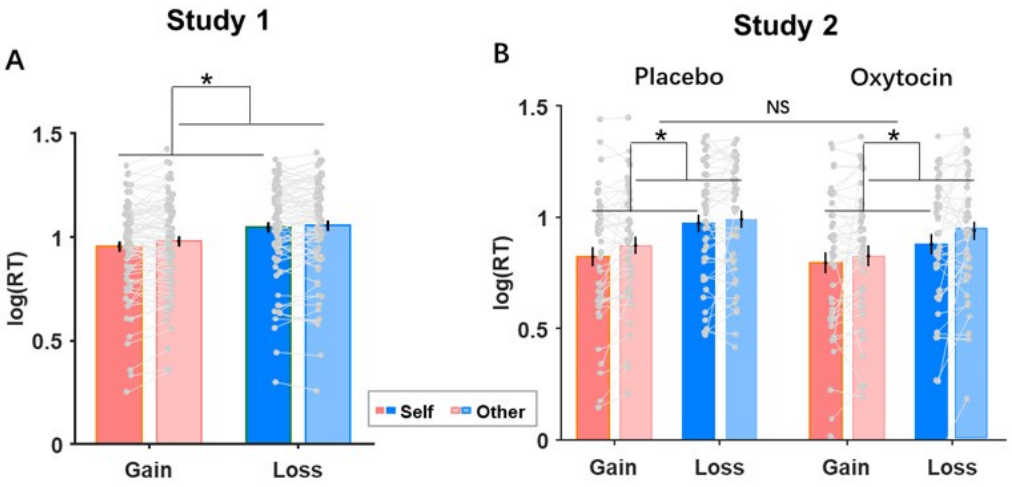
> (6) Do the authors collect reaction times (RT) information? Did the decision context and oxytocin modulate RT? Based on their procedure, it seems that the authors adopted a speeded response task, therefore the RT may reflect some psychological processes independent of choice. It is also possible (and recommended) that the authors use the drift-diffusion model to quantify latent psychological processes underlying moral decision-making. It would be interesting to see if their manipulations have any impact on those latent psychological processes, in addition to explicit choice, which is the endpoint product of the latent psychological processes. There are some examples of applying DDM to this task, which the authors could refer to if they decide to go down this route (Yu et al, 2021, How peer influence shapes value computation in moral decision-making. Cognition.)

We did collect the RT information for this experiment. As demonstrated in the figure below, participants exhibited significantly longer RT in the loss context compared to the gain context (Study1: the main effect of decision context: $F_{1,79}$=20.043, $P$ < 0.001, $\eta^2$ =0.202; Study2-placebo: $F_{1.45}$=17.177, $P$ < 0.001, $\eta^2$ =0.276). In addition to this effect of context, decisions were significantly slower in the other-condition compared to the self-condition

(Study1: the main effect of recipient: $F_{1,79}$=4.352, $P$ < 0.040, $\eta^2$ =0.052; Study2-placebo: $F_{1,45}$=5.601, $P$ < 0.022, $\eta^2$ =0.111) which replicates previous research findings (Crockett et al., 2014). However, the differences in response time between recipients was not modulated by decision context (Study1: context × recipient interaction: $F_{1,79}$=1.538, $P$ < 0.219, $\eta^2$ =0.019; Study2-placebo: $F_{1,45}$=2.631, $P$ < 0.112, $\eta^2$ =0.055). Additionally, the results in the oxytocin study (study 2) revealed no evidence supporting any effect of oxytocin on reaction time. Neither the main effect (treatment: placebo vs. oxytocin) nor the interaction effect of oxytocin on response time was statistically significant (main effect of OXT treatment: $F_{1,45}$=2.380, $P$ < 0.230, $\eta^2$ =0.050; treatment × context: $F_{1,45}$=2.075, $P$ < 0.157$\eta^2$ =0.044; treatment × recipient:

