Harm to self outweighs benefit to others in moral decision making

Lukas J. Volz[a,b,1], B. Locke Welborn[a,1], Matthias S. Gobel[a], Michael S. Gazzaniga[a,2], and Scott T. Grafton[b]

[a]SAGE Center for the Study of the Mind, University of California, Santa Barbara, CA 93106; and [b]Department of Psychological and Brain Sciences, University of California, Santa Barbara, CA 93106

Contributed by Michael S. Gazzaniga, June 12, 2017 (sent for review May 1, 2017; reviewed by Sarah-Jayne Blakemore and Adina Roskies)

How we make decisions that have direct consequences for ourselves and others forms the moral foundation of our society. Whereas economic theory contends that humans aim at maximizing their own gains, recent seminal psychological work suggests that our behavior is instead hyperaltruistic: We are more willing to sacrifice gains to spare others from harm than to spare ourselves from harm. To investigate how such egoistic and hyperaltruistic tendencies influence moral decision making, we investigated trade-off decisions combining monetary rewards and painful electric shocks, administered to the participants themselves or an anonymous other. Whereas we replicated the notion of hyperaltruism (i.e., the willingness to forego reward to spare others from harm), we observed strongly egoistic tendencies in participants' unwillingness to harm themselves for others' benefit. The moral principle guiding intersubject trade-off decision making observed in our study is best described as egoistically biased altruism, with important implications for our understanding of economic and social interactions in our society.

morality | decision making | altruism | egoism | social cognition

Altruistic and egoistic tendencies in making individual decisions strongly influence the harmonious functioning of our society. Economists typically characterize human decision makers as profoundly selfish when distributing rewards among themselves and others (1, 2). By contrast, when allocating physical harm to themselves and others, seminal work by Crockett et al. has shown that people are hyperaltruistic (3, 4). Specifically, people are willing to forego more reward to reduce others' pain than to spare themselves from harm, suggesting that they value others' welfare more than their own (3). The tension between selfish and (hyper)altruistic views of human decision making calls into question whether the principles guiding moral decisions are the same when distributing rewards versus harms.

Instead, it is possible that harm introduces a new dimension to moral choice, such that trade-off decisions integrating rewards and harms are not interchangeable. Support for this hypothesis stems from human neuroscience, which emphasizes that different neural circuitries may underlie the representations of reward and harm (5, 6). In addition, it is not clear whether harm for self and harm for others are similarly integrated in moral decision making.

Indeed, moral philosophers have distinguished between contexts in which agents harm others and those in which they fail to benefit others (7). Moreover, accepting physical harm to the self to benefit others (but not the inverse) has sometimes been considered supererogatory: morally praiseworthy although not required (8, 9).

Therefore, the modality of the cost for the self, i.e., whether we forgo monetary reward or accept harm in the form of painful electric shocks may lead to fundamentally different trade-off decisions. Importantly, Crockett et al. (3, 4) defined hyperaltruism based on the observation that participants forgo reward to spare others from harm. However, the complementary case, i.e., in which participants accept harm to reward others, was not tested. Given that reward and harm for the self may be distinctly represented (5, 6), altruistic behavior involving harm for the self may fundamentally differ from hyperaltruism, which only comprises reward for the self and harm for the other (3, 4).

We here sought to address whether moral behavior critically depends on the modality of consequences and resolve the tension between the contradictory interpretations of egoistic and hyperaltrustic tendencies. We adapted a recent approach–avoid paradigm from research on the representation of reward and harm in the monkey brain (10, 11). In this paradigm, participants accepted or rejected offers in which varying monetary rewards were available, but only in exchange for painful electric shocks at varying intensities. Using this comprehensive framework, we evaluated participants' willingness to accept trade offs when differing amounts of reward for self or other were combined with different intensities of harm for self or other.

Sixty-three participants judged offers in two intrapersonal conditions (reward and harm for the same person) and two interpersonal conditions (reward for one person, harm for the other). In the "self" condition, participants received both the offered amount of reward and associated harm—reflecting their individual trade-off preferences indicated by their choices. In the "other" condition, participants judged on behalf of another person, who received both reward and harm. In the harm-for-other condition, participants imposed harm on the other person to secure rewards for the self, whereas in the harm-for-self condition, participants endured harm to gain rewards for the other person (Fig. 1). In each condition, the same set of offers, comprising specific combinations of reward and harm (Fig. 2), was presented in a randomized order. Of note, participants' actual trade-off decisions were implemented,

Significance

Principles guiding decisions that affect both ourselves and others are of prominent importance for human societies. Previous accounts in economics and psychological science have often described decision making as either categorically egoistic or altruistic. Instead, the present work shows that genuine altruism is embedded in context-specific egoistic bias. Participants were willing to both forgo monetary reward to spare the other from painful electric shocks and also to suffer painful electric shocks to secure monetary reward for the other. However, across all trials and conditions, participants accrued more reward and less harm for the self than for the other person. These results characterize human decision makers as egoistically biased altruists, with important implications for psychology, economics, and public policy.

Author contributions: L.J.V., B.L.W., and M. S. Gobel designed research; L.J.V., B.L.W., and M. S. Gobel performed research; L.J.V., B.L.W., and M. S. Gobel analyzed data; and L.J.V., B.L.W., M. S. Gobel, M. S. Gazzaniga, and S.T.G. wrote the paper.

