eLife

Reviewed Preprint
v1 • November 14, 2024
Not revised

Neuroscience

Oxytocin salvages context-specific hyperaltruistic preference through moral framing

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 study presents a **valuable** finding that altruistic tendency during moral decision-making is context-dependent (present in the gain domain but absent in the loss domain) and its absence in the loss domain can be restored by the neuropeptide oxytocin. However, the evidence supporting this claim is somewhat **incomplete** and would benefit from better overall framing and clarity on its approaches. Overall, this study will be of interest to social scientists and neuroscientists who work on moral decision-making and oxytocin.

https://doi.org/10.7554/eLife.102756.1.sa3

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 a 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 salvage 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 nullified 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 were willing to pay more to reduce other's pain than their own pain, yielding a "hyperaltruistic" preference in moral dilemmas[4–7]. However, the mechanistic account of the hyperaltruistic phenomenon remains unknown, which hinders the identification of 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 preference. 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[8–10].

Classic moral dilemmas often involve the tradeoff between personal material wellbeing and the adherence of social norm (or moral principle)[11,12]. 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[13–15]. Other research, instead, found that it was more common for people to help others from losing money than to help them to gain money [16,17]. 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 pit against the physical pain experienced by others or 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 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 loss aversion or highlighted vigilance in the face of potential loss[18–23]. Indeed, elevated attention to losses can alternate subjects' sensitivities to the general reinforcement structure and have distinct effects on their arousal and performance[22]. Additionally, prospective losses can elicit negative emotion[24,25], 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, which may yield opposing effects on hyperaltruism [26–30]. Finally, how the moral dilemma is framed can be a critical factor influencing moral decision-making[31–33]. 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 for the harm frame compared to the help frame [20,34]. 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, modulates a variety of emotional, cognitive and social behaviors in humans through distributed receptors in various brain areas[35,36]. 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) [37–44]. However, the evidence for how oxytocin influences hyperaltruism remains scarce[45]. This may be partly due to the fact that the prosocial effect of oxytocin seems to be both personality trait and decision context dependent[8,46–49]. For example, recent studies have shown that oxytocin both promotes in-group cooperation and defensive aggression toward competing out-group members [50–52]. 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[38,53–55]. 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 (monetary valence, gain versus loss) 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 of higher monetary gains (or smaller monetary loss in the loss context) were associated with more painful electric shocks. We found that in line with results reported in previous studies, subjects showed clear hyperaltruistic preference in the gain context. However, shifting 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 hyperaltruism in the gain context. However, oxytocin salvaged subjects' hyperaltruistic preferences in the loss context. Importantly, the oxytocin administration prompted subjects to more likely perceive the task structure as harming others, which in turn mediated their hyperaltruistic preferences.

Results

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 with regard to different shock recipients. A larger choice difference between other- and self-recipients would indicate a stronger hyperaltruistic tendencies in both the gain and loss contexts. The main effects of the shock recipient (self vs. other) and context (gain vs. loss) were both significant (recipient: $F_{1,79} = 6.625$, $P = 0.012$, $\eta^2 = 0.077$; context: $F_{1,79} = 8.379$, $P = 0.005$, $\eta^2 = 0.096$), 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 × context 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 eradicated in the loss context ($F_{1,79} = 0.650$, $P = 0.423$, $\eta^2 = 0.008$). We also adopted an influential modeling approach where subjects' behavior could be characterized by the harm (shock) aversion parameter κ depicting the relative weight subjects assigned to Δm and Δs, the objective differences in money and shocks between the more and less painful options, respectively (See Methods for details) [4–6]. Consistent with previous literatures, we found that the harm aversion parameter κ for 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 conditions disappeared in the loss context($\kappa_{other} - \kappa_{self} > 0$ in the gain context: t_{79}=3.869, $P < 0.001$; κ_{other} and κ_{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 individual's sensitivities to money (Δm) and electric shock (Δs). Subjects' choices

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.

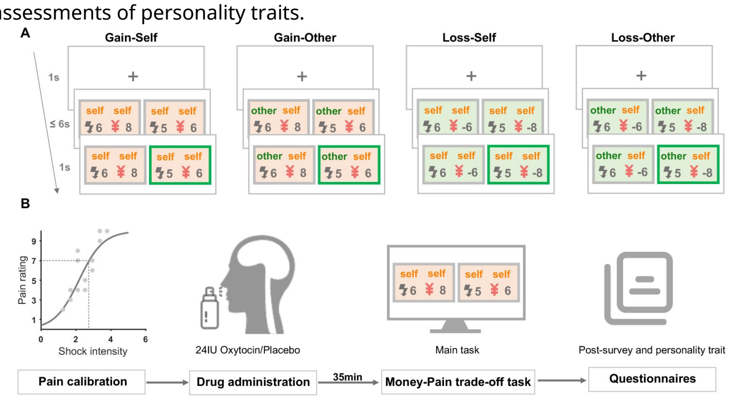
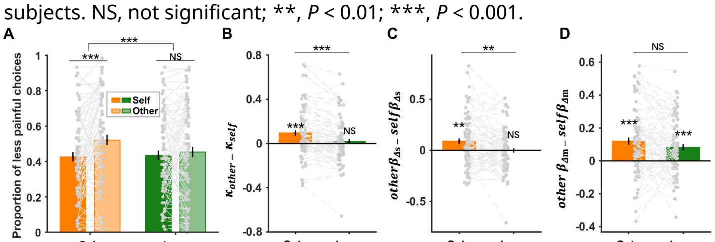


Fig. 2.

