# Correction

## PSYCHOLOGICAL AND COGNITIVE SCIENCES

Correction for "Harm to others outweighs harm to self in moral decision making," by Molly J. Crockett, Zeb Kurth-Nelson, Jenifer Z. Siegel, Peter Dayan, and Raymond J. Dolan, which appeared in issue 48, December 2, 2014, of *Proc Natl Acad Sci USA* (111:17320–17325; first published November 17, 2014; 10.1073/pnas.1408988111).

The authors note that they inadvertently omitted references to two articles by FeldmanHall et al. and Vlaev, respectively. The authors would like to cite the articles in the following sentence added to the first paragraph of the article: "Past studies have examined people's judgments in hypothetical scenarios, but recent work suggests hypothetical judgments cannot accurately predict real decisions (45, 46)."

The complete references appear below.

45. FeldmanHall O, et al. (2012) What we say and what we do: The relationship between real and hypothetical moral choices. *Cognition* 123(3):434–441.

46. Vlaev I (2012) How different are real and hypothetical decisions? Overestimation, contrast and assimilation in social interaction. *J Econ Psychol* 33(5):963–972.

www.pnas.org/cgi/doi/10.1073/pnas.1424572112

Downloaded by guest on April 21, 2021

# Harm to others outweighs harm to self in moral decision making

Molly J. Crockett[a,b,1], Zeb Kurth-Nelson[a,c], Jenifer Z. Siegel[a,b], Peter Dayan[d], and Raymond J. Dolan[a,c]

[a]Wellcome Trust Centre for Neuroimaging, University College London, London WC1N 3BG, United Kingdom; [b]Department of Experimental Psychology, University of Oxford, Oxford OX1 3UD, United Kingdom; [d]Gatsby Computational Neuroscience Unit, University College London, London WC1N 3AR, United Kingdom; and [c]Max Planck–University College London Centre for Computational Psychiatry and Ageing, London WC1B 5EE, United Kingdom

Edited by Daniel Kahneman, Princeton University, Princeton, NJ, and approved October 22, 2014 (received for review May 14, 2014)

**Concern for the suffering of others is central to moral decision making. How humans evaluate others' suffering, relative to their own suffering, is unknown. We investigated this question by inviting subjects to trade off profits for themselves against pain experienced either by themselves or an anonymous other person. Subjects made choices between different amounts of money and different numbers of painful electric shocks. We independently varied the recipient of the shocks (self vs. other) and whether the choice involved paying to decrease pain or profiting by increasing pain. We built computational models to quantify the relative values subjects ascribed to pain for themselves and others in this setting. In two studies we show that most people valued others' pain more than their own pain. This was evident in a willingness to pay more to reduce others' pain than their own and a requirement for more compensation to increase others' pain relative to their own. This "hyperaltruistic" valuation of others' pain was linked to slower responding when making decisions that affected others, consistent with an engagement of deliberative processes in moral decision making. Subclinical psychopathic traits correlated negatively with aversion to pain for both self and others, in line with reports of aversive processing deficits in psychopathy. Our results provide evidence for a circumstance in which people care more for others than themselves. Determining the precise boundaries of this surprisingly prosocial disposition has implications for understanding human moral decision making and its disturbance in antisocial behavior.**

altruism | morality | decision making | valuation | social preferences

Moral decisions often involve sacrificing personal benefits to prevent the suffering of others. Disregard for others' suffering is a core feature of antisocial and criminal behaviors (1) that confer tremendous economic and psychological costs on society (2). However, little is known about how people evaluate the costs of others' suffering, compared with their own suffering. Here, we address this question in two experiments by asking subjects to trade off profits for themselves against pain for themselves or an anonymous other person.

An initial prediction arises from studies of economic exchange in humans. It has been widely shown that people value others' monetary outcomes, evident in a willingness to donate money to anonymous strangers (3) and cooperate in social dilemmas (4). Nevertheless, these data overwhelmingly indicate people care about the monetary outcomes of others far less than their own monetary outcomes (3, 5). This suggests that people will evaluate the cost of others' pain to be greater than zero, but much lower than the cost of their own pain.

An alternative perspective emerges from studies investigating empathy. An aversion to the suffering of others is a powerful motivator for humans (6) and for our close primate relatives (7). Indeed, observing others in pain engages brain networks similar to those that respond to one's own pain (8). The magnitude of responses in these regions correlates with self-reported empathy (8) and predicts the likelihood that people will endure pain themselves to lessen the pain of others (9). The empathy perspective predicts that people will value others' pain similarly to how they value their own pain, to the extent that they empathize with the other person. Note, however, that this perspective predicts that the cost of pain for another will be no more than the cost of pain for oneself.

A third hypothesis stems from the observation that people dislike causing bad outcomes, particularly when those outcomes affect others (10, 11). In *The Theory of Moral Sentiments* Adam Smith argues that the "indelible stain" of guilt is worse than pain: "For one man ... unjustly to promote his own advantage by the loss or disadvantage of another, is more contrary to nature, than death, than poverty, than pain, than all the misfortunes which can affect him" (12). This moral sentiment, reflected in powerful social norms that proscribe harming others (5, 13), could lead some people to evaluate the cost of others' pain as higher than their own in a setting where they feel a degree of responsibility for that pain.

We developed a paradigm that enabled us to quantify how people evaluate the subjective costs of pain for themselves and others in a neutral social context. Pairs of individuals participated in each experimental session under conditions of complete anonymity. At the start of the experimental session a standard procedure determined each subject's pain threshold for an electrical shock stimulus delivered to the left wrist (14). We then used this thresholding procedure to create a shock stimulus for each subject that was mildly painful, but not intolerable. Importantly, this procedure enabled us to match the subjective intensity of shocks for all subjects, and all subjects were made aware of this fact (14).

