Inference of trustworthiness from intuitive moral judgments

Jim A.C. Everett[1], David Pizarro[2], and M.J. Crockett[1]

[1] Department of Experimental Psychology, University of Oxford, 9 South Parks Road, Oxford, OX1 3UD, UK

[2] Department of Psychology, Cornell University, 234 Uris Hall, Ithaca, NY 14853, USA

Corresponding author:
Jim A.C. Everett
Department of Experimental Psychology
University of Oxford
9 South Parks Road
Oxford, OX1 3UD, UK
E-mail: jim.everett@psy.ox.ac.uk

Keywords: morality, intuition, partner choice, deontological, utilitarian

Abstract

Moral judgments play a critical role in motivating and enforcing human cooperation, and research on the proximate mechanisms of moral judgments highlights the importance of intuitive, automatic processes in forming such judgments. Intuitive moral judgments often share characteristics with deontological theories in normative ethics, which argue that certain acts (such as killing) are absolutely wrong, regardless of their consequences. Why do moral intuitions typically follow deontological prescriptions, as opposed to those of other ethical theories? Here we test a functional explanation for this phenomenon by investigating whether agents who express deontological moral judgments are more valued as social partners. Across five studies we show that people who make characteristically deontological judgments are preferred as social partners, perceived as more moral and trustworthy, and are trusted more in economic games. These findings provide empirical support for a partner choice account of why intuitive moral judgments often align with deontological theories.

Inference of trustworthiness from intuitive moral judgments

Moral judgments are stitched into the fabric of human nature, steering us toward cooperation and away from exploitation. Research has highlighted a central role for intuitive, automatic cognitive processes in forming such judgments (Greene, 2014; Haidt, 2001). And intriguingly, such intuitive or automatic (and their counterpart, deliberate, or controlled) processes in moral judgment have been argued to align with two opposing perspectives dominating ethical discussion: *deontology* (Kant, 1797/2002; Scanlon, 1998) and *consequentialism* (Bentham, 1879/1983; Mill, 1863). Consequentialist theories like Utilitarianism focus solely on the impartial maximization of aggregate welfare as the criterion for a moral act: "actions are right in proportion as they tend to promote happiness, wrong as they tend to produce the reverse of happiness" (Mill, 1863). In contrast, deontological theories hold that the rightness or wrongness of an action is not entirely dependent on its consequences, and instead typically focus on notions of duties, rights, and obligations. Inspired by this debate, researchers have explored how people respond to moral dilemmas with two options that align with either consequentialism (e.g., sacrificing a single innocent life to save many others) or deontology (e.g. refusing to sacrifice an innocent life regardless of the consequences). A wealth of behavioral and neurobiological evidence has shown that participants' intuitive and automatic judgments tend to be characteristically deontological, while characteristically consequentialist judgments are often the result of slow, deliberative cognitive processes (Greene, 2007; Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008; Koenigs et al., 2007). But it remains unclear *why* moral intuitions more often align with deontology, rather than the option that would maximize the utility of outcomes. This question is even more puzzling when one considers that deontological judgments might promote inferior social outcomes (Greene, 2014).

One approach to explaining why moral intuitions often align with deontology comes from mutualistic partner choice models of the evolution of morality. These models posit a cooperation market such that agents who can be relied upon to act in a mutually beneficial way are more likely to be chosen as cooperation partners, thus increasing their own fitness (Alexander, 1987; Baumard, André, & Sperber, 2013; Krebs, 2008; Noë & Hammerstein, 1994; Trivers, 1971). People tend to select the most cooperative individuals as partners, and those who contribute less than others are gradually left out of cooperative exchanges (e.g. Barclay, 2004, 2006; Rockenbach & Milinski, 2011). To the extent that individuals who make certain types of moral judgments are favored in a cooperative market because these judgments signal a commitment to cooperation, so too will these judgments come to be favored as defaults. In other words, deontological moral intuitions may represent an evolutionarily prescribed prior that was selected for through partner choice mechanisms. Why might deontologists be preferred as social partners? Two features of deontological intuitions seem important, given their relevance for social exchange: the prohibition of certain acts or behaviours, and the expression of socially valued emotional responses.

First, deontologists' prohibition of certain acts or behaviours may serve as a relevant cue for inferring trustworthiness because the extent to which someone claims to follow rule or action-based judgments may be associated with the reliability of their moral behavior. One piece of preliminary evidence for this comes from a study showing that agents willing to punish third parties who violate fairness principles are trusted more, and actually are more trustworthy (Jordan, Bloom, & Rand, in press). Moreover, the typical deontological reason for why specific actions are wrong is that they violate duties to respect persons and honor social obligations - features that are crucial when selecting a social partner. An individual who claims that stealing is always morally wrong and believes themselves morally obligated to act in accordance with

this duty, seems much less likely to steal from me than an individual who believes that the stealing is sometimes morally acceptable depending on the consequences. Actors who express characteristically deontological judgments may therefore be preferred to those expressing consequentialist judgments because these judgments may be more reliable indicators of stable cooperative behaviour. Consistent with this, recent research has shown that, compared to people who make consequentialist arguments, people who make deontological arguments are perceived by others as less self-interested and as expressing more moral views (Kreps & Monin, 2014). And recent theoretical work has demonstrated that "cooperating without looking" – i.e., without considering the costs and benefits of cooperation – is a subgame perfect equilibrium (Hoffman, Yoeli, & Nowak, 2015). Therefore, expressing characteristically deontological judgments could constitute a behavior that enhances individual fitness in a cooperation market because these judgments are seen as reliable indicators of a specific valued behavior – cooperation.

Second, deontological judgments often align more strongly with socially valued emotional responses, such as empathy and harm aversion, than do consequentialist judgments. As some have argued, making consequentialist judgments generally involves the suppression of prepotent (deontological-leaning) emotional responses in order to reach a more calculated analysis of the consequences to be derived from various actions (Greene, 2014). Research shows that characteristically deontological judgments are positively associated with harm aversion, and negatively associated with antisocial personality traits (Bartels & Pizarro, 2011; Cushman, Gray, Gaffey, & Mendes, 2012; Kahane, Everett, Earp, Farias, & Savulescu, 2015). People who are more likely to endorse the sacrifice of one person to save many others appear also to be those people who are less averse to harming others in everyday contexts where there is no obvious greater good (Kahane et al., 2015). If prospective partners in the cooperation

market intuit this, they may prefer deontologists. In other words, expressing a deontological judgment may communicate that a person has a set of socially valued emotional responses (i.e., an aversion to directly harming others) that make them an attractive social partner. Consistent with this, recent studies have shown that individuals who made deontological decisions in moral dilemmas are rated as being more empathic and having a superior moral character compared to those who make consequentialist decisions (Uhlmann, Zhu, & Tannenbaum, 2013).

Overview

On both theoretical and empirical grounds it seems plausible that deontological moral intuitions may have been selected for through partner choice mechanisms. However, the central claims behind this account – that people who express deontological moral intuitions are perceived as more trustworthy and favored as cooperation partners – has not been empirically investigated. In this paper we fill this gap, asking first whether deontological agents are preferred as social partners, and, if so, which features of characteristically deontological moral intuitions confer greater selective social value.

In order to do so, across five experiments we examined participants' perceptions of an agent who made either characteristically deontological judgments ("killing people is just wrong, even if it has good consequences") or consequentialist judgments ("it is better to save five lives rather than one"). Following the bulk of previous research on moral intuitions, we used hypothetical sacrificial dilemmas as a way of directly pitting deontological and consequentialist intuitions against one another. In addition, we used several different dilemmas (some in which intuitions lean deontological; and some in which intuitions lean consequentialist) and complementary measures of perceived prosociality (self-reported

character ratings of morality and trust; behavioral data from economic games; and partner choice questions inspired by evolutionary biology) to test our hypotheses.

Study 1

We first investigated individuals' perceptions of agents who made either deontological or consequentialist judgments in a sacrificial dilemma that typically evokes deontological intuitions in most respondents (Cushman, Young, & Hauser, 2006; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001). Participants were presented with information about two agents who were asked to judge whether it was morally appropriate to push a man off a footbridge to stop an oncoming train from hitting five others, thereby killing him but saving the five. The consequentialist agent judged it morally appropriate to push the man, while the deontological agent did not. We then had our participants rate the morality and trustworthiness of each agent on a scale (Study 1a), play a hypothetical trust game with the agents (Study 1b), and, finally, play a trust game involving real monetary stakes with the agents (Study 1c).

Study 1a

Methods

Ethics Note. All the studies reported in this manuscript received approval from the UCL Research Ethics Committee (4418/002) and the University of Oxford Central University Research Ethics Committee (MS-IDREC-C1-2015-098).

Participants. 200 American participants (71 female; M_{age} = 34, SD = 10.69) were recruited through Amazon Mechanical Turk (MTurk), and were paid $0.80 for their time. Five participants took the survey more than once, and so were excluded from subsequent analyses (final N = 195). For all studies reported in this manuscript, we used G*Power 3.1 (Faul,

Erdfelder, Buchner, & Lang, 2009) to calculate the minimum sample size needed. This power analysis showed that a within-subjects design with a conservative small effect size (d=0.2) would require a minimum sample of 199 participants.