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; **, $P < 0.01$; ***, $P < 0.001$.



were regressed against Δm and Δs (differences in money and shock magnitude between the more and less painful options, see Methods for further details). This analysis yielded significant relative harm sensitivity (difference of regression coefficients of Δs in the other- and self-condition: (*other* $\beta_{\Delta s}$- *self* $\beta_{\Delta s}$)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 the decision context effect was also significant (t_{79} = 3.630, P = 0.001; **Fig. 2C**). On the contrary, decision context did not have an effect on subjects' relative sensitivities (*other* $\beta_{\Delta m}$- *self* $\beta_{\Delta m}$) towards monetary (Δm) difference (t_{79} = 1.480, P = 0.143), though the relative money sensitivities were significant in both contexts (gain: t_{79} = 5.241, P < 0.001 ; loss: 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[56,57]. 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 is related to both IH and IB across decision contexts. As expected, moral preference, defined as κ_{other} - κ_{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). Interestingly, by examining the separate contributions of Δm and Δs to moral preference, we found that the decision context modulated the relationship between IH and subjects' relative harm sensitivities (gain context: β = -0.053 ± 0.016, t_{152} = -3.224, P = 0.002; loss context: β = -0.007 ± 0.016, t_{152} = -0.453, P = 0.651; context × IH: β = 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: β = 0.019 ± 0.023, t_{152} =0.804, P = 0.423; loss context: β = 0.010 ± 0.023, t_{152} =0.425, P = 0.672; context × IB: β=-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: β = -0.022 ±0.015, t_{152}=-1.476, P = 0.142; loss context: β = -0.023 ±0.015, t_{152}= -1.523, P = 0.130; context × IH: β = -0.001± 0.021, t_{152} = -0.034, P = 0.973; **Fig. 3D**), nor with IB (gain context: β = -0.008±0.021, t_{152} = -0.369, P = 0.713 ; loss context: β = -0.009 ± 0.021, t_{152} = -0.414, P = 0.679 ; context × IB: β = -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 instrumental harm (IH) implies that one person's benefit contingent on the pain inflicted on another person is internally deemed immoral, no matter whether the pain is applied to others or subjects themselves [1]. Hyperaltruistic preference, therefore, is likely to be associated with subjects' IH dispositions.

Oxytocin salvaged 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: context × recipient interaction: $F_{1,45}$ = 9.103, P = 0.004, η^2 =

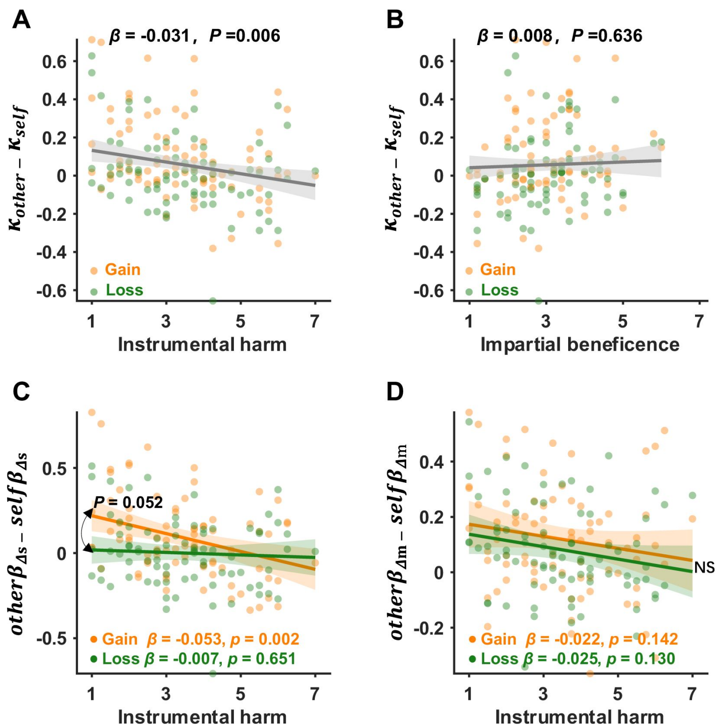


Fig. 3.

Individual difference in moral preferences.

(**A-B**) Hyperaltruism ($\kappa_{other} - \kappa_{self}$) 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. NS, not significant.