Reviewers: A.R., Dartmouth College; and S.-J.B., Institute of Cognitive Neuroscience, University College London.

The authors declare no conflict of interest.

[1]L.J.V. and B.L.W. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: michael.gazzaniga@psych.ucsb.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1706693114/-/DCSupplemental.

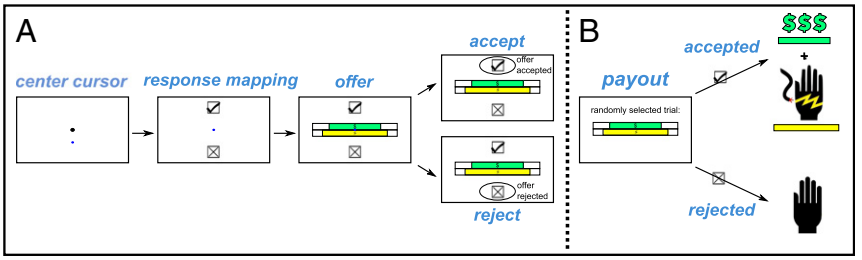

Fig. 1. (*A*) Experimental design. After centering the cursor and noting the response mapping, subjects made a reject/accept decision for a reward/harm combination depicted by the length of the green (reward) and yellow (harm) bars. (*B*) Payout trials were randomly selected and subjects received the indicated amount of reward and harm, if they had accepted the given trial.

resulting in the allocation of electric shocks and monetary reward for the self and the next participant.

These conditions lead to clear predictions regarding egoistic and altruistic behavior. Relative to the self condition, egoistic behavior should result in accepting fewer offers in the harm-for-self and more offers in the harm-for-other condition—and vice versa for altruistic behavior (Fig. 2*A*). Additionally, condition-specific reaction times (RTs) were compared to determine whether they might engage different computational processes. Two hypotheses could be tested: (*i*) Computational costs depend on the number of persons involved in the decision (a decision for two takes longer than for one) and (*ii*) computational costs depend on identity of the person(s) receiving shock or reward (deciding if another person will receive the shock takes longer than deciding if self receives the shock).

Results

Decision Frequencies. To assess general tendencies toward egoistic and altruistic behavior, we first investigated the acceptance rates of offers, which significantly differed across conditions [$F_{(3,62)} = 21.762$, $P < 0.001$]. Directly comparing conditions, we found that participants accepted significantly fewer offers in the harm-for-other condition compared with the self condition [$t_{(62)} = 3.186$, $P = 0.002$], replicating hyperaltruism as introduced by Crockett et al. (3) (Figs. 2*B* and 3*A*). That is, participants were more willing to forego reward to spare the other from harm than to spare themselves from harm. However, we also found evidence for a pronounced egoistic bias. Participants accepted significantly fewer offers in the harm-for-self condition compared with both the self [$t_{(62)} = 7.733$, $P < 0.001$] and harm-for-other [$t_{(62)} = 3.662$, $P < 0.001$] conditions. Thus, participants were willing to suffer some degree of shock to financially benefit another person. However, they caused significantly more physical harm to another person when selecting monetary reward for themselves compared with the physical harm they were willing to undergo to gain monetary reward for another person [$t_{(62)} = 3.662$, $P < 0.001$]. Hence, the magnitude of egoistic harm avoidance outweighed altruistic tendencies.

In addition, harm-for-other and other conditions showed no significant difference in acceptance rate [$t_{(62)} = 1.917$, $P = 0.06$]. Therefore, when harming others, whether the recipient of the reward was the self or the other did nor primarily influence decisions.

Similar acceptance rates between conditions do not necessarily lead to comparable amounts of accumulated reward and harm in different conditions. However, when summing the amount of