Next, the two subjects were randomly assigned to the roles of "decider" and "receiver" (Fig. 1*A*, Fig. S1, and *Materials and*

## Significance

Concern for the welfare of others is a key component of moral decision making and is disturbed in antisocial and criminal behavior. However, little is known about how people evaluate the costs of others' suffering. Past studies have examined people's judgments in hypothetical scenarios, but there is evidence that hypothetical judgments cannot accurately predict actual behavior. Here we addressed this issue by measuring how much money people will sacrifice to reduce the number of painful electric shocks delivered to either themselves or an anonymous stranger. Surprisingly, most people sacrifice more money to reduce a stranger's pain than their own pain. This finding may help us better understand how people resolve moral dilemmas that commonly arise in medical, legal, and political decision making.

Author contributions: M.J.C., Z.K.-N., J.Z.S., P.D., and R.J.D. designed research; M.J.C. and J.Z.S. performed research; M.J.C., Z.K.-N., and P.D. contributed new reagents/analytic tools; M.J.C., Z.K.-N., J.Z.S., and P.D. analyzed data; and M.J.C., Z.K.-N., J.Z.S., P.D., and R.J.D. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. Email: molly.crockett@psy.ox.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1408988111/-/DCSupplemental.

*Methods*). Deciders made a series of choices between a smaller amount of money plus a smaller number of shocks, vs. a larger amount of money plus a larger number of shocks. The decider always received the money, but the shocks were allocated to the decider in half of the trials (Fig. 1 *B* and *D*) and to the receiver in the other half (Fig. 1 *C* and *E*). Deciders had to choose in each trial whether to switch from the highlighted default option to an alternative option by pressing a key within a time limit of 6 s. In half of the trials, the alternative option contained more money and shocks than the default (Fig. 1 *B* and *C*), whereas in the other half the alternative option contained less money and fewer shocks than the default (Fig. 1 *D* and *E*).

To avoid habituation and preserve choice independence no money or shocks were delivered during the task. Instead, one trial was randomly selected and implemented at the end of the experiment. All procedures were fully transparent to participants, and no deception was used in the paradigm. The experiment was designed such that the deciders' choices with respect to the receiver most likely reflected pure aversion to the pain suffered by this anonymous other person, rather than conscious concerns about reputation or reciprocity (*SI Materials and Methods*).

We deployed this paradigm in two separate studies. Our main dependent measure, derived from a computational model of deciders' choices, was a pair of subject-specific harm aversion parameters that characterized the subjective cost of pain for self and other. Initially we were agnostic about the distribution of harm aversion in our study population. Consequently, in study 1 ($n = 39$) we used a staircasing procedure that generated, for each individual decider, a personalized set of 152 choices that aimed to maximize the precision with which we could estimate his or her harm aversion parameters (*SI Materials and Methods*). In study 2 ($n = 41$), we presented deciders with a fixed set of 160 choices that was optimized to cover the full range of harm aversion observed in study 1 (*SI Materials and Methods*). In addition, by decorrelating money and shock magnitudes across trials we could examine how response times varied as a function of each of these factors.

## Results

We first directly compared deciders' choices for themselves and for the receiver, examining the proportion of trials on which deciders chose the more harmful option and the total number of shocks delivered. We can address this question optimally using the data from study 2, where deciders faced identical choice options in each experimental condition. Strikingly, most deciders were "hyperaltruistic," apparently valuing the receiver's pain more than their own (15, 16). Deciders were less likely to harm the receiver than themselves [$F_{(1,40)} = 7.033$, $P = 0.011$, Fig. S2*A*] and chose to deliver fewer shocks to the receiver than to themselves [$F_{(1,40)} = 6.30$, $P = 0.016$, Fig. S2*B*]. Deciders were hyperaltruistic both when they had the option to increase shocks for a profit [proportion of harmful choices: $t_{(40)} = 2.195$, $P = 0.034$; total number of shocks: $t_{(40)} = 2.027$, $P = 0.049$] and when they had the option to pay to decrease shocks [proportion of harmful choices: $t_{(40)} = 2.696$, $P = 0.010$; total number of shocks: $t_{(40)} = 2.6517$, $P = 0.011$].

To examine these findings in more detail, we fit a range of computational models to the choice data and found that deciders' choices were most parsimoniously explained by a model that allowed for distinct valuation of harm to self and other, together with a factor that accounted for loss aversion for both shocks and money:

$$\Delta V = (1-\kappa)\mathcal{L}_m\Delta m - \kappa\mathcal{L}_s\Delta s$$

$$\kappa = \begin{cases} \kappa_{self} & \text{if self trial} \\ \kappa_{other} & \text{if other trial} \end{cases}$$

$$\mathcal{L}_m = \begin{cases} 1 & \text{if } \Delta m > 0 \\ \lambda & \text{if } \Delta m < 0 \end{cases}$$

$$\mathcal{L}_s = \begin{cases} 1 & \text{if } \Delta s < 0 \\ \lambda & \text{if } \Delta s > 0 \end{cases},$$

where $\Delta V$ is the subjective value of switching from the default to the alternative option, $\Delta m$ and $\Delta s$ are the differences in money and shocks between the default and alternative options, $\lambda$ is a loss aversion parameter that captures the difference in subjective value between gains (increases in money or decreases in shocks) and losses (decreases in money or increases in shocks) (17), and $\kappa_{self}$ and $\kappa_{other}$ are harm aversion parameters that capture the subjective cost of pain for self and others. When $\kappa = 0$, deciders are minimally harm-averse and will accept any number of shocks to increase their profits; as $\kappa$ approaches 1, deciders become maximally harm-averse and will pay increasing amounts of money to avoid a single shock (Fig. S3). Trial-by-trial value differences were transformed into choice probabilities using a softmax function (18). This model explained deciders' choices well across both studies, correctly predicting 70% of deciders' choices in study 1 [95% confidence interval (CI) (65–74); note that the staircase procedure acts to create choices that are explicitly hard to predict] and 90% of deciders' choices in study 2 [95% CI (88–92); with the predictability benefiting from the fixed set of choices]. Bayesian model comparisons (19, 20) indicated that this model was favored over a range of alternative models, including economic models of social preferences (*SI Results* and Table S1).