$F_{1,45}$=0.266, $P$ < 0.609, $\eta^2$ =0.006; treatment × context × recipient: $F_{1,45}$=2.909, $P$ < 0.095, $\eta^2$ =0.061).;

**Author response image 4.**




We also agree that it would be interesting to also investigate how the OXT might impact the dynamics of the decision process using a drift-diffusion model (DDM). However, we have already showed in the original manuscript that the OXT increased subjects' relative harm sensitivities. If a canonical DDM is adopted here, then such an OXT effect is more likely to correspond to the increased drift rate for the relative harm sensitivity, which we feel still aligns with the current framework in general. In future studies, including further manipulations such as time pressure might be a more comprehensive approach to investigate the effect of OXT on DDM related decision variables such as attribute drift rate, initial bias, decision threshold and attribute synchrony.

> *(7) This is just a personal preference, but I would avoid metaphoric language in a scientific paper (e.g., rescue, salvage, obliterate). Plain, neutral English terms can express the same meaning clearly (e.g., restore, vanish, eliminate).*

Again, we thank the reviewer for the suggestion and have since modified the terms.

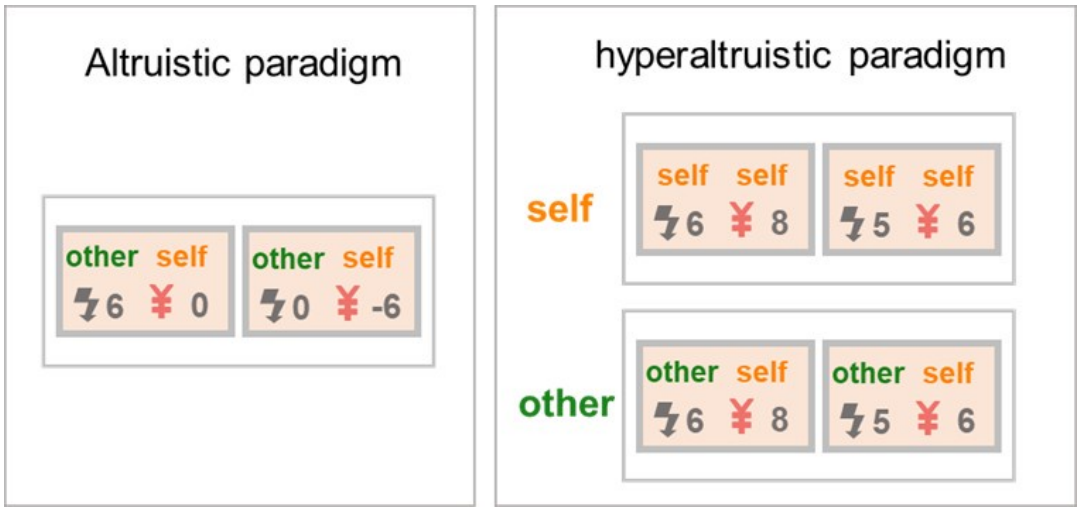
> ***Reviewer #3:***
>
> *The primary weakness of the paper concerns its framing. Although it purports to be measuring "hyper-altruism" it does not provide evidence to support why any of the behavior being measured is extreme enough to warrant the modifier "hyper" (and indeed throughout I believe the writing tends toward hyperbole, using, e.g., verbs like "obliterate" rather than "reduce"). More seriously, I do not believe that the task constitutes altruism, but rather the decision to engage, or not engage, in instrumental aggression.*

We agree with the reviewer (and reviewer # 2) that plain and clear English should be used to describe our results and have since modified those terms.