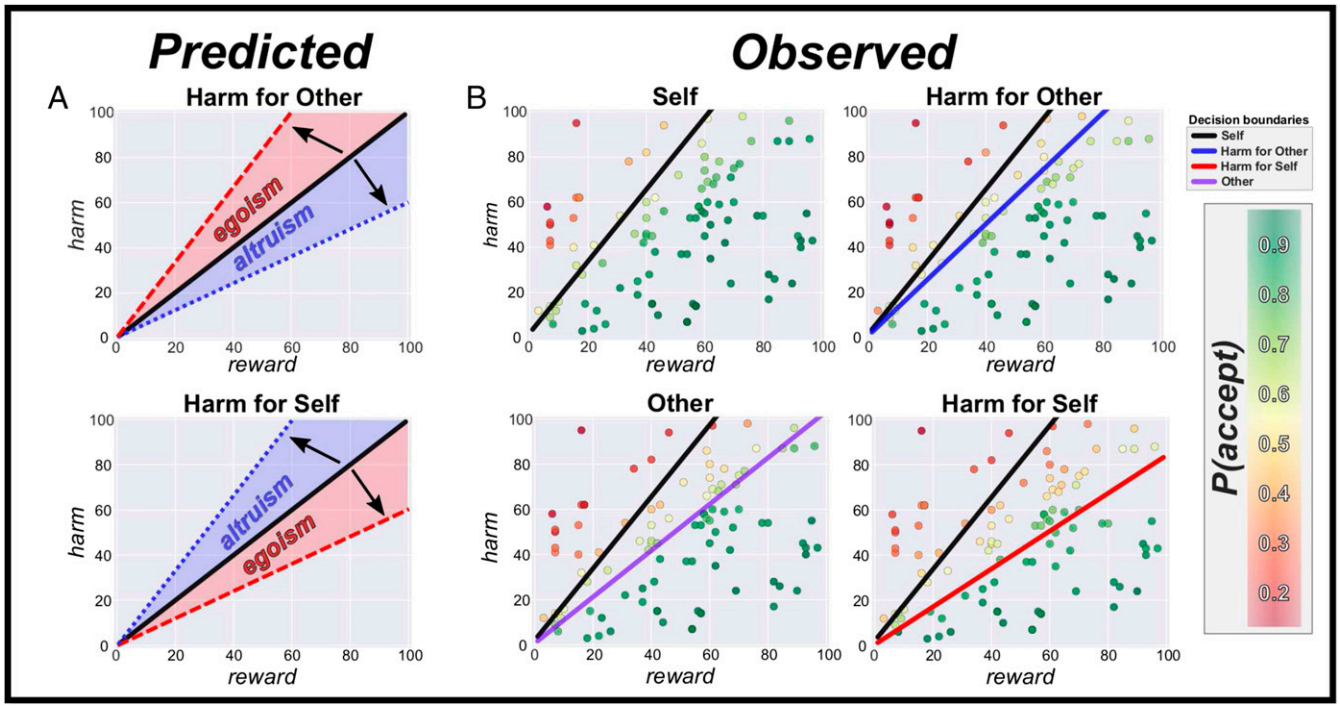

Fig. 2. (*A*) Predictions. Egoistic decision making would result in an increase in accepted trials in the harm-for-other and a decrease in accepted trials in the harm-for-self condition, whereas the opposite was predicted for altruistic behavior. (*B*) Results. The decrease in accepted offers in the harm-for-other condition replicated the notion of hyperaltruism (3, 4). By contrast, decreased willingness to accept offers in the harm-for-self condition is in line with egoistically motivated behavior. Decision boundaries depict the turning point of rejecting an offer (derived from mixed logistic regression, red dots) to accepting an offer (green dots). The displayed dots represent the reward/harm offers presented to the subjects and their mean response to a given offer by our cohort.

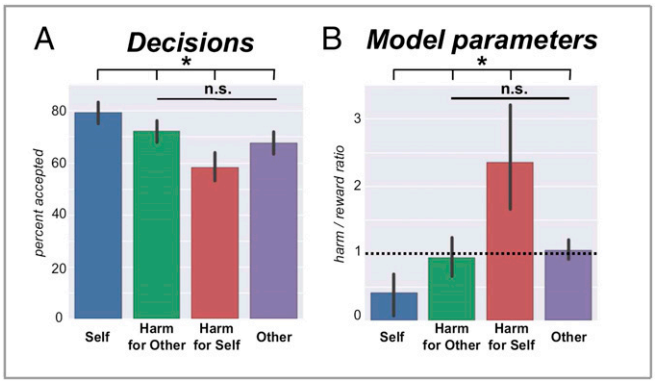

Fig. 3. (*A*) Decisions. The number of accepted trials significantly differed across conditions [$F_{(3)} = 21.762$, $P < 0.001$], showing a significantly lower acceptance rate for the harm-for-other compared with the self condition, replicating the notion of hyperaltruism. At the same time, acceptance rate in the harm-for-self condition was significantly lower compared with all other conditions, highlighting egoistic tendencies. (*B*) Model parameters. Whereas decisions in the self condition were driven by reward rather than harm (as reflected by a harm/reward ratio <1), participants primarily based their decisions on the amount of harm rather than reward in the harm-for-self condition (as reflected by a harm/reward ratio >1).

reward and harm across accepted trials, both the accumulated reward and the accumulated harm differed across conditions [Fig. S1; sum of reward: $F_{(3,62)} = 23.606$, $P < 0.001$; sum of harm: $F_{(3,62)} = 27.762$, $P < 0.001$] in a way that mirrored the acceptance rates.

Additional factors that potentially influence egoistic or altruistic decision-making tendencies were assessed. Including participants' gender as an additional factor revealed no main effect [$F_{(1,62)} = 0.946$, $P = 0.335$] or interaction [$F_{(3,168)} = 0.076$, $P = 0.969$] for gender on acceptance rates, whereas the main effect of condition on acceptance rates [$F_{(3,186)} = 20.531$, $P < 0.001$] and all significant post hoc differences were preserved (all $P < 0.05$). Accordingly, including participants' age as an additional factor did not result in a significant main effect for age [$F_{(3,168)} = 0.032$, $P = 0.860$] or age × condition interaction [$F_{(3,168)} = 0.152$, $P = 0.211$]. Similarly, including the order of conditions did not affect the difference in acceptance rates between conditions, as suggested by the absence of a main effect order [$F_{(3,62)} = 0.867$, $P = 0.463$] or order × condition interaction [$F_{(9,186)} = 1.384$, $P = 0.198$] when including the order of conditions as an additional factor in the ANOVA.