Design. Participants were told that they would be randomly paired with two other MTurk workers who had already completed the survey, and that they would see the other workers' judgment along with their reasons for their judgment to the following moral dilemma (the footbridge Dilemma: Foot, 1967; Thomson, 1976):

> A runaway trolley is heading down the tracks toward five workers who will all be killed if the trolley proceeds on its present course. Adam is on a footbridge over the tracks, in between the approaching trolley and the five workers. Next to him on this footbridge is a stranger who happens to be very large. The only way to save the lives of the five workers is to push this stranger off the bridge and onto the tracks below where his large body will stop the trolley. The stranger will die if Adam does this, but the five workers will be saved.

Participants were then presented with the judgments of the two other workers (i.e. "Person A said that Adam should [not] push the large man to save the five workers") and the reasons they gave for their judgment (i.e., "it is better to save five lives rather than one" vs. "killing people is just wrong, even if it has good consequences"). Half of the participants first read about the agent who endorsed pushing the large man (the consequentialist agent) then read about the agent who rejected the sacrifice (the deontological agent). This order was reversed for the other half of participants. After reading the responses of both agents, participants participants were asked to rate perceived morality (*1 = extremely immoral / bad; 7 = extremely*

moral / good) and perceived trust (*1 = extremely untrustworthy; 7 = extremely trustworthy*) of the two agents in the order in which they were presented. At the end of the study participants were asked to make their own judgment about whether they thought Adam should push the stranger (Yes or No), and how wrong it would be for Adam to push the stranger (*1 = not at all wrong, 7 = very wrong*).

Results and Discussion

In line with previous research, only a minority of participants (29%) endorsed the consequentialist option, with most participants (72%) indicating that it would be wrong to push the large man even if it would save five lives.

Because the data were non-normally distributed, we used a series of Wilcoxon Signed Rank tests. In line with predictions, the agent who gave a deontological response was perceived as being more moral (Z=-8.31, p<.001, d=1.1) and more trustworthy (Z=-8.13, p<.001, d=1.07) than the agent who gave a consequentialist response.

We next investigated whether these results merely reflected a similarity effect whereby individuals simply prefer those who make similar judgments to them. If so, the observed overall preference for deontologist agents would be a result of the fact that most participants themselves made deontological judgments in the footbridge dilemma. However, the data did not support a mere similarity effect: our results were robust to controlling for participants' own moral judgments such that participants who made a deontological judgment (the majority) strongly preferred a deontological agent, while participants who made a consequentialist judgment (the minority) showed no preference between the two agents either on perceived morality (Z=-0.78, p=.44, d=0.07) or trustworthiness (Z=-0.24, p=.81, d=0.04). We also conducted analyses looking at only people who either said that the consequentialist was "not

at all wrong" or "very wrong" (points 1 and 7 on the wrongness scale respectively). Again, while people who said the consequentialist action was "very wrong" thought the deontological agent more moral and trustworthy, those who said the consequentialist action was "not at all wrong" perceived no difference between the two agents in morality (Z=-0.65, p=.52) or trustworthiness (Z=-0.95, p=.34). In the interest of brevity we only report the main control analyses in the main text. Please see the Supplementary Materials for full Ms, SDs, and significance tests for analyses broken down by participant judgment for all the studies reported in this paper.

Study 1b

Method

Participants. 360 American participants (114 female; M_{age}=37, SD=11.97) were recruited through MTurk and paid $1.00. Participants were excluded from analyses if they did not complete the survey fully (N=3), failed simple comprehension checks involving the structure of the trust game (N=74), took the survey more than once (N=10), or had taken part in Study 1a (N=54), leaving a final sample of 219 participants.

Design. Study 1b used the same basic design as Study 1a, but with the addition of a new dependent measure—in addition to the character and trustworthiness ratings, participants played a *trust game* (TG) with the deontological and consequentialist agents. In a typical TG, there are two players: an investor and a trustee. The investor is given some money and told that they may send a proportion (from zero to the full amount) of this money to the trustee, and that the experimenter will multiply the money sent by some amount. Once the trustee receives the money, they are told that they may send back a portion of it to the investor, again ranging from zero to the full amount. In this study (and all subsequent ones), participants always played the

role of investor. The amount of money participants transferred to the agent (from ($0.00 to $0.30) was used as an indicator of trustworthiness, as was how much money they believed they would receive back from the agent (0% - 100%). Finally, as an explicit measure of partner choice we asked participants to indicate "If you had a choice and could select one of the other people from earlier in this study (Person A or Person B), which one would you rather have in this game with you? Would you rather play with Person A or Person B?"

Results and Discussion

As in previous research, most participants (73%) endorsed the characteristically deontological judgment and indicated that it would be wrong to push the large man even if it would save five lives. And consistent with Study 1a, participants reported the deontological agent to be more moral (Z=-8.90, p<.001, d=1.10 Fig. 1A) and more trustworthy (Z=-8.70, p < .001, d=1.05; Fig. 1B; see Supplementary Tables 1-2 for Ms and SDs) than the consequentialist agent. Furthermore, participants transferred more money to deontological agents (63% of endowment) than consequentialist agents (40%: Z=-7.74, p<.001, d=0.73; Fig. 1C) and believed that deontological agents (43%) would return more money than consequentialist agents (24%: Z=-7.19, p<.001, d=0.73; Fig. 1D). As predicted, there was also a significant difference in preferred partner (p<.001), with 80% of participants preferring to play a trust game with an agent who made a deontological judgment.

Again, these results held when controlling for participants' own judgments. While deontological participants showed a strong preference for the deontological agent, consequentialist participants reported no difference in perceived morality (Z=-0.03, p=.98, d=0.07) or trust (Z=-0.27, p=.79, d=0.03,) of the two agents, and transferred the same amount to both agents (Z=-1.49, p=.14, d=0.27). Moreover, consequentialist participants actually

predicted that a deontologist agent would return *more* money than would a fellow consequentialist (Z =-2.02, p =.04, d=0.34). As in Study 1a, we next focused at participants who gave extreme responses on the wrongness scale. Again, while people who said the consequentialist action was "very wrong" transferred more to the deontological agent, those few who said the consequentialist action was "not at all wrong" showed no difference in the amount of money they transferred (Z=-0.03, p=.98) or predicted returns (Z=-0.06, p=.96); though they did perceive the utilitarian agent to be more moral (Z =-2.12, p=.03). Most convincingly, a full 35% of participants who thought the consequentialist action was "not at all wrong" indicated they would have preferred to play with a deontologist partner (compared to just 6% of those who said it was "very wrong" saying they would prefer a consequentialist agent). As in Study 1a, these results cannot be explained simply through participants distrusting those who disagree with them on moral issues - the majority of participants preferred the deontologist agent, and even the minority that endorsed a consequentialist position showed no consistent preference for either agent.

Study 1c

Two potential limitations of the studies thus far, however, deserve consideration. First, it could be argued that because the responses in the TG in Studies 1a-b were only hypothetical, it is not clear whether participants would show differential responses to a consequentialist or deontological target in a TG with real monetary incentives. Second, it could be argued that demand characteristics might play a role given the within-subjects design of the previous studies - people might have believed that the experimenters expected them to respond differently to the consequentialist and deontological target because we asked them about both. To address these potential concerns, in Study 1c we sought to replicate the findings of Studies

1a and 1b, and extend them by having participants play an incentivized trust game in a between-subjects design.

Method

Participants. 190 American participants (55 female; M_{age}=34, SD=11.11) were recruited through MTurk and paid $0.50. In this and subsequent studies, participants could only complete the survey in full if they answered simple comprehension checks at the start of the survey correctly. Participants were excluded from analyses if they took the survey more than once, or had participated in one of the previous studies reported in this paper (N=46), leaving a final sample of 144 participants. Again, most participants endorsed the deontological option (72%). In this study and all others involving real TGs in the manuscript, participants were paid bonuses according to the average amount returned by actual deontologist and consequentialist agents.

Design. In Study 1c we followed the design of the two previous studies, but instead of having participants play a hypothetical TG with two agents (one agent who made a consequentialist judgment, and one who made a deontological judgment), participants played a real incentivized TG in a between-subjects design, so that they interacted only with one agent who either endorsed the deontological or consequentialist action.

Results and Discussion

Data were non-normally distributed and so we tested our hypotheses using Mann-Whitney U tests. Consistent with our predictions, participants transferred more money to deontological agents (59%) than consequentialist agents (40%: U=1952, p=.01, d=0.46) and predicted that deontological agents would return more money (28%) than consequentialist

agents (21%: U=2099, p=.05, d=0.26). As predicted, there was a significant difference in preferred partner, p<.001, with 74% of participants preferring to play a trust game with an agent who made a deontological judgment.

As before, results held when controlling for participants' own moral judgments. Deontologist participants, again, showed a strong preference for the deontologist agent, and consequentialist participants showed no preference between either agent in transfer amounts (U=160, p=.42, d=0.25) or predicted returns (U=173, p=.67, d=0.13). We were unable to conduct analyses separately for those who endorsed the end-points of the wrongness scale because only 4 participants in our sample said that the consequentialist action was "not at all wrong".

Study 2

Studies 1a-c demonstrated that people perceive agents who provide deontological responses to a sacrificial moral dilemma (compared to those who provide consequentialist responses) as more trustworthy, both in their self-reports and in their actual behavior. Yet it remains unclear whether this preference results from deontological judgments signaling a commitment to cooperation, or from consequentialist judgments indicating a reduced commitment to cooperation. These two potential explanations can be teased apart when using process dissociation, a technique that can assess the degree to which individuals' responses are driven by being high or low in deontology or by being high or low in consequentialism (Conway & Gawronski, 2013). In Study 2 we attempted to similarly tease apart these possibilities, by investigating whether it is the *presence* of deontological intuitions that is crucial for inferring trustworthiness, or the *absence* of consequentialist intuitions.