However, the term "hyperaltruism", which is the main theme of our study, was originally proposed by a seminal paper (Crockett et al., 2014) and has since been widely adopted in related studies (Crockett et al., 2014, 2015, 2017; Volz et al., 2017; Zhan et al., 2020). The term "hyperaltruism" was introduced to emphasize the difference from altruism (Chen et al., 2024;

FeldmanHall et al., 2015; Hu et al., 2021; Hutcherson et al., 2015; Lockwood et al., 2017; Xiong et al., 2020). Hyperaltruism does not indicate extreme altruism. Instead, it simply reflects the fact that "we are more willing to sacrifice gains to spare others from harm than to spare ourselves from harm" (Volz et al., 2017). In other words, altruism refers to people's unselfish regard for or devotion to the welfare of others, and hyperaltruism concerns subject's own cost-benefit preference as the reference point and highlights the "additional" altruistic preference when considering other's welfare. For example, in the altruistic experimental design, altruism is characterized by the degree to which subjects take other people's welfare into account (left panel). However, in a typical hyperaltruism task design (right panel), hyperaltruistic preference is operationally defined as the difference ($\kappa_{other}$ - $\kappa_{self}$) between the degrees to which subjects value others' harm ($\kappa_{other}$) and their own harm ($\kappa_{self}$).

**Author response image 5.**




> *I found it surprising that a paradigm that entails deciding to hurt or not hurt someone else for personal benefit (whether acquiring a financial gain or avoiding a loss) would be described as measuring "altruism." Deciding to hurt someone for personal benefit is the definition of instrumental aggression. I did not see that in any of the studies was there a possibility of acting to benefit the other participant in any condition. Altruism is not equivalent to refraining from engaging in instrumental aggression. True altruism would be to accept shocks to the self for the other's benefit (e.g., money). The interpretation of this task as assessing instrumental aggression is supported by the fact that only the Instrumental Harm subscale of the OUS was associated with outcomes in the task, but not the Impartial Benevolence subscale. By contrast, the IB subscale is the one more consistently associated with altruism (e.g,. Kahane et al 2018; Amormino at al, 2022) I believe it is important for scientific accuracy for the paper, including the title, to be re-written to reflect what it is testing.*

Again, as we mentioned in the previous response, hyperaltruism is a term coined almost a decade ago and has since been widely adopted in the research field. We are afraid that switching such a term would be more likely to cause confusion (instead of clarity) among audience.

Also, from the utilitarian perspective, the gain or loss (or harm) occurred to someone else is aligned on the same dimension and there is no discontinuity between gains and losses. Therefore, taking actions to avoid someone else's loss can also be viewed as altruistic behavior, similar to choices increasing other's welfare (Liu et al., 2020).

> *Relatedly: in the introduction I believe it would be important to discuss the non-symmetry of moral obligations related to help/harm--we have obligations not to harm strangers but no obligation to help strangers. This is another reason I do not think the term "hyper altruism" is a good description for this task--given it is typically viewed as morally obligatory not to harm strangers, choosing not to harm them is not "hyper" altruistic (and again, I do not view it as obviously altruism at all).*

We agree with the reviewer's point that we have the moral obligations not to harm others but no obligation to help strangers (Liu et al., 2020). In fact, this is exactly what we argued in our manuscript: by switching the decision context from gains to losses, subjects were less likely to perceive the decisions as "harming others". Furthermore, after the administration of OXT, making decisions in both the gain and loss contexts were more perceived by subjects as harming others (Fig. 6A).

> *The framing of the role of OT also felt incomplete. In introducing the potential relevance of OT to behavior in this task, it is important to pull in evidence from non-human animals on origins of OT as a hormone selected for its role in maternal care and defense (including defensive aggression). The non-human animal literature regarding the effects of OT is on the whole much more robust and definitive than the human literature. The evidence is abundant that OT motivates the defensive care of offspring of all kinds. My read of the present OT findings is that they increase participants' willingness to refrain from shocking strangers even when incurring a loss (that is, in a context where the participant is weighing harm to themselves versus harm to the other). It will be important to explain why OT would be relevant to refraining from instrumental aggression, again, drawing on the non-human animal literature.*