Because the timing of the payout of monetary reward and electric shocks might influence participants' decisions (e.g., learning or habituation; ref. 12), two different payout schedules were compared between subjects (intermittent payout: $n = 31$, end-of-experiment payout: $n = 32$; see *Materials and Methods* for further details). Including the payout schedule (intermittent or end of experiment) as a between-subjects factor did not show a main effect [$F_{(3,186)} = 0.006$, $P = 0.936$] and no payout schedule × condition interaction [$F_{(3,186)} = 0.385$, $P = 0.764$] was evident, whereas a significant effect of condition was still observed [$F_{(3,62)} = 21.546$, $P < 0.001$]. Hence, participants' decisions did not differ when administering payout intermittently (every 20 trials) compared with performing the payout for all randomly selected trials at the end of the experiment. Of note, the number of payout trials and the selection procedure to identify payout trials were the same in both versions of the experiment.

Mixed Logistic Regression Models of Moral Decision Making. To evaluate the individual sensitivities to reward and harm for self and other, the likelihood of accepting a given offer was modeled using a mixed logistic regression framework. In accordance with Amemori and Graybiel (10), we adopted the conditional logit model as a framework for modeling discrete choices. The models included the fixed-effect factors reward, harm, and intercept, as well as a random effect for subject and converged with significant fixed effects for all factors (all $P < 0.001$). The computational models allowed us to disentangle how the magnitude of offered reward and harm drove decision making across different conditions, as reflected by the ratio of harm/reward parameter estimates (see *Materials and Methods* for further details).

The ratio of harm/reward parameter estimates differed significantly across conditions [$F_{(3,62)} = 9.496$, $P < 0.001$], with post hoc t tests indicating significant pairwise differences for all conditions (all $P < 0.011$) except between the harm-for-other and other conditions [$t_{(62)} = 0.767$, $P = 0.446$].

In the self condition, participants were influenced by reward more than harm, indexed by a harm/reward parameter ratio significantly smaller than 1 (Fig. 3*B*). A harm/reward ratio of ∼1 indicated a balanced sensitivity to harm and reward in the harm-for-other and other conditions. In stark contrast, choices in the harm-for-self condition were primarily driven by avoiding harm, reflected by a significantly higher harm/reward ratio, suggesting egoistic motivations.

We then compared the consistency of how participants weighted harm and reward across conditions. Significant intercorrelations were apparent between the self, harm-for-other, and other conditions. However, the parameters estimated for the harm-for-self condition did not correlate with parameters from any other condition (Fig. 4*A*). In other words, subjects that were more altruistic in the harm-for-other condition were not necessarily more willing to suffer harm to benefit the other in the harm-for-self condition.

Representational Similarity Analysis. A model-independent analysis of condition-specific decision patterns was performed using representational similarity analysis (RSA) (13, 14). This model-free approach results in a multivariate estimate of whether similar offers were accepted or rejected across conditions. Whereas similar acceptance rates can in principle result from accepting different sets of offers across conditions, dissimilarity of decision patterns readily reflects such differences.

Whereas no difference was observed between the self, harm-for-other, and other conditions (all $P > 0.1$), the decision pattern in the harm-for-self condition was significantly different compared with all other conditions (all $P < 0.001$, Fig. 4*B*). In other words, a different set of offers was accepted/rejected in the harm-for-self condition compared with all other conditions, suggesting that cost–benefit trade-off decisions were derived from distinct processes and criteria in the harm-for-self condition.

Reaction Times. The comparison of mean reaction times (RTs) across conditions may help to disambiguate two competing hypotheses: (*i*) Computational costs depend on the number of persons, i.e., they are greater and hence take longer when making decisions for two persons rather than one; alternatively (*ii*) computational costs depend on the identity of the person(s) receiving the shock or alternatively, the reward. For example, they are greater and hence take longer when another person receives the shock compared with when self receives the shock.

RTs significantly differed across conditions [$F_{(3,62)} = 4.549$, $P = 0.004$; Fig. 5*A*]. Interestingly, the RT in the self condition was significantly shorter compared with all other conditions (all $P < 0.001$), whereas no significant differences were evident between all other conditions (all $P > 0.828$). In other words, reasoning about another person significantly increased the RTs. Because the RT in the "other" condition was comparable to the harm-for-other and harm-for-self condition, representing two distinct recipients simultaneously does not further increase computation time compared with reasoning about the other alone.

When differentiating RTs between accepted or rejected trials (Fig. 5*B*), we observed that decisions were significantly faster for accepted than rejected trials, for all conditions but the harm-for-

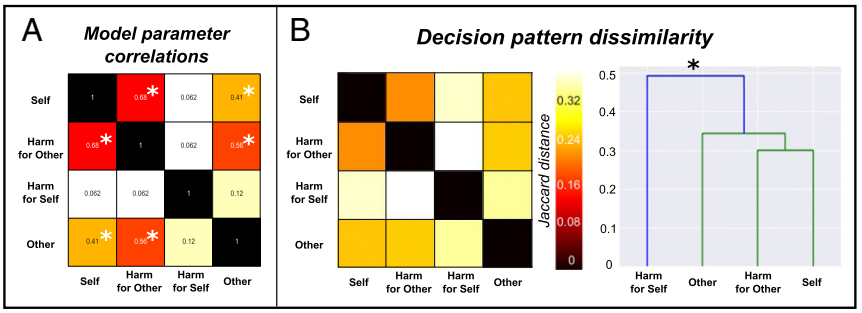

Fig. 4. (*A*) Model parameter correlations. Correlating model parameters (reward/harm ratio) across conditions revealed significant intercorrelations for all conditions except the harm-for-self condition. Hence, participants who were most prone to spare others from harm at the cost of minimizing their own reward, were not necessarily most willing to harm themselves to reward others, underlining the modality dependence of altruism. (*B*) Decision pattern dissimilarity. A completely independent, multivariate analyses of response patterns across conditions revealed a significantly dissimilar decision pattern in the harm-for-self condition compared with all other conditions, highlighting that allocating harm to oneself, although benefiting others, is governed by profoundly different mechanisms compared with decision making in all other conditions.