We reasoned that if deontological agents are preferred over consequentialist agents because they are perceived as more committed to social cooperation, such preferences should be lessened if consequentialist agents reported their judgments as being very difficult to make, indicating some level of commitment to cooperation (Critcher, Inbar, & Pizarro, 2013). From the process dissociation perspective (Conway & Gawronski, 2013), a person who reports that it is easy to make a characteristically consequentialist judgment can be interpreted as being high in consequentialism (because they endorsed the sacrifice) but low in deontology (because there was little decision conflict with competing deontological motives). In contrast, a person who reports it is difficult to make the consequentialist judgment can be interpreted as being high on both consequentialism (because they endorsed the sacrifice) and deontology (because there was decision conflict with simultaneous deontological intuitions to not endorse the sacrifice). To the extent that it is the presence of deontological intuitions that is crucial for inferring trustworthiness, the preference for a deontologist over a consequentialist agent should be lessened when the consequentialist agent reports difficulty in making the judgment. In those cases, the conflict is indicating that the consequentialist agent has deontological intuitions that are making their decision difficult. In contrast, to the extent that it is really the absence of consequentialist intuitions that drive the preference for deontologists, participants should prefer a deontologist who reports ease in making the decision over one who reports it as difficult – because ease of judgment indicates low consequentialist intuitions compared to deontological ones.

Method

Participants. 300 American participants (327 female; M_{age}=34, *SD*=10.86) were recruited through MTurk and paid $0.50. Participants were excluded from analyses if they took the survey more than once (*N*=8), leaving a final sample of 292 participants.

Design. This study had a 2 (Target Judgment: Deontological vs. Consequentialist) x 2 (Difficulty of Judgment: Easy vs. Difficult) between-subjects design. As in previous studies, participants played a TG with a target who gave a deontological or consequentialist response to a sacrificial moral dilemma, but in this study we added information that the target reported that their judgment was either "very difficult" or "very easy" to make.

The dependent measures in this study were identical to those used previously, with the additional inclusion of a partner preference question that asked participants to rank with which of four agents they would have preferred to play the trust game if they had been given a choice: an agent who gave either a deontological or consequentialist response, and who reported that their response was either a difficult or easy to make. The person the participant reported they would have most preferred to play with was scored as "1," and the person they would have least preferred to play with was scored as "4."

Because we were interested in whether there was an interaction effect between target judgment (deontological; consequentialist) and reported difficulty and because there is no standardized way of measuring interaction effects for non-parametric data, we used a square root function to transform the data and then ran a parametric ANOVA test to obtain an interaction effect to complement the main non-parametric simple contrasts. We report both the simple non-parametric contrasts and the parametric ANOVA results.

Results and Discussion

Character Judgments. As in Study 1, deontologists were rated as more moral overall ($F(1,288)=131.24$, $p<.001$). As predicted there was a significant interaction between agent and choice difficulty ($F(1,288)=13.42$, $p<.001$) such that consequentialists were judged less negatively if they reported that the moral decision was difficult ($U=1612$, $p<.001$, $d=0.55$: $t(135)=2.96$, $p=.04$). In contrast, choice difficulty for deontological agents did not have a significant effect on perceived morality ($U=2480$, $p=.07$, $d=0.29$; $t(153)=-1.73$, $p=.09$). We obtained similar results for perceived trustworthiness. Deontologists were seen as more trustworthy overall ($F(1,288)=67.83$, $p<.001$) and there was a significant interaction between agent and choice difficulty ($F(1,288)=5.71$, $p=.02$). While the means were in the expected direction, however, the simple effect of consequentialists being trusted more if the agent reported difficulty in making the decision was not significant ($U=1979$, $p=.10$, $d=0.31$: $t(135)=1.80$, $p=.07$). As before, reported decision difficulty had no significant effect on perceived trust of deontological agents ($U=2501$, $p=.09$; $d=0.25$, $t(153)=-1.54$, $p=.13$), and the means were in the direction of deontologist agents being trusted more if they said the decision was easy. Finally, results held when controlling for participants' own judgments. Consequentialists and deontologists perceived the deontologist agent to be more moral ($U=429$, $p<.001$) and trustworthy ($U=547$, $p=.03$). Overall, then, self-report data supported the claim that preference for deontologists is driven more by the presence of deontological intuitions, rather than the absence of consequentialist intuitions.

Trust Game. Behavioral measures of trust indicated that across both difficulty conditions the deontological agent received higher transfers than a consequentialist one ($U=8846$, $p=.009$, $d=0.31$; $F(1,288)=7.22$, $p=.008$), and was expected to return more ($U=8313$, $p<.001$, $d=0.38$; $F(1,288)=9.80$, $p=.002$). However, while the means were in the expected direction, there were no significant interaction effects with reported difficulty of the

decision, either for transfer amounts (F(1,288)=0.12, p=.73) or predicted returns (F(1,288)=1.05, p=.31). Again, results held when controlling for participant judgment. Participants who made a deontological judgment both transferred more to, and predicted more to be returned from, a deontologist agent, while participants who made consequentialist judgments showed no differences in either transfer amounts (U=737, p=.80, d=0.04) or predicted returns (U=701, p = .54, d=0.13) towards the two agents.

It is intriguing that a discrepancy arose between the self-reported and behavioural results: while participants' self-reported data showed an interaction effect of the judgment and difficulty, behavioural data revealed significant main effects of the judgment only (though means were in the expected direction). It is unclear whether this is due to behavioural measures being noisier, or whether this represents a real distinction between behavioural and self-reported judgments in this domain. Nonetheless, that there was this discrepancy provides further support for part of the motivation of this paper: to explore preferences for deontologists using actual behaviour and not just self-report judgments.

Partner Choice. When choosing between agents who responded with either consequentialist or deontological judgments, the majority of participants preferred to play with an agent who reported a deontological judgment (70%, p<.001). Moreover, when asked to rank their preferences for playing with four potential partners (who gave either a deontological or consequentialist response, and for whom the judgment was either difficult or easy), significant differences in the rankings emerged (Friedman's test: χ^2 (3)=62.47, p<.001). Specifically, participants provided higher rankings on average to deontological agents compared to consequentialist agents, but this preference was mitigated if the consequentialist agent reported difficulty in making the decision (Z=4.21, p<.001) (Figure 2). Results held when controlling for participant judgment such that deontologist participants preferred to play with a

deontologist agent, while consequentialist participants showed no preference ($\chi 2(3) = 1.62$, $p = .66$).

Study 3

Results from Studies 1 and 2 demonstrate that people who make characteristically deontological judgments in the footbridge dilemma are seen as more trustworthy social partners, consistent with a partner choice account of moral intuitions. But might our results thus far depend on specific characteristics of the footbridge dilemma? This dilemma has one important potential limitation relevant for partner choice – sacrificing one to save many involves an act of violent assault: pushing a man off a bridge. It is plausible, therefore, that our observed preference for deontologists is driven solely because this dilemma highlights the possibility that deontologists are simply more averse to physical harm, and not necessarily that they are more reliable cooperators. To rule out the possibility that the preference for deontologists we have observed in the previous studies is merely a preference for agents who do not commit physical assault, rather than a preference for agents who espouse a deontological morality, we sought to replicate our initial findings using a dilemma in which the sacrificial action does not require physical assault:

> A runaway trolley is heading down the tracks toward five workers who will all be killed if the trolley proceeds on its present course. Adam is on a footbridge over the tracks, in between the approaching trolley and the five workers. Next to him on this footbridge is a stranger who happens to be very large. The only way to save the lives of the five workers is to flip a switch to release a trapdoor that will drop the stranger off the bridge

and onto the tracks below where his large body will stop the trolley. The stranger will die if Adam does this, but the five workers will be saved.

Note that the trapdoor case is identical to the footbridge case in all of its relevant structural features: the only difference is that Adam does not push the large man, but instead pushes a button that opens a trapdoor that causes the large man to fall onto the tracks. If the preference for the deontological agent in the footbridge is explained simply by an aversion to physical assault - rather than the consequentialist action per se - we should see no preference for a deontological agent in the trapdoor case. In contrast, if our claim that deontological judgments serve as signals of a cooperative nature is correct, we should find the same pattern we observed using the footbridge version: that overall, a person who makes a deontological judgment is perceived more positively than one who makes a consequentialist judgment.

Method

125 American participants (40 female; M_{age}=33, *SD*=10.17) were recruited through MTurk and paid $1.00. Participants were excluded from analyses if they took the survey more than once (*N*=9) or had participated in a previous study (*N*=16), leaving a final sample of 101 participants. Most participants endorsed the deontological option (61%), but this difference was less pronounced compared to the footbridge case. The structure of this study was identical to that of Study 1b, with the same within-subjects design, dependent measures, and justification for the sacrificial action, but with the trapdoor dilemma instead of the footbridge dilemma.

Results

Results suggested that an aversion to agents who endorse physical assault cannot fully explain the preference observed for deontological over consequentialist agents. Participants rated the deontologist agent in the trapdoor dilemma as being more moral (Z=-5.41, p<.001, d=0.9) and trustworthy (Z=-4.61, p<.001, d=0.75) than a consequentialist one. Participants transferred more money to deontological agents (67%) than consequentialist agents (56%: Z=-2.22, p=.03, d=0.30) and predicted that deontological agents (41%) would return more money than consequentialist agents (34%: Z=-2.63, p=.009, d=0.25). As predicted, there was a significant difference in preferred partner, p<.001, with 65% of participants preferring to play a trust game with an agent who made a deontological judgment.