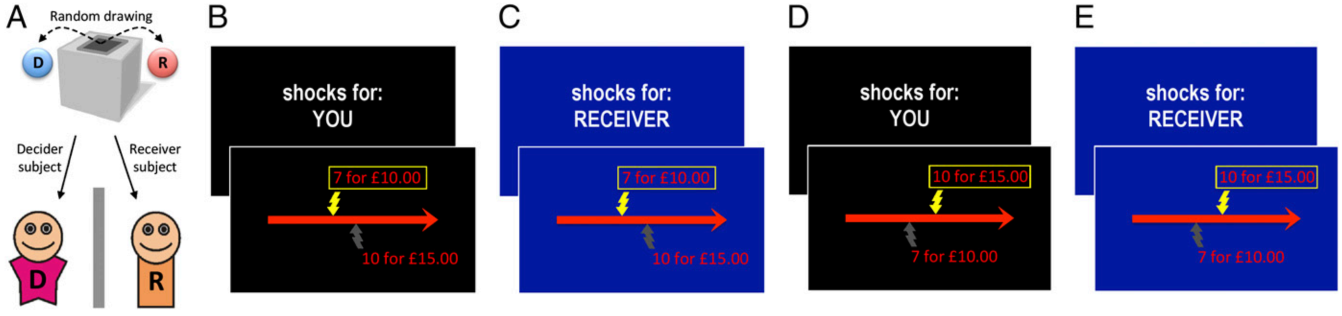

**Fig. 1.** Experimental design. (*A*) Subjects remained in separate testing rooms at all times and were randomly assigned to roles of decider and receiver (Fig. S1). (*B*–*E*) In each trial the decider chose between less money and fewer shocks, vs. more money and more shocks. The money was always for the decider, but in half the trials the shocks were for the decider (*B* and *D*) and in the other half the shocks were for the receiver (*C* and *E*). In all trials, if the decider failed to press a key within 6 s the highlighted default (top) option was registered; if the decider pressed the key, the alternative (bottom) option was highlighted and registered instead. In half the trials, the alternative option contained more money and shocks than the default (*B* and *C*), and in the other half the alternative option contained less money and fewer shocks than the default (*D* and *E*).

Consistent with our observation that deciders were less likely to choose the more harmful option for the receiver than for themselves, and delivered fewer shocks to the receiver than to themselves, parameter estimates for harm aversion indicated that most deciders were hyperaltruistic, placing a higher cost on the receiver's pain than on their own pain [$\kappa_{other} > \kappa_{self}$, study 1: $t_{(38)} = 3.113$, $P = 0.004$, Fig. 2 *A* and *B*; study 2: $t_{(40)} = 2.23$, $P = 0.031$, Fig. 2 *D* and *E*]. Harm aversion for self and others was strongly related to one another (study 1: $r = 0.612$, $P < 0.001$; study 2: $r = 0.590$, $P < 0.001$; Fig. 2 *C* and *F*), suggesting that deciders' aversion to the receiver's pain may be anchored on their aversion to their own pain (6, 21).

Deciders were also loss-averse [$\lambda > 1$, study 1: $t_{(38)} = 4.056$, $P < 0.001$; study 2: $t_{(40)} = 1.959$, $P = 0.057$]. Note that loss aversion in the context of our experiment produces a pattern of choices in which subjects require more money to accept an increase in shocks than they are willing to pay for an equivalent decrease in shocks, an effect consistent with an omission bias in moral decision making (11). Our preferred model contained a single loss aversion parameter that was applied both to decreases in money and increases in shocks, equally for self and other trials. Model comparisons indicated that loss aversion was not significantly different for shocks vs. money, nor for self vs. other (*Supporting Information*). In line with this, hyperaltruistic valuation of pain was evident both for increases and decreases in shocks. We confirmed this by extracting harm aversion parameters for self and other, separately for increasing and decreasing trials. Deciders required more money to increase another's shocks than to increase their own shocks [study 1: $t_{(38)} = 2.823$, $P = 0.008$; study 2: $t_{(40)} = 2.039$, $P = 0.048$], and they were willing to pay more money to decrease another's shocks than to decrease their own shocks [study 1: $t_{(38)} = 3.175$, $P = 0.003$; study 2: $t_{(40)} = 2.703$, $P = 0.01$].