self condition [decision × condition interaction: $F_{(3,168)} = 7.636$, $P < 0.001$, post hoc t tests for decision: all $P < 0.001$, harm-for-self condition: $t_{(62)} = 1.157$, $P = 0.250$]. Hence, the difference in RTs for accepted and rejected trials supports the notion that different mechanisms were influencing decision making in the harm-for-self condition (Fig. 5*B*).

Discussion

Our findings revealed that individuals make both altruistic as well as egoistic choices, depending on the modality of the consequence for the self. Participants were willing to forego monetary reward to spare others from physical harm. However, they were not eager to inflict harm on themselves to secure rewards for the other. Even participants who were less harm averse in the self condition (i.e., accepting most trade offs) were hesitant to accept the same offers when another person received the resulting reward. Importantly, modality-dependent altruism and egoism were uncorrelated within individuals. For example, a participant might forego a large amount of reward to spare someone else from harm—but not suffer even modest harm for someone else's benefit.

Taken together, our current results are inconsistent both with economic models of pure self-interest (1, 2) and with universal altruism. Replicating previous findings, our data support the notion of hyperaltruism when decisions concern harm for others (3, 4); however, the same is not true for decisions involving harm for self. Because the modality of the consequences for the self is so crucial, the capacity for genuine altruism is qualified by an egoistic bias. In the intersubject conditions, the hyperaltruism effect (i.e., avoid harming others) was outweighed by participants' egoistic harm avoidance. The emerging behavioral patterns are best described as modality-dependent egoistically biased altruism.

One explanation for the modality specificity of altruism may lie in the fact that distinct neural circuitry facilitates the representation and integration of reward and harm (6, 10, 11) in humans and other animals. From an evolutionary perspective, distinct neural processing of possible rewards and harms may be explained by the different (social) contexts in which trade-off decisions were encountered. The distribution of rewards might primarily occur in positive situations, e.g., distributing the outcome of a successful collaboration like hunting in a group (15). In such circumstances, the willingness to reduce personal benefit (reward) by sharing with others may facilitate group cohesion (16, 17). Indeed, the tendency to help others without immediate benefit for the self emerges early in human development (18). By contrast, making decisions involving harm might more typically arise in negative contexts, for example, during life-threatening conflicts. Here, reducing harm to the self may be strongly prioritized over potential benefit for others (19). In other words, the moral interchangeability of reward and harm may be a modern concept, appealing to us on grounds of philosophical and economic consistency. From an evolutionary standpoint, however, the moral evaluation of trade-off decisions may have been highly modality dependent: Failing to benefit others and actively harming them were not the same thing.

Consistent with this conjecture, evidence for distinct neural computations of harm for self and others has recently been reported by Crockett et al., who showed that neural activity in lateral prefrontal cortex specifically encodes profit obtained by harming others, but not harming oneself (20). In addition, functional connectivity between lateral prefrontal cortex and dorsal striatum mediated a shift in the reward-related response of the dorsal striatum to monetary gain. These results suggest

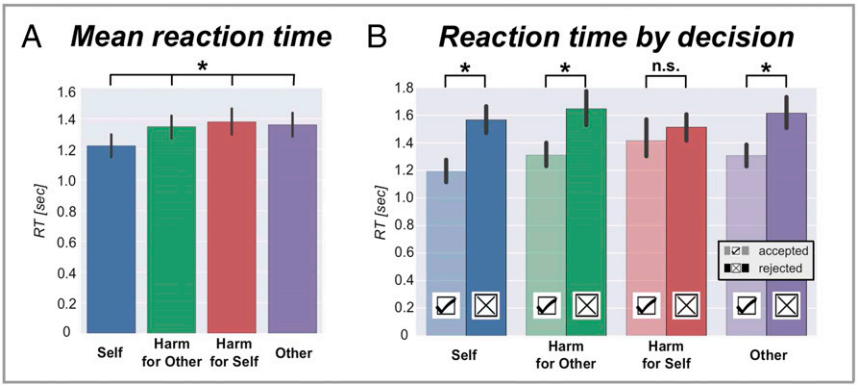

Fig. 5. Reaction time differences between conditions. (*A*) RTs significantly differed across conditions, with subjects taking more time in any condition involving another person compared with the self condition. Of note, RTs did not differ between the conditions involving another person (harm for other, harm for self, and other). (*B*) When differentiating RTs by decisions, accepted trials were found to feature significantly lower reaction times compared with rejected trials for all but the harm-for-self condition, where no significant difference was observed. This systematic difference further supports differential mechanistic underpinnings for decision making in the harm-for-self condition.

that the neural mechanisms of moral decision making are sensitive to the identity of the individual who receives painful electric shocks (20, 21). Importantly, Crockett et al. (20) focused on moral decisions involving harming others for personal gain but did not address decisions involving harming oneself to reward others. Therefore, future neuroscientific work is needed to further characterize the neural underpinnings of recipient- and modality-specific moral decision making.