As with the footbridge dilemma, these results held when controlling for participants' own judgments. While deontological participants showed a strong preference for the deontological agent, consequentialist participants showed no preference between the two agents on perceived morality, (Z=-0.02, p=.99, d=0.02), trustworthiness (Z=-0.01, p=.99, d=0.07), transfers in the TG (Z=-1.42, p=.16, d=0.26), expected returns in the TG (Z=-0.86, p=.39, d=0.25), or partner choice (34%, p=.06). The same pattern was observed when looking at participants who endorsed either end-point on the wrongness scale, again confirming that our results cannot be attributed just to people preferring agents who agreed with them.

Study 4

Study 3 demonstrated that the preference for agents who made a deontological response in the footbridge dilemma cannot be explained purely due to an aversion to agents who physically harm others. In Study 4 we investigated an alternative explanation for why people prefer deontologists as social partners, testing whether this preference can be explained through

the notion of respecting persons and not treating them as 'mere means'. A typical feature of deontological ethics is the notion of respect for individual persons (Kant, 1797/2002; Scanlon, 1998), with corresponding implications for the acceptability of harm as a means to an end. On characteristically deontological approaches, it is morally wrong to use a person (e.g. holding slaves) as a mere means of acquiring some subjective end (e.g. to become wealthy) because to do so would deny the moral status of that person as a free and autonomous being. Relatedly, from a partner choice perspective one of the most fundamental ways of preserving a positive reputation in a cooperation market is to treat others as if they were persons with their own wishes, desires, and needs rather than mere objects to be used as necessary. Recall the footbridge and trapdoor dilemmas, where the death of the man occurs as an intended method (or *means*) to save the five others: the purpose of the large man being pushed or dropped is so that his body will stop the train. This use of someone merely as a means is part of what many deontological philosophers have claimed makes the action unacceptable, and seems consistent with commonsense moral intuitions. It may follow, then, that the use of others as mere means would also be undesirable in social partners.

This intuition has been highlighted using yet another variant of the trolley case in which an individual is faced with the decision to sacrifice one to save many--the switch case. In this formulation, the lives of five workers are saved by diverting the trolley onto another track by pulling a switch:

> A runaway trolley is heading down the tracks toward five workers who will all be killed if the trolley proceeds on its present course. Adam is standing next to a large switch that can divert the trolley onto a different track. The only way to save the lives of the five workers is to divert the trolley onto another track that only has one worker on it. If

Adam diverts the trolley onto the other track, this one worker will die, but the other five workers will be saved.

The switch case differs from the footbridge case in two critical ways (e.g. Crockett, 2013; Cushman, 2013; Greene et al., 2009). First, the harm in the footbridge case is direct and physical. We have already shown, in Study 3, that a mere aversion to agents who physically harm others cannot explain the preference for deontologists because deontological agents are also preferred in the trapdoor case where there is no direct physical violence. In Study 4 we exploited a second difference between the footbridge and switch cases – that of treating others as means to and end – to understand the preference for deontologist agents. Despite the general endorsement many people have that "ends do not justify means," people do typically judge that sacrificing the one man by diverting the train is less morally wrong than sacrificing the man by using his body to stop the train (Foot, 1967; Greene et al., 2001). Such folk intuitions align with the *Doctrine of Double Effect*, which is based on the "distinction between what a man foresees as a result of his voluntary action and what, in the strict sense, he intends" (Foot, 1967). On the doctrine of double effect, causing harm as a *side effect* of - but not a means to - bringing about a good outcome can be morally permissible. Indeed, aversion to violations of the doctrine of double effect might be an important driver of these differences in intuitions in the switch and footbridge cases (Crockett, 2013; Cushman, 2013, 2014). Therefore, to explore whether our observed preference for deontological agents is sensitive to violations of the doctrine of double effect, in Study 4 we looked at partner preference in the switch case.

To the extent that characteristically deontological judgments serve as cooperative signals because they indicate respect for persons, we would expect to see less negativity towards consequentialists in the switch case relative to the footbridge. This is because in the

footbridge case, the sacrificial action is performed with the intention of using the man's body to save the others and thus involves using the man as a mere means rather than respecting him as a person. A reported unwillingness to use others as means to an end is likely to signal, therefore, that one is a more trustworthy social partner. But in the switch case, the sacrificial action does not so obviously involve violating the man's autonomy by treating him as a mere means and so consequently should not signal as clearly that one is a more trustworthy social partner. Put simply, to the extent that people use judgments in moral dilemmas as indications of a person's trustworthiness as a social partner, the preference observed so far for deontological over consequentialist agents should be substantially weaker – or non-existent – in cases where the deontological action does not so obviously involve respecting persons more than the consequentialist action. We tested this in both a within-subjects (Study 4a) and between-subjects (Study 4b) design.

The switch case also enables us to test an alternative explanation for the observed preference for deontological agents: that people prefer as social partners those whose judgments accord with the majority view. In the cases tested thus far, the deontological judgment was also that endorsed by the majority of participants. However, in the switch case, the majority of participants endorsed the consequentialist agent and so if people simply prefer agents whose judgments reflect the majority view, then they should consistently prefer the consequentialist agent in the switch case.

Study 4a

Method

Participants. 161 American participants (88 female; M_{age}=38, SD=13.06) were recruited through MTurk and paid $1.00. Participants were excluded from analyses if they took

the survey more than once (N=7) or if they participated in one of the previous studies (N=32), leaving a final sample of 122 participants. In line with predictions and previous research (and in contrast to our previous studies using the footbridge and trapdoor dilemma), a majority of participants (73%) endorsed the consequentialist option, with only a minority this time endorsing the deontological option (27%).

Design. As in Study 3, the structure and dependent measures for this study were identical to that of Study 1b, but with the switch dilemma.

Results

Results showed that preferences for deontological over consequentialist agents largely depended on an aversion to agents who endorse using persons merely as a means, because these preferences disappeared when the sacrifice occurred as a side effect. In contrast to the previous studies, for the switch dilemma consequentialist agents were rated to be no less moral (Z=-0.73, p=.47, d=0.10) or trustworthy (Z=-1.87, p=.06, d=0.26) than deontological agents. Consequentialist agents did not receive smaller transfers (59%) than deontological agents (53%: Z=-1.86, p=.06, d=0.17) and were not predicted to return less (37%) than deontological agents (34%: Z=-0.82, p=.41, d=0.10). Overall, participants showed no significant preference to play with either a consequentialist or deontologist agent (p=.09).

Results broken down by participant judgment showed that, like the footbridge and trapdoor dilemmas, participants who made a deontologist judgment in the switch dilemma preferred partners who also made a deontologist judgment. In contrast to earlier studies, however, participants who made consequentialist judgment in the switch dilemma preferred consequentialist agents (see Supplementary Materials). This was the case whether using the

binary judgment or looking only at participants who endorsed the end-points on the wrongness scale. These results are in line with predictions such that in those cases where a consequentialist judgment does not clearly violate fairness-based principles about respecting others and not treating them as mere means, people do not infer that the agent is necessarily an untrustworthy social partner.

Study 4b

Method

750 American participants (327 female; M_{age}=34, SD=10.86) were recruited through MTurk in a 2 (Dilemma: footbridge vs. switch) x 2 (Agent Judgment: Deontological vs. Consequentialist) between-subjects design and paid $0.50. This was a direct replication of Study 1c, with the addition of the switch dilemma as a between-subjects factor. Sample size for the replication was determined by having 2.5 times the original sample size (Simonsohn, 2015). Participants were excluded from analyses if they took the survey more than once (N=10), leaving a final sample of 740 participants. Replicating our earlier findings, the majority of participants given the footbridge dilemma gave a deontological response (70%), while in the switch dilemma these proportions reversed such that the majority gave a consequentialist response (72%). To test whether there was an interaction effect between dilemma and agent judgment, as in Study 2 we used a square root function to transform the data and then ran a parametric ANOVA test to obtain an interaction effect to complement the main non-parametric results.

Results

Character Judgments. We first looked at perceived morality. In line with predictions, there was a significant interaction effect between the dilemma and agent judgment (F(1, 736)=58.40, p<.001), Participants perceived the deontological agent to be more moral than the consequentialist agent in both in the footbridge dilemma (U=5857, p<.001, d=1.42; t(375)=-13.17, p <.001) and switch dilemma (U=14390, p=.03, d=0.23; t(361)=-2.17, p=.03), though the effect size was considerably greater in the footbridge dilemma.

Similarly for perceived trust, in line with predictions there was a significant interaction effect between the dilemma and agent judgment (F(1, 736)=45.81, p<.001). Turning to the simple effects, we found that participants perceived the deontological agent to be more trustworthy in the footbridge dilemma (U=7993, p<.001, d=0.89; t(375)=-9.78, p<.001) but not the switch dilemma (U=16146, p=.73, d=0.05; t(361)=-0.62, p=.54).

Trust Game. We next looked at behavior in the TG. As predicted there was a significant interaction effect between the dilemma and agent judgment on transfer amounts (F(1, 736)=6.44, p=.01). Breaking the interaction effect down, it was found that participants trusted the deontological agent more than the consequentialist agent in the footbridge dilemma (U=14906, p=.005, d=0.31; t(375)=-2.81, p<.005) but not the switch dilemma (U=15323, p = .24, d=0.13; t(361)=-0.80, p=.43) (see Figure 3.)