0.168; simple effect: gain context $F_{1,45}$ = 9.356, P = 0.004, η^2 = 0.172; loss context $F_{1,45}$ = 0.005, P = 0.946, η^2 < 0.001; **Fig. 4A**). The main effect of treatment (placebo vs. oxytocin) was significant ($F_{1,45}$ = 41.961, P < 0.001, η^2 = 0.483), indicating that the oxytocin treatment prompted subjects to select the more painful option more often. Interestingly, the treatment × recipient × context interaction effect was also significant ($F_{1,45}$ = 6.309, P = 0.016, η^2 = 0.123), suggesting that the recipient × context interaction might be different due to the placebo and oxytocin treatments. Indeed, with the administration of oxytocin, the hyperaltruistic preference was salvaged in the loss context, without affecting the hyperaltruistic pattern in the gain context (oxytocin: context × recipient: $F_{1,45}$ = 0.480, P = 0.492, η^2 = 0.011; simple effect: gain context $F_{1,45}$ = 25.364, P < 0.001, η^2 = 0.360; loss context: $F_{1,45}$ = 24.408, P <0.001, η^2 = 0.352; **Fig. 4A**). These results together highlight that oxytocin might play an important context-dependent modulatory role in restoring hyperaltruistic moral preference. Our modeling analysis revealed similar results: there was a significant treatment × context interaction effect ($F_{1,45}$ = 6.349, P = 0.015, η^2 = 0.124) on subjects' moral preference (defined as κ_{other} - κ_{self}, **Fig. 4B**). With the administration of oxytocin, however, the hyperaltruistic preference lost in the loss context (placebo: $F_{1,45}$ = 1.295, P = 0.261, η^2 =0.028) was successfully rescued (oxytocin: $F_{1,45}$=17.990, P < 0.001, η^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 treatment

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 to link subjects' choices with continuous independent variables including Δm, Δs and categorical variables including treatment (placebo vs. oxytocin), harm/shock recipient (self vs. other) and decision context (gain vs. loss; see methods for details). This regression analysis yielded a significant Δs × treatment × recipient × context interaction effect (β = 0.125 ± 0.053, P = 0.018, 95% CI = [0.022, 0.228]; **Supplementary Fig. 6**), 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, η^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 conditions (simple effect: $F_{1,45}$ = 0.131, P = 0.719, η^2 = 0.001; **Fig. 5A**). It is worth noting that oxytocin did not alter the gain/loss relative money sensitivity difference (context × treatment interaction: $F_{1,45}$ = 1.933, P = 0.171, η^2 = 0.041 ; **Fig. 5B**), which corresponded to a non-significant Δm × treatment × recipient × context interaction effect (β = -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: β = -0.071 ± 0.024, t_{84} = -3.005, P = 0.004 ; loss context: β = 0.006 ± 0.024, t_{84} = -0.243, P = 0.809; context × IH: β = 0.077 ± 0.033, t_{84} = 2.297, P = 0.024; **Fig. 5C**). However, under the treatment of oxytocin, the context × IH interaction became non-significant (β = -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: β = -0.037 ± 0.015, t_{84} = -2.458, P = 0.016; loss context: β = -0.042 ± 0.015, t_{84} = -2.843, P = 0.006; **Fig. 5D**) (also see **Supplementary Fig. 7** and **Table S2**). 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.

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 between other- and self- recipients) 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; *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

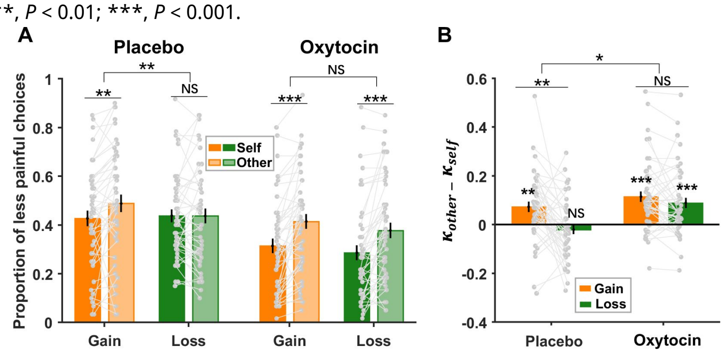
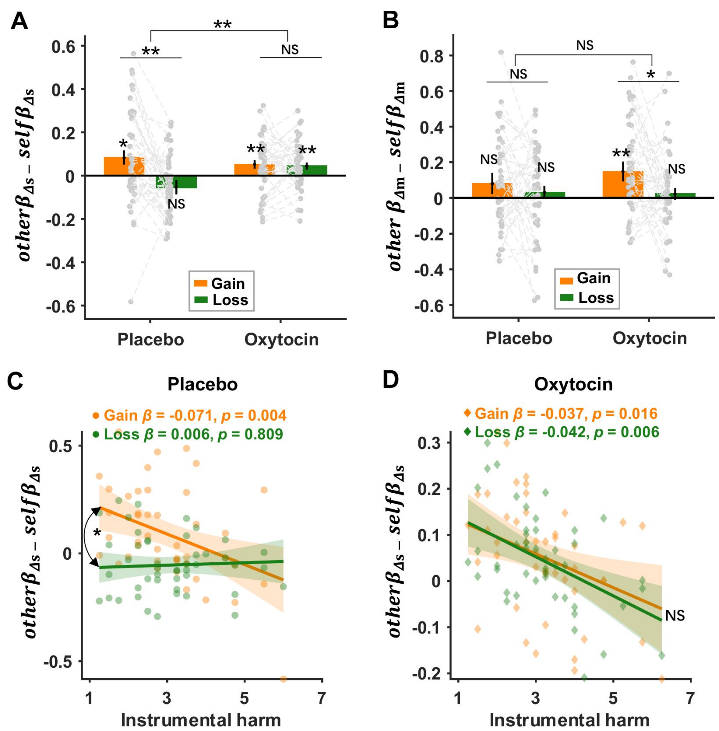


Fig. 5.