An alternative explanation of modality specificity in altruism stems from the uncertainty inherent in estimating how others experience harmful events. Of note, compared with the self condition, participants accepted fewer offers in the "other" condition (Fig. 3*A*), suggesting a priority to reduce harm rather than maximize gain for others. However, it is important to note that uncertainty alone cannot explain hyperaltruism. Without a tendency to care for the well-being of others, uncertainty about their subjective experience of harm would not preclude participants from shocking others for monetary gain.

Appropriate interpretations of empirical findings addressing egoistic and altruistic tendencies in moral decision making are contingent on the specific experimental context. We here used a minimal moral context, providing no information about the other person and excluding opportunities for reciprocity. One might argue that such a minimal context may be limited in its applicability to real-world interactions, where the identity of the other person is typically known and reciprocity is possible. However, the increasing relevance of social media interactions occurring in such minimal moral contexts emphasizes the importance of understanding their normative principles. Moreover, only a minimal moral context allows for an evaluation of egoistic and altruistic tendencies, unbiased by the identity of the other person and idiosyncratic strategies accounting for reciprocity. Therefore, the minimal moral context provides an essential baseline comparison for future research in richer social contexts, for example, when group membership or behavioral dispositions of the other person are known. Moreover, personal and demographic characteristics of participants may impact on moral decision-making behavior. In particular, developmental changes during adolescence and beyond in brain circuitry involved in moral and reward-based decision making might result in age-dependent differences in behavior (for example, see ref. 22). Although age did not impact on acceptance rates in our primarily late adolescent cohort, we cannot be certain that the observed behavioral patterns are identical in other age ranges. Therefore, the effects of age and other personal and demographic characteristics should be addressed by future research.

Ultimately, our results should not be interpreted as evidence contradicting the notion that human beings can be genuinely altruistic. In fact, participants still accepted a considerable fraction of offers in the harm-for-self condition, in the absence of any personal gain. The fact that this genuine altruism was outweighed by participants' willingness to impose harm on others for their own benefit emphasizes rather the crucial modality dependence of altruism. From a practical perspective, this modality dependence highlights that the framing of trade-off decisions may critically bias our responses. For example, the perception of policy making involving a trade off between the interests of different groups, may profoundly differ, depending on whether consequences are presented as a forgoing benefit to spare others from harm versus accepting harm for others' benefit.

In conclusion, the present work suggests that modality-specific processing of rewards and harms may have a strong impact on individuals' propensities to exhibit egoistic and altruistic behavior. Therefore, it is ultimately short sighted to envision humans either as categorically egoistic or altruistic. Rather, we propose an integrated framework of modality-dependent moral decision making, recognizing that the possibility of genuine altruism is embedded in context-dependent egoistic bias, construing decision makers as egoistically biased altruists.

Materials and Methods

Participants and Procedure. Sixty-three healthy subjects were recruited at the University of California, Santa Barbara (UCSB) and they provided written informed consent (41 female, mean age: 19.7 ± 2.7 y). This study was approved by the local ethics committee (UCSB Institutional Review Board no. 34–16-0964). Participants were given complete and comprehensive instructions regarding all aspects of the task both verbally and with accompanying visual demonstrations, and they practiced using the joystick to indicate response options in a trial version of the self condition.

Before completing the task, participants were informed about the implications of decisions for self and other. In particular, participants were informed that they would receive (*i*) both monetary reward and electric shocks resulting from their own choices, as well as (*ii*) monetary reward and electric shocks resulting from the previous participant's decisions (in the harm-for-other and "other" conditions). Hence, monetary rewards accumulated by the previous participant in the harm-for-self and "other" conditions, as well as harm accumulated in the harm-for-other and "other" conditions were administered to the present participant. Importantly, the administration of both reward and harm (from both decisions of the present participant and those of the past participant) took place after the participants had fully completed all conditions of the task. This experimental design avoids any influence of the trade-off decisions of the previous participant on the trade-off decisions of the present participant.

Participants were also informed that their decisions in respective conditions would be administered to the next participant. Importantly, participants' actual trade-off decisions were carried out, i.e., the next participant indeed received the monetary reward as allocated in the harm-for-self and "other" condition and the electric shocks as allocated in the harm-for-other and "other" condition. Participants had no personal interaction with past or future participants and no knowledge of the identity or personal characteristics of past or future participants. This limitation is crucial, given our aim of constructing a minimal moral context, which provides participants with no knowledge about the identity of the other person and affords no opportunities for reciprocity.

To ensure that electric shocks had a similar aversive value across subjects, the shock intensities used in the experiment were individualized using a standardized procedure (for example, see refs. 1, 2). Participants then completed all four conditions in a randomized order, which was counterbalanced across subjects. Each condition comprised 100 trade-off decisions with distinct combinations of reward and harm. These combinations were identical for all subjects and conditions, but were presented in different randomized orders. Apart from the introduction screen (indicating the upcoming condition), visual stimuli were identical across conditions. Every 20 trials, participants were presented a payout screen (Fig. 1*B*), indicating the reward/harm combination from a randomly selected trial. Because every offer could be selected for payout, participants were instructed to treat each offer as if it would be administered. If participants had accepted the selected payout trade off, corresponding reward and harm were administered to the appropriate target(s). If participants rejected the randomly selected trial, neither reward nor harm was administered.