For predicted returns, there was again a significant interaction effect between the dilemma and agent judgment (F(1, 736)=8.99, p<.003). Again, we found that participants predicted the deontological agent to return more than the consequentialist agent in the TG only for the footbridge dilemma (U=14449, p<.001, d=0.35; t(375)=-3.75, p<.001) but not for the switch dilemma (U=15896, p=.55, d=0.15; t(361)=-0.61, p=.54).

Overall, results from the TG showed that participants perceived an agent who made a deontological judgment to be more trustworthy than a consequentialist agent (as indexed by

transfer amounts and predicted returns) only when the consequentialist agent endorsed treating others as mere means.

Partner Choice. Results showed that overall, across both dilemmas there was no significant preference for either a deontological (53%) or consequentialist (47%) target. However breaking this down by dilemma type, we found that 70% of participants in the footbridge dilemma condition preferred to play with a deontologist target (p<.001), while this preference was reversed in the switch dilemma condition with 64% of participants preferring to play with a consequentialist target (p < .001).

These results largely held when controlling for participants' judgments. In line with results from Study 1, for the footbridge dilemma consequentialist participants perceived the deontologist agent to be more moral (U=1101, p=.005, d=0.55), and showed no difference in perceived trust (U=1282, p=.09, d=0.31), transfer amounts (U=1446, p=.46, d=0.12), or predicted returns (U=1563, p=.99, d=.04). Again, results cannot be attributed to participants simply preferring those who agree with them on moral problems. For the switch dilemma, consequentialist participants showed no difference in perceived morality (U=8511, p=.99, d=0.03) or trust (U=7606, p=.12, d=0.23), and while consequentialist participants did transfer more to a consequentialist agent (U=6896, p=.006, d=0.34), there was no difference in predicted returns (U=7843, p=.25, d=0.19) (see Supplementary Materials for all Ms, SDs, and significance tests). It is somewhat unclear why there was only a single effect of consequentialist participants trusting a consequentialist agent more in terms of transfers (but not rated trust or predicted returns), only in the between-subjects design (and not Study 4a). But nonetheless, the overall pattern of results is consistent with our predictions and the findings presented here in Studies 1-3.

Discussion

Results from Study 4 were consistent with deontological agents being preferred as social partners to the extent that such judgments honor implicit obligations to not treat others as mere means, suggesting that deontological judgments communicate trustworthiness. While deontological agents were preferred over a consequentialist agent when the (consequentialist-endorsed) sacrificial act clearly involved using others as mere means (Studies 1-3), no such preference was observed when the (consequentialist-endorsed) sacrificial action was not strongly associated with using others instrumentally (Study 4). Furthermore, the fact that participants did not show a consistent tendency to perceive the consequentialist agent to be more trustworthy in the switch case – in which the majority response is to endorse the consequentialist sacrifice - demonstrates that the preferences for the deontological agent observed in Studies 1-3 do not reflect a mere preference for agents whose judgments accord with the majority view.

Study 5a

We have so far implicitly focused (along with the vast majority of psychological work) on the deontological theory of Kant, who held that moral law consists of a set of maxims, or rules, that are categorical in nature, and that we are bound by duty to act in accordance with these categorical imperatives (Kant, 1797/2002). It is this (very simplified) Kantian view - whereby certain acts are intrinsically morally right or wrong - that predominates in the moral psychology literature when deontology is discussed. But this simplified account often ignores the critical roles of justice, duties, obligations, and rights that are central features of (neo-)Kantian ethics. And we have already presented evidence that people might attend to these features when selecting social partners, whereby people strongly prefer agents whose moral

judgment does not violate the implicit duties we have to one another (Studies 1-3) but show no preference when a moral judgment doesn't so obviously violate these duties or obligations (Study 4).

But there are other deontological approaches that extend Kantian thinking: for example, those focusing on the idea of social contracts and the ways our actions can be justified to one another (Gauthier, 1986; Hobbes, 1668/1994; Parfit, 1984; Rawls, 1971; Scanlon, 1998). Of particular interest is recent theoretical work that has argued for the evolution of a *contractualist morality* by partner choice mechanisms (Baumard & Sheskin, 2015). Moral contractualism is a non-consequentialist ethical theory developed by Scanlon (Scanlon, 1998), on which moral actions are those that would result if we were to make fair and binding agreements – i.e., social contracts - from a point of view that respects our equal moral importance as rational autonomous agents. Baumard and Sheshkin (2015) argue for a contractualist account of commonsense moral intuitions whereby people are likely to endorse a sacrificial action when such actions align with a principle of fairness (as in the switch dilemma), but reject a sacrificial action in most other cases, where such actions would violate implicit contractual obligations between persons (as in the footbridge or trapdoor dilemmas). People do not seem to be intuitively applying consequentialist principles to these two dilemmas, and indeed as Scanlon himself wrote, "the implications of act utilitarianism are wildly at variance with firmly held moral convictions" (Scanlon, 1982). In the switch dilemma, all the individuals are on the train tracks and it is merely chance that the train is headed down one set of tracks. Given this element of chance, there is a sense in which all individuals have an equal right to be saved, and so to divert the train to save more people is a moral action entirely consistent with the mutualistic logic of partner choice. In contrast, pushing a stranger off a footbridge does not fit this model of fairness because there is no way in which the train would have gone onto the footbridge and

so the stranger was already safe from danger. Baumard and Sheshkin (2015) argue, therefore, that a simple and commonsense fairness principle respecting persons' autonomy can explain the variance in moral intuitions across dilemmas: "When someone has something (e.g., safety from being in the potential path of a trolley), respect it; when people are on a par (e.g., they are all in the potential path of a trolley), then do not favor anyone in particular" (p.45). The contractualist account of Baumard and Sheskin is posited to explain why people typically endorse the sacrificial action in the switch, but not footbridge, dilemma.

Partner choice mechanisms, then, may have selected for moral intuitions that are consistent with the demands of justice and our mutual obligations to one another ('contractualist') as opposed to those that draw solely on specific acts being wrong regardless of the context ('categorical'). Consistent with this notion, results from Studies 2-4 suggest that it is the deontological feature of respecting persons and honoring social contracts, rather than committing to abstaining from specific actions *per se*, that signals trustworthiness – and this is consistent with the central feature being contract-based (as in Scanlon's Contractualism) rather than merely rule-based (e.g. a very simplified categorical-based Kantian ethic). In Study 5 we sought to test this hypothesis directly. To the extent that evolution may have favored a specifically contractualist morality (c.f. (Baumard & Sheskin, 2015)), when a characteristically categorical-based judgment prohibiting a certain act conflicts with a characteristically contractualist judgment endorsing the same act, the agent who makes the contractualist decision should be seen as a more trustworthy social partner.

We explored this in Study 5 using the so-called "soldier's dilemma." In this dilemma, a soldier is badly injured and caught in a trap, with the enemy fast approaching. The soldier cannot escape, and begs the troop leader not to leave him behind, as he will be cruelly tortured to death. Should the troop leader stab the soldier in the heart to prevent his suffering at the

hands of the enemy? In this dilemma, and in contrast to those previously used, it is the *endorsement* – rather than rejection - of the sacrificial action that is consistent with a contractualist deontological ethical analysis. Many people, if they were the fallen soldier, would want to be put out of their misery to prevent further suffering. Moreover, killing the soldier actually relieves his suffering and respects what the soldier himself wants. To the extent that partner choice mechanisms have promoted intuitive deontological judgments more along contractualist rather than categorical lines, participants should report a preference for agents who endorse the sacrificial action (contractualist) compared to those who reject this action (categorical).

Method

150 American participants (72 female; M_{age}=33, SD=9.93) were recruited through MTurk and paid $0.50. Participants were excluded from analyses if they took the survey more than once (N=5), leaving a final sample of 145 participants. Study 5 used the same between-subjects design and dependent measures as previous studies, but with the following dilemma:

> "Harry is the leader of a small group of soldiers, and all of the group is out of ammunition. Harry is on his way back from a completed mission deep in enemy territory when one of his men steps in a trap set by the enemy. The soldier's leg is badly injured and caught in the trap. Harry cannot free him from this trap without killing him. However, if Harry leaves the soldier behind, the enemy troops will find the soldier and torture him to death. The soldier begs Harry not to leave him behind to be cruelly tortured to death. The enemy troops are closing in on their position and it is not safe for Harry or his men to remain with the trapped comrade any longer. In order to prevent

this man's needless suffering at the hands of the enemy, Harry could kill the soldier himself by stabbing him in the heart. Should Harry stab the soldier in the heart to prevent his suffering at the hands of the enemy?"

In this dilemma, participants were told that the agent who endorsed the sacrifice of the soldier justified their judgment by saying "it is acceptable to kill someone if it reduces their suffering", while the agent who rejected the sacrifice justified their judgment by saying "killing people is just wrong, even if it has good consequences". We refer to these as the contractualist and categorical agents respectively.

Results

In the soldier's dilemma, only a minority of participants (25%) endorsed the categorical-deontological option and said it would be wrong to sacrifice the man, while the majority of participants endorsed the contractualist-deontological option by stating that it would be morally right to acquiesce to the soldier's wishes and kill him (75%).