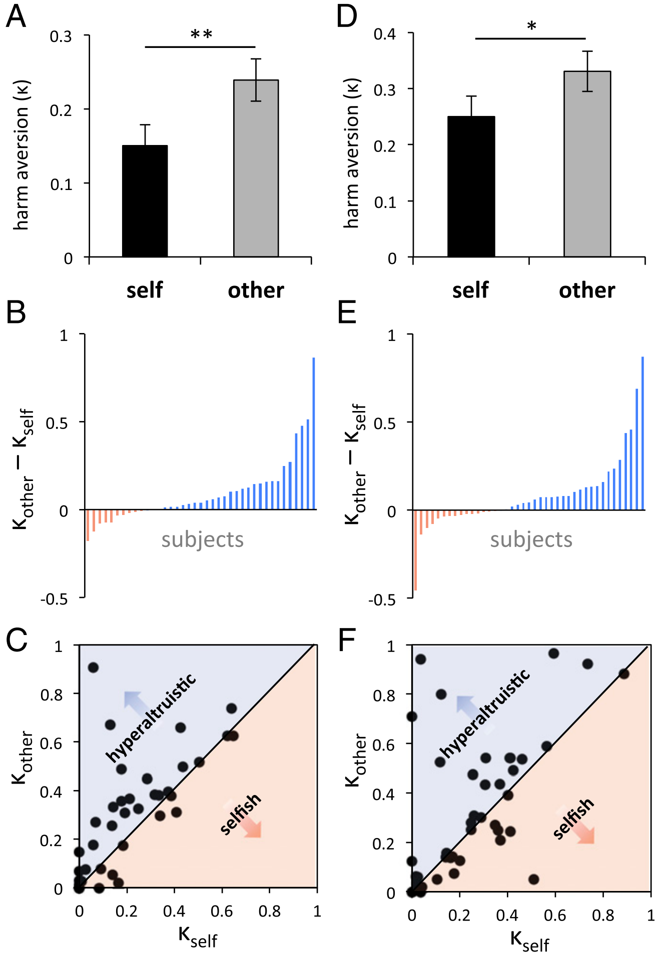

**Fig. 2.** Harm aversion for self and other in study 1 (*A*–*C*) and study 2 (*D*–*F*). (*A* and *D*) Estimates of harm aversion for self and other. Error bars represent SEM difference between $\kappa_{self}$ and $\kappa_{other}$. (*B* and *E*) Distribution of hyperaltruism ($\kappa_{other} - \kappa_{self}$) across subjects. (*C* and *F*) Correlations between harm aversion for self and other. *$P < 0.05$, **$P < 0.01$.

As noted, these results stand in sharp contrast to economic studies of social preferences, which show that people value others' monetary outcomes far less than their own (3). To examine deciders' valuation of others' monetary outcomes (rather than pain), we provided them with an opportunity to donate a proportion of their earnings to charity at the end of the experiment. Consistent with previous findings (3), deciders valued others' monetary outcomes less than their own, donating only a minority of their earnings (study 1: mean ± SE = 25 ± 5%; study 2: mean ± SE = 16 ± 3%; Fig. S4). We note that valuation of others' monetary outcomes in the context of charitable donation is not directly comparable to valuation of others' pain in the current experiment, and an interesting question for future study would be to investigate the valuation of pain and money for the same target individual. Nevertheless, our observation of relatively selfish charitable donation behavior in the present experiment suggests that the valuation of others' outcomes is highly context-dependent.

Recent studies have investigated the relationship between deliberation speed and prosocial motivation (4, 22), showing that that people who consider helpful decisions more quickly have stronger prosocial preferences (4). Here, we examined whether people who consider harmful decisions more quickly have weaker prosocial preferences. In the current study, we predicted that harm aversion would generally relate to slower responding, because anticipating aversive outcomes often leads to behavioral inhibition (23). To investigate this, we fit a linear model to subjects' response time data from study 2 and examined how shocks, money, and shock recipient (self or other) affected response times (Table S2). In line with our predictions, subjects were slower in cases where responding resulted in larger shock increases [$t_{(40)} = 5.022$, $P < 1e^{-5}$] and when the maximum number of possible shocks was large [$t_{(40)} = 6.6937$, $P < 5e^{-8}$]. Both of these slowing effects correlated significantly with subjects' personal harm aversion parameter $\kappa_{self}$ (shock increases: $r = 0.651$, $P < 0.0001$; maximum shocks: $r = 0.391$, $P = 0.011$). These results indicate that the prospect of harm gives people pause, and more so to the extent that people are harm-averse.

Choice data in both studies showed that harm aversion was stronger for others than for the self. If harm aversion is associated with slower responses, then subjects should on average respond more slowly when decisions affect others. To test this, we first compared raw response times for self and other trials. Subjects were indeed slower to respond in other trials than in self trials [$t_{(40)} = 2.079$, $P = 0.044$; Fig. 3*A*], an effect that remained significant after controlling for other task parameters [$t_{(40)} = 2.499$, $P = 0.017$] and for differences in subjective value between choice options [$t_{(40)} = 2.388$, $P = 0.022$; *SI Results*]. Response times related to choices, with more hyperaltruistic subjects showing a greater degree of slowing on choices for others relative to choices for self ($r = 0.419$, $P = 0.006$; Fig. 3*B*). The relationship between slowing for others and hyperaltruistic preferences was robust even when controlling for slowing related to differences in the subjective values of the choice options (*SI Results*).

Antisocial behavior in psychopathy is linked to blunted affective responses to aversive stimuli (24). We examined individual differences in psychopathic traits and their relationship to harm aversion for self and others, pooling the data from both studies to maximize power. A link between psychopathy and harm aversion was evident; psychopathic traits were negatively correlated with harm aversion, both for self and others ($\kappa_{self}$: $r = -0.327$, $P = 0.003$; $\kappa_{other}$: $r = -0.399$, $P = 0.0002$). However, psychopathy was only correlated with hyperaltruism at trend level ($\kappa_{other} - \kappa_{self}$; $r = -0.19$, $P = 0.091$). We also examined the relationship between psychopathic traits and response times in study 2. Psychopathic traits