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. NS, not significant; *, $P < 0.05$; **, $P < 0.01$



Oxytocin obliterates 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 a “harming” condition (inflicting pain on the receiver to increase monetary gain/avoid monetary loss) or as a “helping” condition (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 (χ^2 = 2.827, P = 0.028, Bonferroni correction), 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 (χ^2 = 2.665, P = 0.046, Bonferroni correction) and removed contextual disparities in harm perception (context × treatment interaction: χ^2 = 7.410, P = 0.006), 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. 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 (perceived harm report) 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 rescued participants’ hyperaltruistic preference by restoring the mediation effect 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 personal gain was traded against moral norms such as no deception or no infliction of suffering on others[1,11,58]. 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 pay more 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

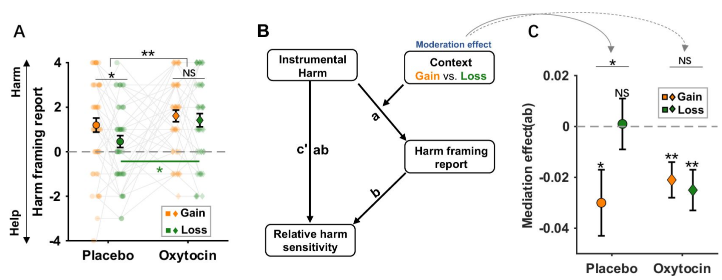


Fig. 6.

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.

rescued the hyperaltruistic preference in the loss context by biasing subjects to perceive the decision context to be more harming others for self-interest. Finally, moral framing, or how subjects perceive the decision context as helping or harming others, mediated the association between 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, similar to the well-known trolley dilemma, might be better suited to study the specific relationship between IH attitude and moral preference[11]. 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 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 preference.

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[18]. 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 (κ_{other} - κ_{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[20,34]. 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 leaded 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 taps 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[59], which proposes that oxytocin attenuates people's avoidance of negative social or nonsocial stimuli[41,59–61]. Other studies conducted on both animals and humans suggest that oxytocin increases pain tolerance and attenuates acute pain experience [62–64]. 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 [50–52]. 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. 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. 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[4–7,65]. Based on the small to medium effect size (Cohen's d = 0.2), 75 subjects were needed to detect a significant effect ($\alpha = 0.05$, $\beta = 0.9$, 2 (gain vs. loss) × 2 (self vs. other shock recipient) within-subject design 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 therefore was preregistered (*https://osf.io/fhwa9*) via the Open Science Framework. Again, based on the effect size (Cohen's d = 0.21) of oxytocin effect on social behaviors reported previously a minimum of 44 subjects($\alpha = 0.05$, $\beta = 0.9$, 2×2×2 within-subject design ANOVA) was required for the within-subject design[66]. We recruited 46 male subjects only (mean age = 21.74 ± 2.33 years) in study 2, similar to most previous studies to avoid sex-related potential confounds in oxytocin effects on social cognition and behavior[67,68]. 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

Every time, two subjects (a pair) arrived with a 5-minute interval and entered into separate testing rooms without them 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[4,16]. 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 time 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 [4,69]. 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[47,49]. For each treatment, the oxytocin or the 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, and 3 indicating neutrality. 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: t_{45} = 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 task 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. The experimental task was coded with Psychtoolbox (version 3.0.17) in Matlab (2018b). A more painful option is always associated with more monetary gain (or less monetary loss). We first created 60 gain trials using the procedure reported in earlier research [4]. The monetary gain ranged from ¥0.2 to ¥20 (¥0.2 increment) and the electric shocks ranged from 1 to 20 shocks for the more painful option, whereas the gain ranged from ¥0.2 to ¥19.8 and the shocks ranged from 1 to 19 for the less painful option. After the monetary gain and shock number was randomly determined for the less painful option, the monetary gain and shock amounts were constructed by randomly drawing a shock increment integer Δs (from 1 to 19) such that the more painful option shock delivery did not exceed 20 ($S_m \leq$ 20). Similarly, a random gain increment Δm was drawn from a uniform distribution (from ¥0.2 to ¥19.8) such that the more painful option monetary gain did not exceed 20 ($M_m \leq$ ¥20). 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 for all the 4 decision-context and shock- recipient combinations (gain-self, gain-other, loss-self and loss other) each thus yielding a total of 240 trials delivered across two blocks (gain and loss blocks). We followed Crockett et al. (2017)'s method to optimize the trial set[6]. Across trials, Δs and Δm was uncorrelated (r = -0.0012, P = 0.986). We randomly generated each subject's trial set using the above method. 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 across each loss-self and loss-other trials). 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). Subjects received a ¥60 showing up fee in study 1 and ¥220 in study 2 in addition to their performance-based compensation. In total, the average renumeration for each subject was ¥97 in study 1 and ¥318 in study 2.

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) [70]. 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**). Multi- linear regression analysis was performed to examine the relationships between IB, IH and 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 ≤ κ ≤ 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 $\kappa_{gain-self}$, $\kappa_{gain-other}$, $\kappa_{loss-self}$, and $\kappa_{loss-other}$ respectively. Δm and Δs denote 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).