We thank the reviewer's comments and agree that the current understanding of the link between our results of OT with animal literature can be at best described as vague and intriguing. Current literature on OT in animal research suggests that the nucleus accumbens (NAc) oxytocin might play the critical role in social cognition and reinforcing social interactions (Dölen et al., 2013; Dölen & Malenka, 2014; Insel, 2010). Though much insight has already been gained from animal studies, in humans, social interactions can take a variety of different forms, and the consociate recognition can also be rather dynamic. For example, male human participants with self-administered OT showed higher trust and cooperation towards in-group members but more defensive aggression towards out-group members (De Dreu et al., 2010). In another human study, participants administered with OT showed more coordinated out-group attack behavior, suggesting that OT might increase in-group efficiency at the cost of harming out-group members (Zhang et al., 2019). It is worth pointing out that in both experiments, the participant's group membership was artificially assigned, thus highlighting the context-dependent nature of OT effect in humans.

In our experiment, more complex and higher-level social cognitive processes such as moral framing and moral perception are involved, and OT seems to play an important role in affecting these processes. Therefore, we admit that this study, like the ones mentioned above, is rather hard to find non-human animal counterpart, unfortunately. Instead of relating OT to instrumental aggression, we aimed to provide a parsimonious framework to explain why the "hyperaltruism" disappeared in the loss condition, and, with the OT administration, reappeared in both the gain and loss conditions while also considering the effects of other relevant variables.

We concur with the reviewer's comments about the importance of animal research and have since added the following paragraph into the revised manuscript (Line 86~90) as well as in the discussion:

"Oxytocin has been shown to play a critical role in social interactions such as maternal attachment, pair bonding, consociate attachment and aggression in a variety of animal models[42,43]. Humans are endowed with higher cognitive and affective capacities and exhibit far more complex social cognitive patterns[44]."

> *Another important limitation is the use of only male participants in Study 2. This was not an essential exclusion. It should be clear throughout sections of the manuscript that this study's effects can be generalized only to male participants.*

We thank the reviewer's comments. Prior research has shown sex differences in oxytocin's effects (Fischer-Shofty et al., 2013; Hoge et al., 2014; Lynn et al., 2014; Ma et al., 2016; MacDonald, 2013). Furthermore, with the potential confounds of OT effect due to the menstrual cycles and potential pregnancy in female subjects, most human OT studies have only recruited male subjects (Berends et al., 2019; De Dreu et al., 2010; Fischer-Shofty et al., 2010; Ma et al., 2016; Zhang et al., 2019). We have modified our manuscript to emphasize that study 2 only recruited male subjects.

> *Recommendations:*
>
> *I believe the authors have provided an interesting and valuable dataset related to the willingness to engage in instrumental aggression - this is not the authors' aim, although also an important aim. Future researchers aiming to build on this paper would benefit from it being framed more accurately.*
>
> *Thus, I believe the paper must be reframed to accurately describe the nature of the task as assessing instrumental aggression. This is also an important goal, as well-designed laboratory models of instrumental aggression are somewhat lacking.*

Please see our response above that to have better connections with previous research, we believe that the term hyperaltruism might align better with the main theme for this study.

> *The research literature on other aggression tasks should also be brought in, as I believe these are more relevant to the present study than research studies on altruism that are primarily donation-type tasks. It should be added to the limitations of how different aggression in a laboratory task such as this one is from real-world immoral forms of aggression. Arguably, aggression in a laboratory task in which all participants are taking part voluntarily under a defined set of rules, and in which aggression constrained by rules is mutual, is similar to aggression in sports, which is not considered immoral. Whether responses in this task would generalize to immoral forms of aggression cannot be determined without linking responses in the task to some real-world outcome.*