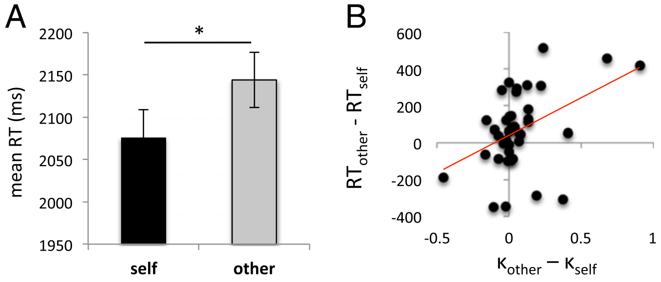

**Fig. 3.** Slowing when deciding for others predicts hyperaltruistic valuation of pain. (*A*) Deciders were slower to decide when choices affected others, relative to when choices affected themselves. Error bars represent the SEM. $^*P < 0.05$. (*B*) The degree of slowing when deciding the fate of others (relative to oneself) predicted the degree of hyperaltruism, ($r = 0.419$, $P = 0.006$).

were negatively correlated with slowing related to shock increases ($r = -0.36$, $P = 0.02$) and the maximum number of shocks ($r = -0.38$, $P = 0.01$) but were not related to slowing for others ($r = 0.09$, $P = 0.56$).

Finally, we observed sex differences, with males showing less harm aversion than females, both for self [$t_{(78)} = -3.849$, $P < 0.001$] and for others [$t_{(78)} = -2.594$, $P = 0.011$]. Because psychopathic traits are more prevalent in males, both in past research (25) and the current sample [$t_{(78)} = 3.157$, $P = 0.002$] we reasoned that sex differences in harm aversion may be moderated by psychopathy. This was indeed the case for harm aversion toward others. When jointly examining the effects of psychopathy and sex on harm aversion for others, sex effects on harm aversion for others disappeared [$F_{(1,77)} = 2.313$, $P = 0.132$], whereas the relationship between psychopathy and harm aversion for others remained significant [$F_{(1,77)} = 9.855$, $P = 0.002$].

## Discussion

We describe an experimental setting in which people cared more about an anonymous stranger's pain than their own pain, despite the fact that their decisions were completely anonymous, with no future possibility of being judged adversely or punished. This counterintuitive finding is inconsistent with previous studies of social preferences, where most people value others' monetary outcomes much less than their own (3). Our results are equally unpredicted by previous work on empathy (6, 8), which implies people will pay at most the same amount to prevent others' pain as their own pain. The observation that people apparently valued others' pain more than their own pain in our experiments indicates there must be additional factors besides empathy that influence choices in this setting.

At a proximate level, there are several potential explanations for our findings that are not mutually exclusive. First, people might adopt the self-serving view that they themselves are more able to tolerate pain than others (26), thus making it appropriate to pay more to reduce others' pain. To explore this possibility we examined deciders' ratings of the estimated unpleasantness of shocks for themselves and for the receiver. We found that deciders tended to estimate that shocks would indeed be slightly more unpleasant for the receiver than themselves [study 1: $t_{(19)} = 2.689$, $P = 0.015$; study 2: $t_{(40)} = 1.839$, $P = 0.073$; Fig. S5]. However, the differences in these explicit estimates did not predict the extent to which deciders valued the receiver's pain more than their own pain (study 1: $r = -0.014$, $P = 0.952$; study 2: $r = 0.135$, $P = 0.401$). Furthermore, this explanation cannot account for our finding that people are slower to decide the fate of others than themselves and that slowing relates to hyperaltruistic valuation of pain.

An alternative explanation stems from reports that people dislike being responsible for bad outcomes, particularly when they affect others (10, 11). Thus, even if people find others' pain inherently less aversive than their own pain, the added cost of moral responsibility in the current setting could make people value others' pain more than their own. Computing the cost of moral responsibility presumably takes time, and this computation is invoked for decisions affecting others but not for decisions affecting oneself. The responsibility explanation is therefore consistent with our observation that people were slower when deciding for others than for themselves, and the extent to which they were slower predicted the degree of hyperaltruistic valuation of pain. The idea that prosocial behavior involves careful deliberation is an old one (12) and is even reflected in our everyday language—we describe morally praiseworthy people as "thoughtful" and "considerate," whereas selfish people are described as "thoughtless" and "inconsiderate." This distinction also informs moral judgments. Those who make harmful decisions quickly are judged more negatively than those who make harmful decisions after prolonged deliberation (22). Our data indicate there is some truth in such judgments given the finding that those who consider harmful decisions more quickly are also more selfish.

A third possibility, which is orthogonal to a responsibility account and can also explain the relationship between hyperaltruistic behavior and response times, arises from the fact that predicting how our decisions will affect others is inherently uncertain (21). If deciders assume that the receiver's mapping from number of shocks to subjective unpleasantness is nonlinear, then this uncertainty could induce a form of risk premium in the moral costs of imposing what might be intolerable pain on another. To avoid these moral costs, people may adopt a risk-averse, conservative strategy, leading them to systematically err on the side of reducing others' pain at their own expense. Intriguingly, many deciders expressed this very logic when explaining their choices retrospectively. For example, one characteristic decider remarked, "I knew what I could handle but I was less sure about the other person and didn't want to be cruel." In policy making this attitude is enshrined in the precautionary principle, which prohibits actions that carry a risk of causing harm and imposes on decision makers the burden of proving actions are harmless (27).

Because uncertainty in decision making is generally reflected in longer response times (28, 29), this explanation for hyperaltruistic valuation of pain is also consistent with our finding that hyperaltruism was greater in deciders who were slower to decide the fate of receivers than of themselves. However, the uncertainty account makes an additional prediction concerning behavioral noise. Uncertainty in decision making is often reflected in choice noisiness, or more formally the fidelity with which decision values are translated into choices (28). If uncertainty when choosing for others relates to hyperaltruism, then hyperaltruism should be greater in deciders with noisier choices for the receiver than for themselves. An exploratory analysis supported this prediction (*SI Results* and Fig. S6). Further research is required to tease apart the roles of responsibility and uncertainty in moral decision making.