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

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 [71]:

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

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:

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

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 significant interaction effect (β_3^{M}, β_4^{Y}, β_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[72].

We utilized the bootstrapping procedure with 5000 bootstrap resamples to analyzed the data[73,74]. This analysis was conducted separately for the placebo and oxytocin sessions. Our results showed that the interaction coefficients (placebo: $\beta_4^{Y} = 0.036 \pm 0.029$, $t_{84} = 1.267$, $P = 0.209$; $\beta_5^{Y} = 0.0242 \pm 0.017$, $t_{84} = 1.415$, $P = 0.161$; Oxytocin: $\beta_4^{Y} = 0.006 \pm 0.010$, $t_{84} = 0.319$, $P = 0.751$; $\beta_5^{Y} = 0.003 \pm 0.012$, $t_{84} = 0.246$, $P = 0.806$) in model Y were not significant, while β_4^{M} was significant in the placebo session ($\beta_3^{M} = 0.784 \pm 0.343$, $t_{86} = 2.288$, $P = 0.025$) but not the oxytocin session ($\beta_3^{M} = -0.137 \pm 0.291$, $t_{86} = -0.472$, $P = 0.638$, 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', 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.

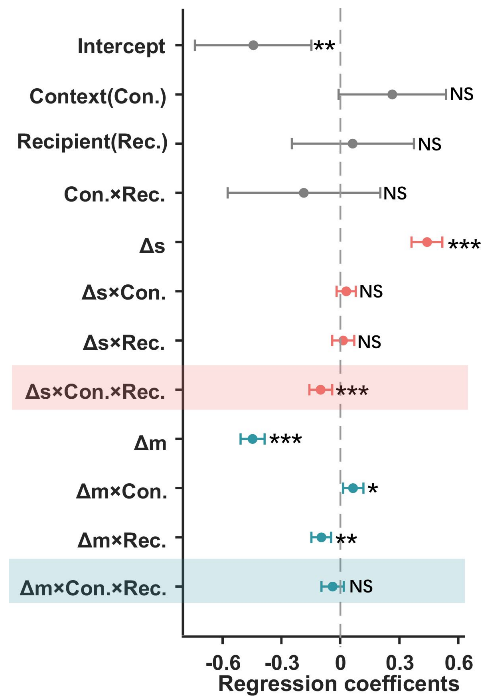
Data availability

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

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.

Supplementary figures and tables



Supplementary Fig. 1.

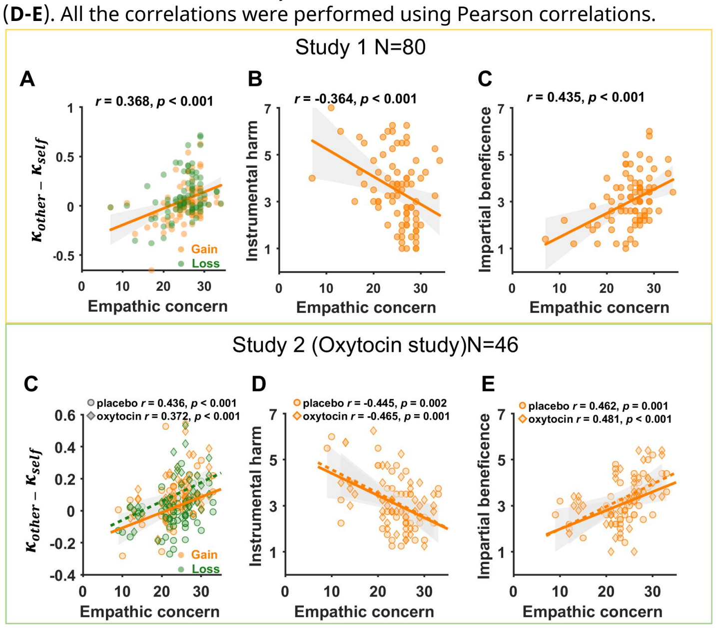
Mixed-effect logistic regression analysis results for experiment 1(regression model 1).

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.

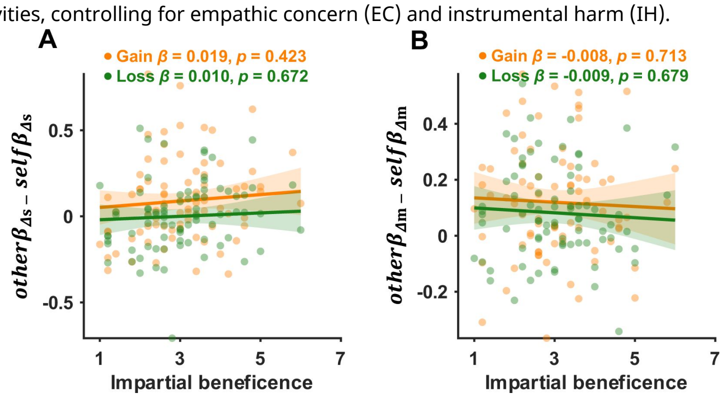
Correlation between EC and hyperaltruism ($\kappa_{other} - \kappa_{self}$), instrumental harm (IH), impartial beneficence (IB) both in the Study1(**A-C**) and Study 2 (**D-E**). All the correlations were performed using Pearson correlations.



Supplementary Fig. 3.