Participants preferred agents whose judgments were consistent with a contractualist, rather than categorical, morality. In the TG, participants entrusted more money to a contractualist-deontological agent (64%) endorsing the sacrificial action (51%: U=2110, p=.04, d=0.33), and predicted this agent (37%) to return more back to them (28%: U=2073, p=.03, d=0.33), relative to the categorical-deontological agent condemning the sacrificial action (Figure 4). Moreover, most participants (59%) preferred to play with the contractualist agent over the categorical agent (p<.001). These results held when controlling for participant's own judgments, whereby participants making the contractualist judgment trusted the contractualist agent more in terms of both transfer amounts (U=1128, p=.03, d=0.42) and

predicted returns (U=1019, p=.004, d=-0.55), while participants making the categorical judgment showed no preference between the two agents. Yet overall, character ratings were less consistent: while the contractualist and categorical agents were rated as equally trustworthy overall (U=2192, p=.08, d=0.28), the categorical agent was rated as more moral than the contractualist agent (U=1949, p=.006, d=0.47).

Study 5b

In Study 5a we found results consistent with the claim that individuals who make judgments that can be seen as honoring implicit social contracts and obligations to one another are preferred over those whose actions do not. There are, however, a few aspects of Study 5a that preclude drawing strong conclusions. Study 5b was designed to address ambiguities in the design of Study 5a, and as a secondary question, investigate whether participants attend more to agents' justifications for a given action, or to agents' endorsement of the action itself.

The first potential problem with the design of Study 5a was that the specific action that the soldier himself wanted was potentially ambiguous. In the original dilemma, participants read that the "soldier begs Harry not to leave him behind to be cruelly tortured to death". Given that the commander "cannot free him from this trap without killing him", it is implicit that the course of action the soldier is advocating is a mercy-killing. However, if participants did not interpret it in this way, the claim that participants attend to whether an agent respects a person's wishes and autonomy is weakened. Therefore, in Study 5b we made salient – and manipulated across conditions – what action the soldier himself wanted: in one condition he implored "Please, kill me. I don't want to suffer at the hands of torturers", while in a second condition he said "Please, don't kill me. I don't want to die out here in the field". Given that notions of

consent and respect for autonomy are central to deontological ethics – and exactly what one should seek in a social partner - the preference for the agent who endorsed the stabbing of the soldier should only be observed when the endorsed action aligns with the soldier's injunctions.

A second potential problem with Study 5a is that the justification to stab the soldier to prevent his suffering conflated both contractualist and consequentialist reasons. In Study 5a, the agent's justification for the judgment that it would be morally right to stab the soldier was that it would reduce "*their* suffering". But this is potentially ambiguous as an indication of a contractualist vs. a consequentialist style of thinking because both theories aim at reducing suffering. Put simply, it is not clear whether participants preferred the agent because their judgment was contractualist-consistent, or because they inferred consequentialist motives. The key difference between these two approaches is that consequentialist theories aim to maximize overall aggregate happiness, while contractualist theories focus more on specific individuals and the obligations we have to them in a given context. Aside from this, it remains to be seen whether participants are focusing more on the specific action endorsed (e.g. stab or not stab), or the justifications espoused for that action (e.g. respecting wishes vs. reducing overall suffering). Therefore, in Study 5b as well as manipulating whether the soldier gave consent we manipulated agents' justifications for the action: one agent focused on relieving suffering, regardless of whether consent was given or not (consequentialist: "It is acceptable to kill someone if it reduces overall suffering"); one agent focused on killing being wrong, regardless of whether consent was given or not (categorical: "Killing people is just wrong, even if it has good consequences"); and two agents focused explicitly on autonomy and respecting the soldier's wishes (contractualist: ("It's right [wrong] to kill the soldier if that's [not] what they want, and it's the commander's duty to respect that"). This enabled us to explore not only our primary question of whether people prefer agents who endorse acting in accordance with the

soldier's wishes, but also a secondary question of whether participants attend primarily to the justifications or merely the endorsement of action. To the extent that people focus only on the action, we should see a preference for the agent who acts in accordance with the soldier's wishes, regardless of the justification given. But to the extent that people are concerned with justifications, we should see a pattern whereby people prefer agents differentially based on their justification – even when they all endorse the same consent-conforming action.

A final – and more minor – potential problem with Study 5a that we sought to correct is that in the original dilemma the soldier is presumed to be awake and conscious when he is stabbed, and this might invoke feelings of harm aversion in participants. Therefore, in Study 5b we made it clear that the soldier would be unconscious by the time that he would be stabbed and so would not feel any immediate pain or suffering.

Method

454 American participants (185 female; M_{age}=34, SD=11.28) were recruited through MTurk and paid $0.70. Participants were given a modified version of the soldiers' dilemma from Study 5a, where we manipulated whether the soldier asked to be killed or not in a between-subjects design. As a secondary manipulation, we varied the justification that the MTurker agent gave, such that participants believed they were playing with an agent who had read the dilemma and said that they thought the soldier either should or should not be sacrificed, and gave either a characteristically consequentialist, contractualist, or categorical-based justification (see Table 1). Therefore, this study had a 2 (Soldier's Consent: Yes, No) x 3 (Agent Justification: consequentialist; contractualist; categorical) design. The dilemma was given as follows:

> ‘Harry is the leader of a small group of soldiers, and all of the group is out of ammunition. Harry is on his way back from a completed mission deep in enemy territory when one of his men steps in a trap set by the enemy. The soldier's leg is badly injured and caught in the trap. Harry cannot free him from this trap without killing him. The enemy is advancing and they will undoubtedly find the soldier and torture him to death. The enemy troops are closing in on their position and it is not safe for Harry or his men to remain with the trapped comrade any longer. Harry offers to stab the soldier in the heart after he's unconscious to kill him quickly and prevent him suffering at the hands of the torturers. Just before he passes out due the pain, the soldier pleads to Harry "Please, kill me. I don't want to suffer at the hands of torturers" ["Please, don't kill me. I don't want to die out here in the field"].’

Results

The majority of participants reported that they thought the morally right action was the one that the soldier wanted (71%), highlighting the importance of consent and respecting wishes to participants’ own moral judgments. Breaking this down, when participants read that the soldier wanted to be killed, 88% indicated that they thought it morally right to stab the soldier, with this dropping to 44% for participants who read that the soldier asked not to be killed. Overall, then, while participants did appear to have intuitions along categorical lines (“killing is wrong”), their judgments were more in line with a respect-based contractualist analysis (“honor people’s autonomy and respect their wishes”). This replicates our finding from Study 5a that participants were more likely to say they endorsed the sacrifice, as well as suggesting this pattern was indeed driven by participants’ consideration of the soldier’s wishes (which were potentially ambiguous in Study 5a).

Did participants prefer agents whose moral judgments accorded with the soldier's wishes? In line with predictions, results suggested that they did: Regardless of whether the decision made was to sacrifice or not, participants preferred the agent who endorsed the action that conformed to the soldier's wishes over those who endorsed the action that did not. Agents whose judgments conformed to the soldier's wishes were rated as more moral (U=19348, p=.002) and trustworthy (U=17497, p<.001), and received higher transfers in the TG (U=20322, p=.01). There were, however, no significant effects on predicted returns (U=20868, p=.08) or partner choice (p=.33), although the means went in the predicted direction.

We next considered whether participants' preferences for the agent who honored the soldier's wishes depended on whether they themselves endorsed the action that the soldier wanted. Results showed that, for participants who themselves endorsed the action that violated the soldier's wishes, there were no significant effects in the ratings of agents who either conformed to or violated the soldier's wishes: on morality (U=1738, p=.28), trust (U=1746, p=.30), transfers (U=1896, p=.76), or predicted returns (U=1770, p=.35). However, for participants who themselves endorsed the action that the soldier wanted, there were significant differences in judgments of the agent: for morality (U=8405, p<.001), trust (U=7097, p<.001), transfers (U=9469, p=.002), and predicted returns (U=9472, p=.006). Such results parallel those found in the previous studies, where it is deontological participants – for whom notions of consent, respect and duty outweigh concerns based solely on maximizing happiness – who show a preference for other deontologists, while consequentialist participants show no such preference. Moreover, these results highlight that the critical feature does appear to be contractualist notions of consent and respect, rather than mere categorical concerns about

specific actions being forbidden. As such, these results are consistent with evolution favoring moral intuitions more along contractualist than categorical lines.

Finally, we investigated the extent to which participants were concerned with the justifications that the agent gave. If the observed preference for deontologists is driven primarily by their endorsement of a favored *action* that honors consent and mutual obligations, we should see a preference for an agent who acts in accordance with the soldier's wishes, regardless of the justification given. But if people are specifically concerned with deontological *justifications*, we should see a pattern whereby people prefer agents differentially based on their justification – even when they all endorse the same consent-conforming action. To this end, we first looked at the conforming cases where the agent endorsed the action that the soldier wanted (i.e. the consequentialist in the consent condition, the categorical in the non-consent condition, and the contractualist in both conditions). Results from a Kruskal-Wallis test suggested that when the action conformed with what the soldier explicitly asked for, participants did not judge agents differently based on the justifications given: for ratings of morality (H=5.52, p=.06), trust (H=0.49, p=.78), transfers (H=0.11, p=.95), or predicted returns (H=1.08, p=0.58). Put simply, as long as the agent's judgment respected the soldier's wishes, participants did not care much about the reasoning why. A different picture emerged for the non-conforming cases where the agent endorsed an action that the soldier explicitly did not want (i.e. the consequentialist agent in the non-consent condition, and the categorical agent in the consent condition). While there were no effects on behaviour in the trust game, participants judged that an agent who *refused* to sacrifice the soldier against his wishes (i.e. the categorical-consent condition) as significantly more moral (H=16.76, p<.001) and trustworthy (H=5.93, p=.02) than the agent who endorsed the sacrifice of the soldier against his wishes (i.e.

the consequentialist-non-consent condition). In other words, results indicated that participants attended to the justification for the action only when it violated the soldier's wishes.