Recent theoretical accounts of prosocial behavior posit that prosociality is reflexive and automatic, whereas selfishness involves deliberation (4, 30, 31). However, this conclusion arises primarily out of studies involving rewarding outcomes for others. Our data suggest a broadening of the theory, in that the relationship between deliberation and prosocial behavior may depend on valence. Specifically, those with stronger prosocial preferences may be faster in rewarding contexts but slower in aversive contexts. This account gels with past studies showing that people who help others quickly are judged more positively than those who hesitate, but people who harm others quickly are judged more negatively than those who hesitate (22).

One unresolved issue is whether the relationship between deliberation and prosocial behavior depends on the valence of the action (helping vs. harming) or the valence of the outcome (positive or negative). These two factors are often confounded, in that studies of helpful actions often use only positive outcomes (e.g., charitable donations or contributions to public goods), and studies of harmful actions often use only negative outcomes (e.g., in scenarios involving killing one to save many). We surmise that the valence of the outcome plays a more important role and

PSYCHOLOGICAL AND COGNITIVE SCIENCES

could reflect the influence of a reflexive, Pavlovian system that promotes automatic approach and withdrawal responses to appetitive and aversive stimuli (32). In the current study involving aversive outcomes, those with stronger prosocial preferences were slower to respond, regardless of whether they chose to help or harm. Because these data are correlational, however, it is difficult to draw strong conclusions about the causal influence of deliberative processes (33). We further note that, in contrast with some previous studies (4), our experimental design does not distinguish between the time required to evaluate the choice options and that required to register the decision. Stronger evidence comes from a recent study showing that that time pressure increased prosocial behavior in an extraction game where actions were harmful but outcomes were rewarding (34). Disentangling how the valence of actions vs. outcomes shapes the relationship between deliberation and prosocial preferences is an important direction for future study.

We sought to provide a set of neutral baseline conditions under which to examine the extent to which people value others' pain relative to their own pain. There are many reasons why people might show hyperaltruistic behavior outside the laboratory, including reputational concerns, the possibility of being punished, and a prior relationship with the object of harm. We attempted to remove any conscious, explicit concerns about reputation and reciprocity in several ways. First, in our instructions to the decider we emphasized the confidentiality of their decisions. Deciders completed the task alone in the testing room and knew that all participants' identities were concealed from one another. These steps were taken to minimize the influence of any conscious motivation to preserve one's reputation in the eyes of the receiver or the experimenters and motivations related to established social relationships. Furthermore, deciders knew that the roles were fixed, and that the receiver would not have any opportunity to retaliate against the decider's choices. These steps were taken to minimize any potential influence of conscious motivations for reciprocity and avoiding punishment, although we acknowledge that unconscious or habitual motivations for reputation or reciprocity could potentially spill over into putatively anonymous decisions (34–36).

An open question is whether our observation of hyperaltruistic valuation of pain would generalize to harms of greater magnitude. For ethical reasons, we were limited in the level of pain we could deliver in the laboratory, but moral dilemmas in the real world often involve more drastic consequences than a limited number of mildly painful electric shocks, spanning financial ruin and poverty to disfigurement and even death. Determining how people evaluate severe harms remains an empirical challenge. Using incentivized laboratory measures is clearly not ethical. However, asking people to state compensation prices for hypothetical severe harms is unlikely to yield reliable data, as evidenced by an early study in which people stated, on average, that they would require a compensation of $100,000 (in 1937 dollars) to eat a live earthworm, but only $57,000 to have their little toe cut off (37). Nevertheless, we can speculate about boundary conditions for our observed effects. It is possible that hyperaltruistic valuation of suffering is limited to harms of little consequence, where being nice confers large moral benefits but requires only small personal costs. An alternative possibility is that for more severe harms hyperaltruistic behavior may be evident for harmful actions but not harmful omissions. For example, one might predict that people would require more compensation to break someone else's leg than to break their own leg but would not be willing to pay more to prevent someone else from breaking their leg than to prevent themselves from suffering the same fate. Still another possibility is that hyperaltruistic behavior in the case of severe harms depends critically on a custodial social relationship between the agent and the victim. Cases of parents making great sacrifices for their children are not uncommon, and maritime customs have long dictated that captains of sinking ships should place the lives of their passengers above their own. Exploring the boundaries of hyperaltruistic valuation of suffering is an important question for future research and might require observational or field methods given the limits of imposing harmful outcomes in the laboratory.

At the ultimate level of explanation, we suggest that an apparent disposition to value others' suffering more than one's own in some settings is likely to have selective value. One account of altruism posits that people behave altruistically primarily because they value others' outcomes (38). However, many studies suggest that altruistic behavior is not motivated solely by concern for others' outcomes (39–41). These observations, together with the current results, suggest that people behave altruistically because they value altruistic actions (which are often visible to others) in addition to outcomes (which are sometimes obscured or delayed). This idea is consistent with theories highlighting the importance of reciprocity and partner choice in the evolution of prosocial behavior (36, 42). Social norms that proscribe harm to others are widespread, and violation of these norms often results in punishment (5, 13). Those who are more cautious when deciding about others' pain would thus be less likely to suffer the costs of such punishments. Whether this disposition is "innate" in the sense of an evolutionarily prescribed prior on the costs of social harms, or learned through social experience, is an important topic for future studies.