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

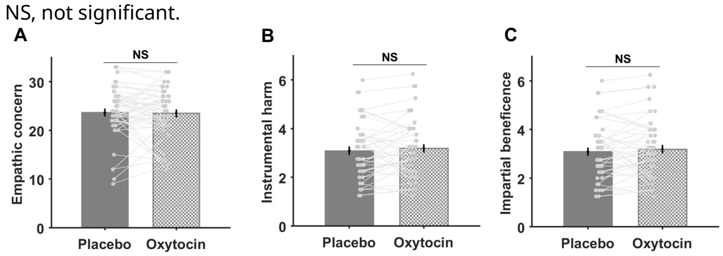
IB showed no significant correlation with subjects' relative harm sensitivities ($other\beta_{\Delta s} - self\beta_{\Delta s}$) **(A)**, or relative monetary sensitivities ($other\beta_{\Delta m} - self\beta_{\Delta m}$) **(B)**. The regression coefficients showed the association between IB and relative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



Supplementary Fig. 4.

Oxytocin and personality traits in study 2.

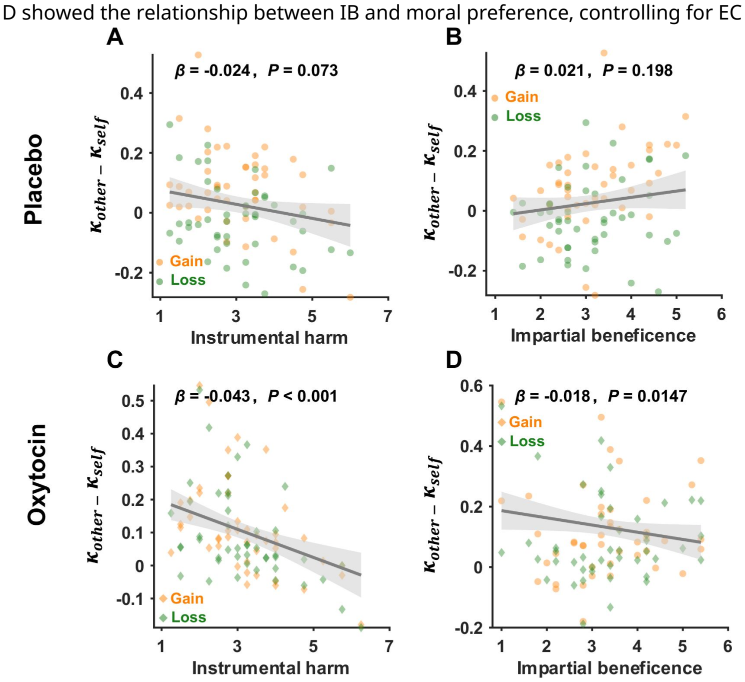
(**A-C**) Oxytocin did not affect subjects' ratings on empathic concern (EC), instrumental harm (IH) or impartial beneficence (IB). Error bars represent SE. NS, not significant.

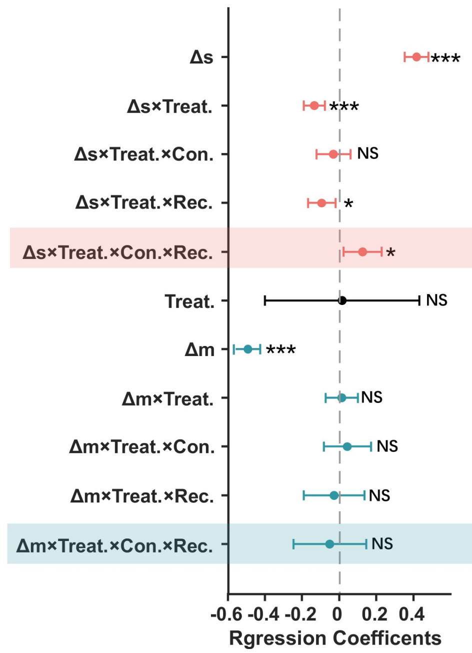


Supplementary Fig. 5.

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 Δm × Treat.× Con.× Rec. interaction and Δm × Treat.× Con.× Rec. interaction was further explained in **Fig. 5A** & B. 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.

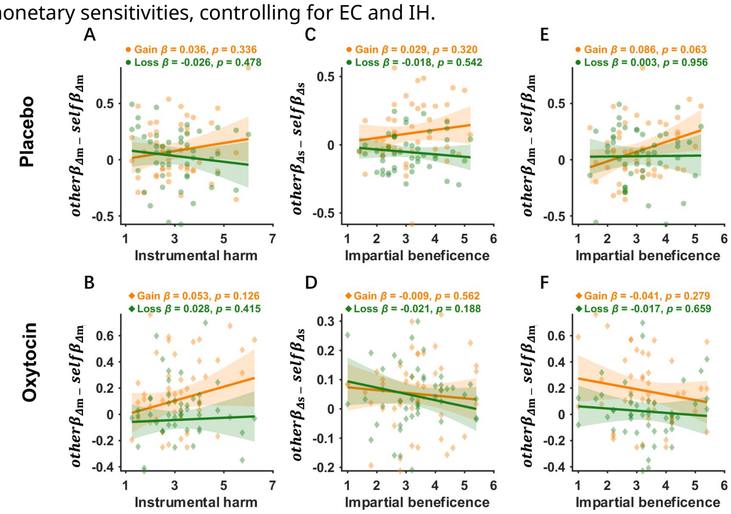


Table S1.