We agree with the reviewer that "aggression in a lab task …. is similar to aggression in sports". Our starting point was to investigate the boundary conditions for the hyperaltruism (though we don't deny that there is an aggression component in hyperaltruism, given the experiment design we used). In other words, the dependent variable we were interested in was the difference between "other" and "self" aggression, not the aggression itself. Our results showed that by switching the decision context from the monetary gain environment to the loss condition, human participants were willing to bear similar amounts of monetary loss to spare others and themselves from harm. That is, hyperaltruism disappeared in the loss condition. We interpreted this result as the loss condition prompted subjects to adopt a different moral framework (help vs. harm, Fig. 6A) and subjects were less influenced by their instrumental harm personality trait due to the change of moral framework (Fig. 3C). In the following study (study 2), we further tested this hypothesis and verified that the administration of OT indeed increased subjects' perception of the task as harming others for

both gain and loss conditions (Fig. 6A), and such moral perception mediated the relationship between subject's personality traits (instrumental harm) and their relative harm sensitivities (the difference of aggression between the other- and self-conditions). We believe the moral perception framework and that OT directly modulates moral perception better account for subjects' context-dependent choices than hypothesizing OT's context-dependent modulation effects on aggression.

> *The language should also be toned down--the use of phrases like "hyper altruism" (without independent evidence to support that designation) and "obliterate" rather than "reduce" or "eliminate" are overly hyperbolic.*

We have changed terms such as "obliterate" and "eliminate" to plain English, as the reviewer suggested.

Reference


Abu-Akel, A., Palgi, S., Klein, E., Decety, J., & Shamay-Tsoory, S. (2015). Oxytocin increases empathy to pain when adopting the other- but not the self-perspective. Social Neuroscience, 10(1), 7–15.

Barchi-Ferreira, A., & Osório, F. (2021). Associations between oxytocin and empathy in humans: A systematic literature review. Psychoneuroendocrinology, 129, 105268.

Berends, Y. R., Tulen, J. H. M., Wierdsma, A. I., van Pelt, J., Feldman, R., Zagoory-Sharon, O., de Rijke, Y. B., Kushner, S. A., & van Marle, H. J. C. (2019). Intranasal administration of oxytocin decreases task-related aggressive responses in healthy young males. Psychoneuroendocrinology, 106, 147–154.

Chen, J., Putkinen, V., Seppälä, K., Hirvonen, J., Ioumpa, K., Gazzola, V., Keysers, C., & Nummenmaa, L. (2024). Endogenous opioid receptor system mediates costly altruism in the human brain. Communications Biology, 7(1), 1–11.

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

Crockett, M. J., Siegel, J. Z., Kurth-Nelson, Z., Dayan, P., & Dolan, R. J. (2017). Moral transgressions corrupt neural representations of value. Nature Neuroscience, 20(6), 879–885.

Crockett, M. J., Siegel, J. Z., Kurth-Nelson, Z., Ousdal, O. T., Story, G., Frieband, C., Grosse-Rueskamp, J. M., Dayan, P., & Dolan, R. J. (2015). Dissociable Effects of Serotonin and Dopamine on the Valuation of Harm in Moral Decision Making. Current Biology, 25(14), 1852–1859.

De Dreu, C. K. W., Greer, L. L., Handgraaf, M. J. J., Shalvi, S., Van Kleef, G. A., Baas, M., Ten Velden, F. S., Van Dijk, E., & Feith, S. W. W. (2010). The Neuropeptide Oxytocin Regulates Parochial Altruism in Intergroup Conflict Among Humans. Science, 328(5984), 1408–1411.

De Dreu, C. K. W., Gross, J., Fariña, A., & Ma, Y. (2020). Group Cooperation, Carrying-Capacity Stress, and Intergroup Conflict. Trends in Cognitive Sciences, 24(9), 760–776.

Dölen, G., Darvishzadeh, A., Huang, K. W., & Malenka, R. C. (2013). Social reward requires coordinated activity of nucleus accumbens oxytocin and serotonin. Nature, 501(7466), 179–184.

Dölen, G., & Malenka, R. C. (2014). The Emerging Role of Nucleus Accumbens Oxytocin in Social Cognition. Biological Psychiatry, 76(5), 354–355.