The timing of the payout might have a considerable impact on participants' decisions (3). On the one hand, administering reward and harm at the end of the experiment may result in reduced task engagement and render participants skeptical about the veridical nature of the payout. On the other hand, intermittent payout administration may dynamically impact subjective reward/harm trade offs across blocks of trials and thereby create a confound across conditions. To rule out that either of these possibilities considerably impacted on our findings, we administered all payouts at the end of the experiment for 31 participants (2), whereas the other 32 participants received intermittent payouts after every 20 trials. In either case, participants received the actual harm and reward as determined by the previous participant after finishing all four conditions.

Delivery of Painful Shocks. Electric shock configuration and pain thresholding were adapted from Seymour et al. (21). Two adhesive electrodes were placed on the back of the left hand with gel applied to increase skin conductivity. Electric current was delivered to cause an aversive sensation that becomes increasingly painful as the current is increased. For each electric shock, direct current was administered for a duration of 1 s at a frequency of 100 Hz with a 2-ms square waveform. Shocks were administered using a PowerLab 26T device with the LabChart software package (ADInstruments Ltd), which is US Food and Drug Administration approved for human and clinical use.

Shock intensities were individualized with the intention of producing comparable subjective aversiveness across participants. Shocks were administered starting at extremely low currents and increasing at graduated intensities until tolerance was reached. No shocks above this subjectively determined tolerance level were administered. Next, 14 subtolerance intensities were randomly defined and delivered to the participants. For these shocks, participants indicated the subjective pain level using a visual analog scale (VAS) ranging from 0 to 10. A sigmoid function was subsequently fitted to the VAS response values, describing the relationship between shock intensity and pain perception for each individual. Based on this sigmoid function, an intensity equivalent to a pain rating of 8 out of 10 on the VAS was fixed as the maximum shock intensity.

Trade-Off Decision Making. In the present study, we adapted for human subjects the approach–avoidance decision task initially introduced by Amemori and Graybiel for macaques (10, 11), using PsychoPy (23).

On each trial, participants used a custom-built joystick to indicate their decision to accept or reject the offer (for task, see Fig. 1). First, to indicate readiness and to initiate the trial, participants had to move the cursor to the center of the screen. Once the cursor was centered, the response options indicating acceptance and rejection of the given offer were displayed at the top and bottom of the screen. One second later, the offer was presented in the center of the screen in the form of two horizontal bars, with the top green bar indicating the available reward and the bottom yellow bar indicating the available harm. The magnitude of reward and harm was indicated by the length of the respective bars, which independently varied between trials. The maximum amounts of reward ($2) and harm (shock intensity equivalent to 8 VAS) were always indicated by black bounding boxes around the horizontal bars. Subjects were instructed to reject or accept the offer as quickly as possible by using the joystick to move the cursor to the respective response options. The location of the options, i.e., whether the joystick had to be moved toward or away from the subject to accept an offer, was randomized across trials. After indication of their response, a visual feedback was displayed, indicating the acceptance or rejection of the offer.

A randomly selected payout trial was indicated after completion of every 20 trials (Fig. 1*B*). If the randomly selected trial was an offer accepted by the subject, both reward and harm were administered, whereas neither reward nor harm was allocated for rejected offers. For half of the subjects ($n = 31$), the indicated reward and harm were administered immediately, whereas for the remaining participants ($n = 32$), payouts from all conditions were allocated at the end of the experiment. All 100 trials of a given condition were completed before moving on to the next condition, with a randomized order of conditions that was counterbalanced across subjects.

Data Analysis. Data were analyzed using three complementary approaches: (*i*) comparison of response frequencies and reaction times, (*ii*) mixed-effects modeling of trade-off decision making, and (*iii*) RSA comparing decision making patterns across conditions.

Acceptance rates were assessed using one-way analysis of variance (ANOVA) including condition as a factor, using ezANOVA-package for R (24). Post hoc *t* tests were subsequently conducted to compare condition-specific differences. Similarly, reaction times between conditions were compared using the same analytical framework. Moreover, reaction times were separately analyzed for accepted and rejected trials.

To address the respective impact of the reward and harm magnitude on condition-specific decision making, reject/accept decisions were modeled as a function of reward and harm using mixed logistic regressions as implemented in lme4-package for R. In accordance with Amemori and Graybiel (10), we adopted the conditional logit model, which is a popular framework for modeling discrete choices. We used the winning model of the model comparison (4), which characterizes decisions as a linear combination of the available reward x and harm y:

$$f(x, y) = a1\ x + a2\ y + a3,$$

with a reward coefficient *a1*, a harm coefficient *a2*, and an intercept *a3*. Whereas *a1* and *a2* reflect how participants weighted the magnitude of the offered reward and harm, *a3* allowed for flexibility in participants' motivations to maximize reward or avoid harm overall. Separate models were fitted for each condition, including fixed effects for reward and harm alongside a random effect for subject. Resulting parameter estimates were compared across conditions using an ANOVA in line with the analysis of acceptance frequencies. In Figs. 3 and 4, the ratios of parameter estimates a2/a1 across conditions are depicted, representing the relative influence of harm and reward.