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). † 0.1, * $P < 0.05$, ** $P < 0.01$ and *** $P < 0.001$.

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** *Psychol Rev* **125**:131–164 https://doi.org/10.1037/rev0000093

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 https://doi.org/10.1016/j.cognition.2011.05.010

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 https://doi.org/10.1017/S1930297500001662

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 https://doi.org/10.1073/pnas.1408988111

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 https://doi.org/10.1016/j.cub.2015.05.021

6. Crockett M. J., Siegel J. Z., Kurth-Nelson Z., Dayan P., Dolan R. J (2017) **Moral Threshold Model is Universal? Initial Evidence from China's Collectivist Culture** *Psychol Res Behav Manag* **20**:1847–1855 https://doi.org/10.2147/PRBM.S333364

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 https://doi.org/10.1073/pnas.1706693114

8. 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 https://doi.org/10.1016/j.tics.2011.05.002

9. Crespi B. J (2016) **Oxytocin, testosterone, and human social cognition** *Biological Reviews* **91**:390–408 https://doi.org/10.1111/brv.12175

10. Young L. J (2015) **Oxytocin, Social Cognition and Psychiatry** *Neuropsychopharmacology* **40**:243–244 https://doi.org/10.1038/npp.2014.186

11. 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 https://doi.org/10.1126/science.1062872

12. 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 https://doi.org/10.1016/j.neuron.2004.09.027

13. 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 https://doi.org/10.1111/jasp.12852

14. 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 https://doi.org/10.1002/ejsp.2896

15. 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** https://doi.org/10.1016/j.joep.2022.102487

16. 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** *R Soc Open Sci* **2** https://doi.org/10.1098/rsos.150546

17. Li 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 https://doi.org/10.1016/j.ijpsycho.2020.02.003

18. Kahneman D., Tversky A (1979) **Prospect Theory: An Analysis of Decision under Risk** *Econometrica* **47**:263–292 https://doi.org/10.2307/1914185

19. 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 https://doi.org/10.1126/science.1134239

20. 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* https://doi.org/10.1523/jneurosci.1385-19.2020

21. Usher M., McClelland J. L (2004) **Loss Aversion and Inhibition in Dynamical Models of Multialternative Choice** *Psychological Review* **111**:757–769 https://doi.org/10.1037/0033-295X.111.3.757

22. 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 https://doi.org/10.1037/a0029383

23. 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 https://doi.org/10.1037/xge0000406

24. De Martino B., Camerer C. F., Adolphs R. (2010) **Amygdala damage eliminates monetary loss aversion** *Proc Natl Acad Sci U S A* **107**:3788–3792 https://doi.org/10.1073/pnas.0910230107

25. 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 https://doi.org/10.1093/SCAN/NSV139

26. 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 https://doi.org/10.1093/scan/nsw173

27. Carlino E., Frisaldi E., Benedetti F (2014) **Pain and the context** *Nature Reviews Rheumatology* **10**:348–355 https://doi.org/10.1038/nrrheum.2014.17

28. 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** *Soc Neurosci* https://doi.org/10.1080/17470919.2021.1953135

29. Pozo M. *et al.* (2023) **Negative energy balance hinders prosocial helping behavior** *Proceedings of the National Academy of Sciences* **120** https://doi.org/10.1073/pnas.2218142120

30. Ma N., Li N., He X. S., Sun D. L., Zhang X., Zhang D. R (2009) **Rejection of unfair offers can be driven by negative emotions, evidence from modified ultimatum games with anonymity** *PLoS One* **61** https://doi.org/10.1371/journal.pone.0039619

31. McDonald K., Graves R., Yin S., Weese T., Sinnott-Armstrong W (2021) **Valence framing effects on moral judgments: A meta-analysis** *Cognition* **212** https://doi.org/10.1016/j.cognition.2021.104703

32. Cao F., Zhang J., Song L., Wang S., Miao D., Peng J (2017) **Framing Effect in the Trolley Problem and Footbridge Dilemma** *Psychol Rep* **120**:88–101 https://doi.org/10.1177/0033294116685866

33. 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 https://doi.org/10.1016/j.jesp.2017.06.012

34. 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** *Neurosci Bull* https://doi.org/10.1007/s12264-021-00808-3

35. 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 https://doi.org/10.1523/JNEUROSCI.4572-07.2008

36. 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 https://doi.org/10.1038/nrn3044

37. Zak P. J., Stanton A. A., Ahmadi S (2007) **Oxytocin increases generosity in humans** *PLoS One* **2** https://doi.org/10.1371/journal.pone.0001128

38. Barchi-Ferreira A., Osório F (2021) **Associations between oxytocin and empathy in humans: A systematic literature review** *Psychoneuroendocrinology* **129** https://doi.org/10.1016/j.psyneuen.2021.105268

39. 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 https://doi.org/10.1080/17470919.2014.948637

40. 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 https://doi.org/10.1177/0956797612472682

41. Evans S., Shergill S. S., Averbeck B. B (2010) **Oxytocin Decreases Aversion to Angry Faces in an Associative Learning Task** *Neuropsychopharmacology* **35**:2502–2509 https://doi.org/10.1038/npp.2010.110