Evans, S., Shergill, S. S., & Averbeck, B. B. (2010). Oxytocin Decreases Aversion to Angry Faces in an Associative Learning Task. Neuropsychopharmacology, 35(13), 2502–2509.

Fehr, E., & Schmidt, K. M. (1999). A Theory of Fairness, Competition, and Cooperation*. The Quarterly Journal of Economics, 114(3), 817–868.

FeldmanHall, O., Dalgleish, T., Evans, D., & Mobbs, D. (2015). Empathic concern drives costly altruism. Neuroimage, 105, 347–356.

Fischer-Shofty, M., Levkovitz, Y., & Shamay-Tsoory, S. G. (2013). Oxytocin facilitates accurate perception of competition in men and kinship in women. Social Cognitive and Affective Neuroscience, 8(3), 313–317.

Fischer-Shofty, M., Shamay-Tsoory, S. G., Harari, H., & Levkovitz, Y. (2010). The effect of intranasal administration of oxytocin on fear recognition. Neuropsychologia, 48(1), 179–184.

Glenn, A. L., Koleva, S., Iyer, R., Graham, J., & Ditto, P. H. (2010). Moral identity in psychopathy. Judgment and Decision Making, 5(7), 497–505.

Hoge, E. A., Anderson, E., Lawson, E. A., Bui, E., Fischer, L. E., Khadge, S. D., Barrett, L. F., & Simon, N. M. (2014). Gender moderates the effect of oxytocin on social judgments. Human Psychopharmacology: Clinical and Experimental, 29(3), 299–304.

Hu, J., Hu, Y., Li, Y., & Zhou, X. (2021). Computational and Neurobiological Substrates of Cost-Benefit Integration in Altruistic Helping Decision. Journal of Neuroscience, 41(15), 3545–3561.

Hutcherson, C. A., Bushong, B., & Rangel, A. (2015). A Neurocomputational Model of Altruistic Choice and Its Implications. Neuron, 87(2), 451–462.

Insel, T. R. (2010). The Challenge of Translation in Social Neuroscience: A Review of Oxytocin, Vasopressin, and Affiliative Behavior. Neuron, 65(6), 768–779.

Kahane, G., Everett, J. A. C., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. Psychological Review, 125(2), 131–164.

Kahane, G., Everett, J. A. C., Earp, B. D., Farias, M., & Savulescu, J. (2015). 'Utilitarian' judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. Cognition, 134, 193–209.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. Econometrica, 47(2), 263.

Kapetaniou, G. E., Reinhard, M. A., Christian, P., Jobst, A., Tobler, P. N., Padberg, F., & Soutschek, A. (2021). The role of oxytocin in delay of gratification and flexibility in non-social decision making. eLife, 10, e61844.

Kirsch, P., Esslinger, C., Chen, Q., Mier, D., Lis, S., Siddhanti, S., Gruppe, H., Mattay, V. S., Gallhofer, B., & Meyer-Lindenberg, A. (2005). Oxytocin Modulates Neural Circuitry for Social Cognition and Fear in Humans. The Journal of Neuroscience, 25(49), 11489–11493.

Liu, J., Gu, R., Liao, C., Lu, J., Fang, Y., Xu, P., Luo, Y., & Cui, F. (2020). The Neural Mechanism of the Social Framing Effect: Evidence from fMRI and tDCS Studies. The Journal of Neuroscience, 40(18), 3646–3656.

Liu, Y., Li, L., Zheng, L., & Guo, X. (2017). Punish the Perpetrator or Compensate the Victim? Gain vs. Loss Context Modulate Third-Party Altruistic Behaviors. Frontiers in Psychology, 8, 2066.

Lockwood, P. L., Hamonet, M., Zhang, S. H., Ratnavel, A., Salmony, F. U., Husain, M., & Maj, A. (2017). Prosocial apathy for helping others when effort is required. Nature Human Behaviour, 1(7), 131–131.