Discussion

In Study 5 we explored more directly the features of deontological judgments that signal trustworthiness. The data provide further evidence that deontologists are preferred to the extent that they honor the social relationships we have with on another: social relationships that depend vitally on respect of consent, not treating others as mere means, and mutual duties. Results from Study 5b also speak against one potential interpretation of our earlier findings: that a preference for deontologists merely reflects a preference for agents whose judgments are consistent with those of the majority. We found, for example, that 88% of participants reported that when the soldier asked to be killed it would be morally right to sacrifice him – but still thought an agent who refused to sacrifice him in this case was more moral and trustworthy than an agent who endorsed the sacrifice.

General Discussion

Collectively our findings suggest that characteristically deontological judgments in sacrificial moral dilemmas are perceived as signals of trustworthiness to the extent that these judgments indicate respect for persons and commitment to social cooperation, thereby providing the first empirical evidence for a partner choice account of intuitive moral judgments. Across five studies we observed a general pattern whereby people who make deontological judgments are preferred as social partners, seemingly because these judgments involve respect for persons. Using several complementary methods from a range of disciplines, we show that

characteristically deontological judgments in sacrificial dilemmas enhanced an individual's cooperative reputation and thus could be argued to improve their fitness in the cooperation market. In contrast, individuals making consequentialist judgments were seen as less moral and trustworthy, and thus devalued as social partners. This pattern of results cannot be explained merely as a function of people preferring those who make the majority moral decision, because this preference for the non-sacrificial agent was seen even when the majority view is to sacrifice. Nor can our results be explained by people simply preferring agents who express the same moral view as themselves: across all studies these results held even when controlling for participants' own judgments. Furthermore, we show that while this effect holds for agents who report the sorts of deontological judgments consistent with intuitive, commonsense morality, there are predictable exceptions to this pattern of results. When communicating that a consequentialist judgment was made with difficulty, negativity towards agents who made these judgments was reduced. And when a harmful action either did not blatantly violate implicit social contracts, or actually served to honor them, there was no preference for a deontologist over a consequentialist. Our research is consistent with previous claims that socially valued moral intuitions more closely approximate deontological ethics that focus on mutual obligations and implicit social contracts (Baumard & Sheskin, 2015). Notably, these results were consistent across a range of dependent measures borrowed from different disciplines: explicit self-report methods common to social psychological studies (character ratings of morality), behavioral data from economic games (the trust game), and partner choice questions inspired by evolutionary biology (partner preference decisions). That we observed convergent results using these varied measures across our studies lends confidence that characteristically deontological judgments increase one's value as a social partner, with tangible effects both on social perception and behavior.

Deontological Judgments as Signals of Trustworthiness

These results are consistent with a partner choice account of moral intuitions in that they suggest that typically deontological judgments confer an adaptive function by increasing the likelihood of being chosen as a cooperation partner - and so deontological moral intuitions, as a form of 'cooperating without looking' (Hoffman et al., 2015) may represent an evolutionarily prescribed prior that was selected for through partner choice mechanisms. Importantly, while consistent with the findings that Uhlmann and colleagues (2013) reported—that people generally perceive those who make consequentialist judgments to be less moral than those who make deontological judgments—we provided behavioural evidence of the sort that would be predicted by this account.

The addition of behavioural measures also led to the finding that, while having substantial overlap, in some cases explicit character ratings and actual trust behavior diverged. While incentivized trust behavior in the TG supported our predictions across all studies, there were some exceptions for character ratings. In Study 4b, for example, character ratings were more positive for a deontological agent in both the footbridge and switch dilemma, but differences in actual trust behavior were only observed in the footbridge case. A similar divergence was observed in Study 5a, where participants' behavior in the TG showed greater trust of the contractualist agent who endorsed the sacrifice of the solider, but self-report ratings of character indicated no differences between the two agents. It is unclear whether these findings are due to different psychological mechanisms guiding behaviour vs. explicit attitudes, whether behaviour might be seen as the "real" measure of trust, but individuals do not have explicit access to how they might act, or whether there is some other unknown explanation for these dissociations. What is clear is that these results highlight the importance of measuring

social behavior in addition to social perceptions, because the two do not always point in the same direction.

Could expressing deontological judgments be perceived as a 'signal' of trustworthiness? Evolutionary biologists distinguish between two distinct mechanisms for the evolution of reliable signals: a signal can serve as an 'index' of some underlying quality, or a 'handicap' that carries an associated cost (Smith & Harper, 2003). Which of these mechanisms might apply to the expression of characteristically deontological judgments? Supporting the indexing mechanism, deontological judgments are positively associated with physiological indices of harm aversion (Cushman et al., 2012) and negatively associated with antisocial personality traits (Bartels & Pizarro, 2011; Kahane et al., 2015; Koenigs, Kruepke, Zeier, & Newman, 2012). Moreover, it has been argued that deontological judgments reflect moral emotions (Greene, 2014), which enable people to make credible commitments to prosocial behaviors (Frank, 1988). Meanwhile, the handicap mechanism would require endorsement of deontology to carry some cost, and for this cost to be higher for selfish types than for trustworthy types. One possibility is that the rigidity of deontology renders its supporters more vulnerable to being branded as hypocrites, since it necessarily provides less 'wiggle room' for rationalizing defection than does consequentialism, which is more flexible. For a trustworthy agent who will never behave dishonestly, there is little cost to committing to a deontological morality that dictates "always be honest". However, for an agent who will sometimes behave dishonestly when the benefits are sufficiently high, there is an obvious advantage to subscribing to a consequentialist morality that does not condemn dishonesty absolutely, because it opens possibilities for justifying dishonest behavior. Thus, a dishonest agent bears a higher risk of being branded a hypocrite when subscribing to a deontological morality, relative to a trustworthy agent. Future work could usefully address whether either of these signaling

explanations provide alternative explanations for the phenomenon of deontological intuitions.

Finally, it is important to reiterate that our results focus on perceptions of trustworthiness, rather than measuring trustworthiness itself. Our results are, therefore, ambiguous as to whether people that make characteristically deontological judgments are in fact more trustworthy. While judgments in sacrificial dilemmas are reliably perceived as indicating trustworthiness, whether they do so or not in actuality is an open question and so this is a ripe topic for future research.

Who Prefers Deontologists?

The main hypotheses and discussion in this paper have focused on perceptions of agents making deontological and consequentialist agents independent of participants' own judgments in the dilemma. The reason for this simple: to the extent that characteristically deontological judgments improve fitness in the cooperation market, it is not necessary that each and every person prefers an agent that makes a characteristically deontological judgment, but rather that *ceteris paribus,* an agent who makes characteristically deontological judgments will be trusted more overall by a given population. And indeed, in this work and previous work the evidence is that it is deontologists that are preferred overall. But to highlight why it is not necessary for our partner choice account that each and every individual person would favor a deontologist, a comparison with work on the religiosity and anti-atheist prejudice is useful. Because religious belief has been associated with large-scale cooperation and prosociality, religiosity has come to be viewed as a indicator of trustworthiness (Norenzayan et al., 2014) such that religious individuals are trusted more relative to their atheist counterparts by the general population, and that this effect is stronger when judged by other religious individuals (Gervais, Shariff, & Norenzayan, 2011). It is implausible that nonreligious people would consciously agree that

religious belief drives prosociality, for this would imply that they themselves are less moral. Rather, partner preference in religiosity (like, we argue, characteristically deontological judgments) takes place primarily at an ultimate, rather than psychologically proximate, level, and it need not be evident in each and every person. In line with this, results were not driven simply by people disliking those who disagree with them: across studies deontologist agents were preferred by participants endorsing both deontologist and consequentialist judgments, in those cases where consequentialism-endorsing participants showed any preference at all.

While there wasn't a simple matching effect, one interesting finding to emerge is that consequentialist participants were less likely to show a consistent pattern of preference for deontologists. Why is this? One explanation is that these judgments are reflecting the sum of two kinds of preferences: a preference for those individuals who are like them, and a preference for deontologists. For participants who make deontological judgments, these preferences are aligned and are revealed as a strong preference for deontological agents; whereas for participants who make consequentialist judgments these preferences are in conflict and cancel each other out. A second explanation is that participants who make deontological judgments prefer other deontologists (and/or others who are like them), while participants who make consequentialist judgments are more impartial in their preferences. One might be tempted to argue for the second explanation because impartiality towards others is a central feature of consequentialism. However, previous work has shown that participants who make consequentialist judgments in sacrificial dilemmas like those used in the current study are decidedly *not* more impartial than participants who make deontological judgments (Kahane et al., 2015). This evidence, in addition to the large body of evidence that people show a similarity bias in their social judgments (Lydon, Jamieson, & Zanna, 1988), suggests that the first account – that of conflicting preferences – may be more likely. Nonetheless, it will be interesting for

future research to explore this more fully.