A model-independent multivariate analysis of condition-specific decision patterns was performed using RSA (13, 14). The goal of this analysis was to compare the multivariate distance of response patterns, i.e., all 100 binary choices (reject/accept) within each condition. This comparison is possible because the same set of offers was presented in each condition. We used Jaccard distance as a metric for the binary choices. The benefit of analyzing decision patterns using RSA lies in the fact that information on the parameter space is readily represented. For example, similar acceptance rates between conditions may result from accepting highly different sets of offers, reflecting different decision-making principles. Whereas these differences would also be partially reflected by model parameter estimates from the mixed-logistic regression, our RSA approach complements explicit modeling of the parameters. In principle, RSA can detect differences in acceptance patterns that are independent of model assumptions. By combining RSA with mixed-effects logistic regression, we gain greater sensitivity to divergences in participant response across conditions. Similarly, multivariate analysis of distances in response times offers insight into whether or not overlapping sets of trials drive similarity or dissimilarity across conditions in mean response time. The statistical significance of each pairwise distance was assessed via one-sample *t* test.

ACKNOWLEDGMENTS. The authors thank Ken-ichi Amemori, Ann M. Graybiel, Nathanial D. Daw, and John Tooby for helpful discussions and constructive criticism of our work. This work was supported by the Institute for Collaborative Biotechnologies through Contract W911NF-09-D-0001 and Grant W911NF-16-1-0474 from the US Army Research Office; and by the SAGE Center for the Study of the Mind, UCSB.

1. Engel C (2011) Dictator games: A meta study. *Exp Econ* 14:583–610.
2. Henrich J, et al. (2010) Markets, religion, community size, and the evolution of fairness and punishment. *Science* 327:1480–1484.
3. Crockett MJ, Kurth-Nelson Z, Siegel JZ, Dayan P, Dolan RJ (2014) Harm to others outweighs harm to self in moral decision making. *Proc Natl Acad Sci USA* 111: 17320–17325.
4. Crockett MJ, et al. (2015) Dissociable effects of serotonin and dopamine on the valuation of harm in moral decision making. *Curr Biol* 25:1852–1859.
5. FeldmanHall O, et al. (2012) Differential neural circuitry and self-interest in real vs hypothetical moral decisions. *Soc Cogn Affect Neurosci* 7:743–751.
6. Palminteri S, Pessiglione M (2017) Opponent brain systems for reward and punishment learning: Causal evidence from drug and lesion studies in humans. *Decision Neuroscience, An Integrated Approach*, eds Dreher J-C and Trembly L (Academic, Cambridge, MA), Chapter 23, pp 291–303.
7. Kamm F (1985) Supererogation and obligation. *J Philos* 82:118–138.
8. Urmson J (1958) Saints and heroes. *Essays in Moral Philosophy*, ed Melden A (University of Washington Press, Seattle), pp 198–216.
9. Chisholm R (1963) Supererogation and offence: A conceptual scheme for ethics. *Ratio* 5:1–14.
10. Amemori K, Graybiel AM (2012) Localized microstimulation of primate pregenual cingulate cortex induces negative decision-making. *Nat Neurosci* 15:776–785.
11. Amemori K, Amemori S, Graybiel AM (2015) Motivation and affective judgments differentially recruit neurons in the primate dorsolateral prefrontal and anterior cingulate cortex. *J Neurosci* 35:1939–1953.
12. Dayan P, Daw ND (2008) Decision theory, reinforcement learning, and the brain. *Cogn Affect Behav Neurosci* 8:429–453.
13. Kriegeskorte N, Mur M, Bandettini P (2008) Representational similarity analysis: Connecting the branches of systems neuroscience. *Front Syst Neurosci* 2:4.
14. Kriegeskorte N, Diedrichsen J (2016) Inferring brain-computational mechanisms with models of activity measurements. *Philos Trans R Soc B Biol Sci*:1–19.
15. Smith EA (2004) Why do good hunters have higher reproductive success? *Hum Nat* 15: 343–364.
16. Rand DG, Greene JD, Nowak MA (2012) Spontaneous giving and calculated greed. *Nature* 489:427–430.
17. Axelrod R, Hamilton WD (1981) The evolution of cooperation. *Science* 211:1390–1396.
18. Warneken F, Tomasello M (2006) Altruistic helping in human infants and young chimpanzees. *Science* 311:1301–1303.
19. Tomasello M, Melis AP, Tennie C, Wyman E, Herrmann E (2012) Two key steps in the evolution of human cooperation. *Curr Anthropol* 53:673–692.
20. Crockett MJ, Siegel JZ, Kurth-Nelson Z, Dayan P, Dolan RJ (2017) Moral transgressions corrupt neural representations of value. *Nat Neurosci* 20:879–885.
21. Seymour B, Daw ND, Roiser JP, Dayan P, Dolan R (2012) Serotonin selectively modulates reward value in human decision-making. *J Neurosci* 32:5833–5842.
22. Blakemore S-J (2008) The social brain in adolescence. *Nat Rev Neurosci* 9:267–277.
23. Peirce JW (2007) PsychoPy-Psychophysics software in Python. *J Neurosci Methods* 162: 8–13.
24. Lawrence MA (2015) ez: Easy Analysis and Visualization of Factorial Experiments. R package version 4.3.