Social interactions are fraught with uncertainty because, try as we might, we can never truly know what it is like to occupy someone else's shoes (21). Instead, we must rely on our best estimates of others' beliefs and preferences to guide social decision making (21, 43) and tread carefully when their fate rests in our hands. Here, we provide evidence for an apparently hyperaltruistic valuation of others' pain that is associated with slower choices when making decisions that affect others. This disposition cannot be explained by empathy alone, and understanding its boundary conditions has implications for the many medical, legal, and political decisions that involve tradeoffs between financial profits and possible human suffering. Our approach provides novel methods for quantifying moral preferences that have previously been measured mainly via self-report, enabling the development of new computational frameworks for investigating antisocial behavior and its neural antecedents.

## Materials and Methods

**Participants.** Healthy volunteers were recruited from the University College London (UCL) psychology department and the Institute of Cognitive Neuroscience subject pools. Participants with a history of systemic or neurological disorders, psychiatric disorders, medication/drug use, pregnant women, and more than two years' study of psychology were excluded from participation. Individuals who had previously participated in social interaction studies were also excluded owing to concerns that prior experience of being deceived could compromise subjects' belief in our paradigm, which did not use deception. Furthermore, to minimize variability in subjects' experiences with the experimental stimuli, we excluded participants previously enrolled in studies involving electric shocks.

Forty-five pairs of participants took part in study 1. Posttesting, three participants indicated they had been dishonest on the screening questionnaire; two participants expressed doubts as to whether the receiver would receive the shocks, and one participant indicated he could discern the sex of the receiver. All of these participants were in the role of decider and were excluded from further analysis, leaving a total of 39 participants in the role of decider whose data were analyzed in study 1 (19 males, mean age 23.5 y). Forty-two pairs of participants took part in study 2. One participant in the role of decider expressed doubts as to whether the receiver would receive the shocks and was therefore excluded from further analysis. This left a total of 41 participants in the role of decider whose data were analyzed in study 2 (15 males, mean age 23.4 y).

**Procedure.** The study took place at the Wellcome Trust Centre for Neuroimaging in London and was approved by the UCL Research Ethics Committee (4418/001). Participants completed a battery of online trait questionnaires approximately 1 wk before attending a single testing session. Two individuals participated in each session. They arrived at staggered times and were led to separate testing rooms without seeing one another to ensure complete anonymity.

After providing informed consent, participants completed a pain thresholding procedure that has been described in detail elsewhere (14). This procedure allowed us to (*i*) control for heterogeneity of skin resistance between participants, thus enabling us to deliver shocks of matched subjective intensity to different participants; (*ii*) administer a range of potentially painful stimuli in an ethical manner during the task itself; and (*iii*) provide subjects with experience of the shocks before the decision-making task.

Subjects were then randomly assigned to roles of either decider or receiver (*SI Materials and Methods* and Fig. S1). Following role assignment, the receiver subject completed a moral judgment task (data to be reported separately), and the decider subject completed a decision task, which we focus on presently. See *SI Materials and Methods* for task details.

After finishing the decision task, deciders completed self-report measures concerning their experiences during the experiment. Following this, subjects were given an opportunity to make a charitable donation. At the end of the session, one trial was randomly selected and actually implemented. Finally, before departing the laboratory participants completed debriefing questionnaires designed to assess their beliefs about the experimental setup.

**Computational Model of Moral Decision Making.** By formalizing the components of the decision process with a trial-by-trial mathematical model and fitting the model to deciders' choices we could probe how deciders evaluated the costs of their own and the receiver's pain and used these values to make their decisions. We compared a variety of models, each of which explained choices in terms of the value difference ($\Delta V$) between the default and alternative options. For all models, trial-by-trial value differences were transformed into choice probabilities using a softmax function (18):

$$P(\textit{choose alternative}) = \left(\frac{1}{1+e^{-\gamma\Delta V}}\right)(1-2\varepsilon)+\varepsilon,$$

where $\gamma$ is a subject-specific inverse temperature parameter that characterizes the sensitivity of choices to $\Delta V$ and $\varepsilon$ is a lapse rate that captures choice noisiness resulting from factors independent of $\Delta V$ (such as inattention). We optimized subject-specific parameters across trials using nonlinear optimization implemented in MATLAB (MathWorks, Inc.) for maximum likelihood estimation. Estimates were found to be very reliable and were confirmed with multiple random starts of optimization and for smaller models also by calculating the likelihood function at a multidimensional grid of points covering the entire parameter space. Summary statistics were then calculated from these parameter estimates at the group level, treating each parameter estimate as a random effect (44).

We compared models using Bayesian model comparison techniques (19, 20). In individual subjects, we computed Bayesian Information Criterion (BIC) scores for each model fit and summed the BIC scores across subjects to obtain a group BIC score. BIC penalizes models with a greater number of parameters, and the model with the lowest group BIC score is the preferred model.

**ACKNOWLEDGMENTS.** We thank A. de Berker, J. Haushofer, L. Hunt, C. Mathys, R. Rutledge, B. Seymour, P. Smittenaar, G. Story, I. Vlaev, and J. Winston for helpful comments. This work was supported by a Sir Henry Wellcome Postdoctoral Fellowship awarded to M.J.C. (092217/Z/10/Z). Z.K.-N. is supported by a Joint Initiative on Computational Psychiatry and Ageing Research between the Max Planck Society and University College London. P.D. is funded by the Gatsby Charitable Foundation. R.J.D. is supported by the Wellcome Trust (078865/Z/05/Z). The Wellcome Trust Centre for Neuroimaging is supported by core funding from the Wellcome Trust (091593/Z/10/Z).