Losecaat Vermeer, A. B., Boksem, M. A. S., & Sanfey, A. G. (2020). Third-party decision-making under risk as a function of prior gains and losses. Journal of Economic Psychology, 77, 102206.

Lynn, S. K., Hoge, E. A., Fischer, L. E., Barrett, L. F., & Simon, N. M. (2014). Gender differences in oxytocin-associated disruption of decision bias during emotion perception. Psychiatry Research, 219(1), 198–203.

Ma, Y., Liu, Y., Rand, D. G., Heatherton, T. F., & Han, S. (2015). Opposing Oxytocin Effects on Intergroup Cooperative Behavior in Intuitive and Reflective Minds. Neuropsychopharmacology, 40(10), 2379–2387.

Ma, Y., Shamay-Tsoory, S., Han, S., & Zink, C. F. (2016). Oxytocin and Social Adaptation: Insights from Neuroimaging Studies of Healthy and Clinical Populations. Trends in Cognitive Sciences, 20(2), 133–145.

MacDonald, K. S. (2013). Sex, Receptors, and Attachment: A Review of Individual Factors Influencing Response to Oxytocin. Frontiers in Neuroscience, 6. 194.

Markiewicz, Ł., & Czupryna, M. (2018). Cheating: One Common Morality for Gain and Losses, but Two Components of Morality Itself. Journal of Behavior Decision Making. 33(2), 166-179.

Pachur, T., Schulte-Mecklenbeck, M., Murphy, R. O., & Hertwig, R. (2018). Prospect theory reflects selective allocation of attention. Journal of Experimental Psychology: General, 147(2), 147–169.

Radke, S., Roelofs, K., & De Bruijn, E. R. A. (2013). Acting on Anger: Social Anxiety Modulates Approach-Avoidance Tendencies After Oxytocin Administration. Psychological Science, 24(8), 1573–1578.

Saez, I., Zhu, L., Set, E., Kayser, A., & Hsu, M. (2015). Dopamine modulates egalitarian behavior in humans. Current Biology, 25(7), 912–919.

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), 3534.

Tom, S. M., Fox, C. R., Trepel, C., & Poldrack, R. A. (2007). The neural basis of loss aversion in decision-making under risk. Science, 315(5811), 515–518.

Usher, M., & McClelland, J. L. (2004). Loss Aversion and Inhibition in Dynamical Models of Multialternative Choice. Psychological Review, 111(3), 757–769.

Volz, L. J., Welborn, B. L., Gobel, M. S., Gazzaniga, M. S., & Grafton, S. T. (2017). Harm to self outweighs benefit to others in moral decision making. Proceedings of the National Academy of Sciences of the United States of America, 114(30), 7963–7968.

Wu, Q., Mao, J., & Li, J. (2020). Oxytocin alters the effect of payoff but not base rate in emotion perception. Psychoneuroendocrinology, 114, 104608.

Wu, S., Cai, W., & Jin, S. (2018). Gain or non-loss: The message matching effect of regulatory focus on moral judgements of other-orientation lies. International Journal of Psychology, 53(3), 223-227.

Xiong, W., Gao, X., He, Z., Yu, H., Liu, H., & Zhou, X. (2020). Affective evaluation of others' altruistic decisions under risk and ambiguity. Neuroimage, 218, 116996.

Yechiam, E., & Hochman, G. (2013). Losses as modulators of attention: Review and analysis of the unique effects of losses over gains. Psychological Bulletin, 139(2), 497–518.

Zhan, Y., Xiao, X., Tan, Q., Li, J., Fan, W., Chen, J., & Zhong, Y. (2020). Neural correlations of the influence of self-relevance on moral decision-making involving a trade-off between harm and reward. Psychophysiology, 57(9), e13590.

Zhang, H., Gross, J., De Dreu, C., & Ma, Y. (2019). Oxytocin promotes coordinated out-group attack during intergroup conflict in humans. eLife, 8, e40698.

https://doi.org/10.7554/eLife.102756.2.sa0