42. 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 https://doi.org/10.1523/JNEUROSCI.3984-05.2005

43. Wu Q., Mao J., Li J (2020) **Oxytocin alters the effect of payoff but not base rate in emotion perception** *Psychoneuroendocrinology* **114** https://doi.org/10.1016/j.psyneuen.2020.104608

44. 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** https://doi.org/10.7554/eLife.61844

45. 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* https://doi.org/10.1038/s41380-024-02590-w

46. Han S., Ma Y (2015) **A Culture–Behavior–Brain Loop Model of Human Development** *Trends in Cognitive Sciences* **19**:666–676 https://doi.org/10.1016/j.tics.2015.08.010

47. 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 https://doi.org/10.1073/pnas.1604285113

48. 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 https://doi.org/10.1038/npp.2015.87

49. 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 https://doi.org/10.1038/s41593-019-0351-1

50. Zhang H., Gross J., Dreu C. D., Ma Y (2019) **Oxytocin promotes coordinated out-group attack during intergroup conflict in humans** *eLife* https://doi.org/10.7554/eLife.40698

51. 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 https://doi.org/10.1126/science.1189047

52. Dreu C. K. W. D., Gross J., Fariña A., Ma Y (2020) **Group Cooperation, Carrying-Capacity Stress, and Intergroup Conflict** *Trends in Cognitive Sciences* **24**:760–776 https://doi.org/10.1016/j.tics.2020.06.005

53. Patin A., Scheele D., Hurlemann R., Hurlemann R., Grinevich V. (2018) **Oxytocin and Interpersonal Relationships** *Behavioral Pharmacology of Neuropeptides: Oxytocin* Springer International Publishing https://doi.org/10.1007/7854_2017_22

54. 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 https://doi.org/10.1111/j.1749-6632.2009.04504.x

55. Stevens F., Taber K (2021) **The neuroscience of empathy and compassion in pro-social behavior** *Neuropsychologia* **159** https://doi.org/10.1016/j.neuropsychologia.202

56. 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 https://doi.org/10.1073/pnas.1911517117

57. Everett J. A. C. *et al.* (2021) **Moral dilemmas and trust in leaders during a global health crisis** *Nat Hum Behav* https://doi.org/10.1038/s41562-021-01156-y

58. 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 https://doi.org/10.1038/nn.3798

59. 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 https://doi.org/10.1016/j.neubiorev.2014.10.007

60. Harari-Dahan O., Bernstein A (2017) **Oxytocin attenuates social and non-social avoidance: Re-thinking the social specificity of Oxytocin** *Psychoneuroendocrinology* **81**:105–112 https://doi.org/10.1016/j.psyneuen.2017.04.005

61. Wang D., Ma Y (2020) **Oxytocin facilitates valence-dependent valuation of social evaluation of the self** *Communications Biology* **3**:1 https://doi.org/10.1038/s42003-020-01168-w

62. 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 https://doi.org/10.1097/AJP.0b013e31829f57df

63. 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 https://doi.org/10.1016/j.neuroscience.2017.09.041

64. 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** https://doi.org/10.1016/j.yhbeh.2022.105282

65. 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 https://doi.org/10.3758/BRM.41.4.1149

66. 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 https://doi.org/10.1016/j.biopsych.2015.06.016

67. 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 https://doi.org/10.1016/j.psychres.2014.04.031

68. 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 https://doi.org/10.1002/hup.2402

69. Vlaev I., Seymour B., Dolan R. J., Chater N (2009) **The Price of Pain and the Value of Suffering** *Psychological Science* **20**:309–317 https://doi.org/10.1111/j.1467-9280.2009.02304.x

70. Davis M. H (2011) **Interpersonal Reactivity Index [Dataset]** https://doi.org/10.1037/t01093-000

71. Wilkinson G. N., Rogers C. E (1973) **Symbolic Description of Factorial Models for Analysis of Variance** *Journal of the Royal Statistical Society Series C: Applied Statistics* **22**:392–399 https://doi.org/10.2307/2346786

72. 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 https://doi.org/10.3758/BRM.40.3.879

73. Preacher K. J., Rucker D. D., Hayes A. F (2007) **Addressing Moderated Mediation Hypotheses: Theory, Methods, and Prescriptions** *Multivariate Behavioral Research* **42**:185–227 https://doi.org/10.1080/00273170701341316

74. Hayes A. F. (2015) **An Index and Test of Linear Moderated Mediation** *Multivariate Behavioral Research*

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

Reviewing Editor

Xiaosi Gu

Icahn School of Medicine at Mount Sinai, New York, United States of America

Senior Editor

Michael Frank

Brown University, Providence, United States of America

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.

(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"? 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 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?

(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.

(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.

(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.

(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.

(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.

(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.

https://doi.org/10.7554/eLife.102756.1.sa2

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 on themselves or another person in exchange for monetary profits for themselves. In the loss domain, such hyperaltruistic tendency is 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 a 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.

https://doi.org/10.7554/eLife.102756.1.sa1

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 was actually executed.

Weaknesses:

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.

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.

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).

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.

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.

https://doi.org/10.7554/eLife.102756.1.sa0