1. Blair RJR (1995) A cognitive developmental approach to mortality: Investigating the psychopath. *Cognition* 57(1):1–29.
2. Anderson DA (2012) The cost of crime. *Found TrendsR Microecon* 7(3):209–265.
3. Engel C (2011) Dictator games: A meta study. *Exp Econ* 14(4):583–610.
4. Rand DG, Greene JD, Nowak MA (2012) Spontaneous giving and calculated greed. *Nature* 489(7416):427–430.
5. Henrich J, et al. (2010) Markets, religion, community size, and the evolution of fairness and punishment. *Science* 327(5972):1480–1484.
6. Batson CD, Duncan BD, Ackerman P, Buckley T, Birch K (1981) Is empathic emotion a source of altruistic motivation? *J Pers Soc Psychol* 40(2):290–302.
7. Masserman JH, Wechkin S, Terris W (1964) "Altruistic" behavior in rhesus monkeys. *Am J Psychiatry* 121(6):584–585.
8. Singer T, et al. (2004) Empathy for pain involves the affective but not sensory components of pain. *Science* 303(5661):1157–1162.
9. Hein G, Silani G, Preuschoff K, Batson CD, Singer T (2010) Neural responses to ingroup and outgroup members' suffering predict individual differences in costly helping. *Neuron* 68(1):149–160.
10. Kahneman D (2013) *Thinking, Fast and Slow* (Farrar, Straus and Giroux, New York).
11. Ritov I, Baron J (1990) Reluctance to vaccinate: Omission bias and ambiguity. *J Behav Decis Making* 3(4):263–277.
12. Smith A (1759) *The Theory of Moral Sentiments* (A. Millar, London).
13. Fehr E, Fischbacher U (2004) Social norms and human cooperation. *Trends Cogn Sci* 8(4):185–190.
14. Vlaev I, Seymour B, Dolan RJ, Chater N (2009) The price of pain and the value of suffering. *Psychol Sci* 20(3):309–317.
15. Kitcher P (1998) Psychological altruism, evolutionary origins, and moral rules. *Philos Stud* 89(2–3):283–316.
16. Kitcher P (1993) The evolution of human altruism. *J Philos* 90(10):497–516.
17. Kahneman D, Tversky A (1979) Prospect theory: An analysis of decision under risk. *Econometrica* 47(2):263.
18. Daw ND (2011) *Decision Making, Affect, and Learning: Attention and Performance XXIII*, eds Delgado MR, Phelps EA, Robbins TW (Oxford Univ Press, Oxford).
19. Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 6(2):461–464.
20. Burnham KP, Anderson DR (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (Springer, Berlin).
21. Harsanyi JC (1977) Morality and the theory of rational behavior. *Soc Res (New York)* 44(4):623–656.
22. Critcher CR, Inbar Y, Pizarro DA (2013) How quick decisions illuminate moral character. *Soc Psychol Personal Sci* 4(3):308–315.
23. Crockett MJ, Clark L, Robbins TW (2009) Reconciling the role of serotonin in behavioral inhibition and aversion: Acute tryptophan depletion abolishes punishment-induced inhibition in humans. *J Neurosci* 29(38):11993–11999.
24. Blair RJR (2013) The neurobiology of psychopathic traits in youths. *Nat Rev Neurosci* 14(11):786–799.
25. Vitale JE, Newman JP (2001) Using the psychopathy checklist-revised with female samples: Reliability, validity, and implications for clinical utility. *Clin Psychol Sci Pract* 8(1):117–132.
26. Brown JD (1986) Evaluations of self and others: Self-enhancement biases in social judgments. *Soc Cogn* 4(4):353–376.
27. Sunstein CR (2005) *Laws of Fear: Beyond the Precautionary Principle* (Cambridge Univ Press, Cambridge, UK).
28. De Martino B, Fleming SM, Garrett N, Dolan RJ (2013) Confidence in value-based choice. *Nat Neurosci* 16(1):105–110.
29. Pleskac TJ, Busemeyer JR (2010) Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychol Rev* 117(3):864–901.
30. Zaki J, Mitchell JP (2013) Intuitive prosociality. *Curr Dir Psychol Sci* 22(6):466–470.
31. Zaki J, Mitchell JP (2011) Equitable decision making is associated with neural markers of intrinsic value. *Proc Natl Acad Sci USA* 108(49):19761–19766.
32. Crockett MJ (2013) Models of morality. *Trends Cogn Sci* 17(8):363–366.
33. Evans AM, Dillon KD, Rand DG (2014) *Reaction Times and Reflection in Social Dilemmas: Extreme Responses are Fast, but Not Intuitive* (Social Science Research Network, Rochester, NY).
34. Rand DG, et al. (2014) Social heuristics shape intuitive cooperation. *Nat Commun* 5:3677.
35. Delton AW, Krasnow MM, Cosmides L, Tooby J (2011) Evolution of direct reciprocity under uncertainty can explain human generosity in one-shot encounters. *Proc Natl Acad Sci USA* 108(32):13335–13340.
36. Rand DG, Nowak MA (2013) Human cooperation. *Trends Cogn Sci* 17(8):413–425.
37. Thorndike EL (1937) Valuations of certain pains, deprivations, and frustrations. *Pedagog Semin J Genet Psychol* 51(2):227–239.
38. Becker GS (1974) *A Theory of Social Interactions* (National Bureau of Economic Research, Cambridge, MA).
39. Andreoni J (1990) Impure altruism and donations to public goods: A theory of warm-glow giving. *Econ J* 100(401):464.
40. Dana J, Cain DM, Dawes RM (2006) What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organ Behav Hum Decis Process* 100(2):193–201.
41. List JA (2007) On the interpretation of giving in dictator games. *J Polit Econ* 115(3): 482–493.
42. Trivers R (1971) The evolution of reciprocal altruism. *Q Rev Biol* 46(1):35–57.
43. Yoshida W, Dolan RJ, Friston KJ (2008) Game theory of mind. *PLOS Comput Biol* 4(12): e1000254.
44. Holmes A, Friston K (1998) Generalisability, random effects & population inference. *Neuroimage* 7:S754.