Strengths and Limitations

Is it a problem that we look at perceptions of agents who made moral judgments in hypothetical moral dilemmas that are somewhat removed from typical moral problems that people face in everyday life? We argue not, for two reasons. First, recall that the aim of this paper was to investigate partner choice as an ultimate mechanism describing why moral intuitions are characteristically deontological. Because the bulk of psychological research on moral intuitions has used hypothetical sacrificial dilemmas such as trolley problems, we adopted this methodology as a way of directly pitting deontological and consequentialist intuitions against one another – since in practice, deontological and consequentialist theories overlap heavily in terms of what actions they permit or forbid. Second, and just as importantly, it is critical to highlight that while the moral dilemmas may have been hypothetical, our dependent measures were definitely not. One of the central ways in which this work advances upon important prior work by Uhlmann and colleagues (2013) is in the recognition that character judgments do not always map directly onto behavior, which is why we used a behavioral economic methodology where participants made decisions concerning the allocation of real money that had real consequences on how much money they were paid for taking part in the study. Nonetheless, it would be interesting for future work to follow on from this by using other kinds of moral issues that evoke both deontological and consequentialist intuitions.

Will deontologists always be preferred? The evidence presented here suggests that, in general, deontologists will be preferred as social partners, but with the caveat that this is in the very specific context of online economic games. In real life, people select a range of social

partners (e.g. a friend or romantic partner, but also professional relationships such as doctors or lawyers) and what we value most in others can vary as a function of what kind of relationship it is. What one values in a loved one (e.g. warmth) might not be the same things we look for in a lawyer (e.g. competence). Given this, it remains plausible that at least in some contexts, it will be consequentialist agents who will be preferred. It will be generative for future research to explore this in order to understand more deeply how partner choice models can explain moral intuitions in different contexts.

Commonsense Psychology and Ethical Theories

Our work offers a new perspective on the possibility of bridging normative ethical theories with empirical findings in moral psychology. Research on moral judgments in sacrificial moral dilemmas has often tried to explain – or justify – these judgments through the lenses of ethical theories such as Utilitarianism or a simplified categorical-based Kantian deontology. Yet this endeavor has met limited success, as laypeople's' judgments about endorsing or rejecting the sacrifice of one to save others bear little resemblance to the demands of these ethical theories (Kahane et al., 2015). Contra Utilitarianism, commonsense morality does not have the sole aim of maximizing aggregate welfare (Baumard et al., 2013), and contra very simplified forms of categorical-based Kantian ethics does not treat moral rules as absolutely binding (Kahane, 2015). Rather, commonsense morality appears to be pluralist, consisting of a variety of specific fairness and harm-based principles, where (like on some contractualist theories) sometimes it is permissible to overrule some specific deontic principle if following it would lead to great harm. One interesting implication of our work, therefore, is that researchers exploring how folk moral judgments align with normative ethical theories could usefully consider moral contractualism. To whit: it is unlikely that commonsense

intuitions will have a direct mapping onto the philosophical principles of moral contractualism or neo-Kantian ethical theories; but at the very least it seems describing commonsense morality along such contractualist-deontological principles will be less wrong. Our evolved commonsense morality is not utilitarian and not deontological in a simple Kantian-categorical sense, but with it's focus on justice and fairness, it does share important features with contractualist moral theories (Baumard & Sheskin, 2015). Moral contractualism, in addition to aligning well with the moral judgments people typically make, may also help to inform *why* we make these judgments, and under what conditions these judgments can be defended from a normative standpoint - and we look forward to future empirical and theoretical work exploring this.

Acknowledgments

The authors would like to thank Fiery Cushman, David Rand, Joshua Greene, Pat Barclay, Adam Morris, Phil Smolenski, Lucius Caviola, Jenifer Siegel, Filip Gesiarz, Felix Heise, Andreas Kappes, and Nadira Faber for helpful comments. The research was supported by an award to MC from the Oxford University Press John Fell Fund.

References

Alexander, R. (1987). *The Biology of Moral Systems*. New York, NY: Aldine de Gruyter.

Barclay, P. (2004). Trustworthiness and competitive altruism can also solve the "tragedy of the commons." *Evolution and Human Behavior*, *25*(4), 209–220. http://doi.org/10.1016/j.evolhumbehav.2004.04.002

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior*, *27*(5), 325–344. http://doi.org/10.1016/j.evolhumbehav.2006.01.003

Bartels, D. M., & Pizarro, D. A. (2011). The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition*, *121*(1), 154–161.

Baumard, N., André, J.-B., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences*, *36*(01), 59–78.

Baumard, N., & Sheskin, M. (2015). Partner Choice and the Evolution of a Contractualist Morality. In J. Decety & T. Wheatley (Eds.), *The Moral Brain: A Multidisciplinary Perspective* (pp. 35–46). Cambridge, MA: MIT Press.

Bentham, J. (1983). *The Collected Works of Jeremy Bentham: Deontology, Together with a Table of the Springs of Action ; and the Article on Utilitarianism*. Oxford, England: Oxford University Press. (Original work published 1879)

Conway, P., & Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision making: A process dissociation approach. *Journal of Personality and Social Psychology*, *104*(2), 216.

Critcher, C. R., Inbar, Y., & Pizarro, D. A. (2013). How quick decisions illuminate moral character. *Social Psychological and Personality Science*, *4*(3), 308–315.

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences*, *17*(8), 363–366.

Cushman, F. (2013). Action, Outcome, and Value A Dual-System Framework for Morality. *Personality and Social Psychology Review*, *17*(3), 273–292. http://doi.org/10.1177/1088868313495594

Cushman, F. (2014). The Psychological Origins of the Doctrine of Double Effect. *Criminal Law and Philosophy*, 1–14. http://doi.org/10.1007/s11572-014-9334-1

Cushman, F., Gray, K., Gaffey, A., & Mendes, W. B. (2012). Simulating murder: the aversion to harmful action. *Emotion*, *12*(1), 2.

Cushman, F., Young, L., & Hauser, M. (2006). The Role of Conscious Reasoning and Intuition in Moral Judgment Testing Three Principles of Harm. *Psychological Science*, *17*(12), 1082–1089. http://doi.org/10.1111/j.1467-9280.2006.01834.x

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review*, *5*, 5–15.

Frank, R. H. (1988). *Passions Within Reason: The Strategic Role of the Emotions*. Norton.

Gauthier, D. P. (1986). *Morals by Agreement*. Oxford, England: Oxford University Press.

Gervais, W. M., Shariff, A. F., & Norenzayan, A. (2011). Do you believe in atheists? Distrust is central to anti-atheist prejudice. *Journal of Personality and Social Psychology*, *101*(6), 1189–206. http://doi.org/10.1037/a0025882

Greene, J. D. (2007). Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends in Cognitive Sciences*, *11*(8), 322–323.

Greene, J. D. (2014). *Moral Tribes: Emotion, Reason and the Gap Between Us and Them*. London, England: Atlantic Books Ltd.

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention

in moral judgment. *Cognition*, *111*(3), 364–371. http://doi.org/10.1016/j.cognition.2009.02.001

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, *107*(3), 1144–1154.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105–2108.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834. http://doi.org/10.1037/0033-295X.108.4.814

Hobbes, T. (1994). *Leviathan (Edited by Edwin Curley)*. Hackett. (Original work published 1668)

Hoffman, M., Yoeli, E., & Nowak, M. A. (2015). Cooperate without looking: Why we care what people think and not just what they do. *Proceedings of the National Academy of Sciences*, *112*(6), 1727–1732.

Kahane, G. (2015). Sidetracked by trolleys: Why sacrificial moral dilemmas tell us little (or nothing) about utilitarian judgment. *Social Neuroscience*, *0*(0), 1–10. http://doi.org/10.1080/17470919.2015.1023400

Kahane, G., Everett, J. A. C., Earp, B. D., Farias, M., & Savulescu, J. (2015). "Utilitarian" judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition*, *134*, 193–209. http://doi.org/10.1016/j.cognition.2014.10.005

Kant, I. (2002). *Groundwork for the Metaphysics of Morals*. New Haven, CT: Yale University Press. (Original work published 1797)

Koenigs, M., Kruepke, M., Zeier, J., & Newman, J. P. (2012). Utilitarian moral judgment in psychopathy. *Social Cognitive and Affective Neuroscience*, *7*(6), 708–714.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgments. *Nature*, *446*(7138), 908–911. http://doi.org/10.1038/nature05631

Krebs, D. (2008). Morality: An Evolutionary Account. *Perspectives on Psychological Science*, *3*(3), 149–172. http://doi.org/10.1111/j.1745-6924.2008.00072.x

Lydon, J. E., Jamieson, D. W., & Zanna, M. P. (1988). Interpersonal Similarity and the Social and Intellectual Dimensions of First Impressions. *Social Cognition*, *6*(4), 269–286. http://doi.org/http://dx.doi.org.ezp-prod1.hul.harvard.edu/10.1521/soco.1988.6.4.269

Mill, J. S. (1863). *Utilitarianism*. London, England: Parker, Son, and Bourne.

Noë, R., & Hammerstein, P. (1994). Biological markets: supply and demand determine the effect of partner choice in cooperation, mutualism and mating. *Behavioral Ecology and Sociobiology*, *35*(1), 1–11. http://doi.org/10.1007/BF00167053

Norenzayan, A., Shariff, A. F., Gervais, W. M., Willard, A. K., McNamara, R. A., Slingerland, E., & Henrich, J. (2014). The cultural evolution of prosocial religions. *Behavioral and Brain Sciences*, 1–86.

Parfit, D. (1984). *Reasons and Persons*. Oxford University Press.

Rawls, J. (1971). *A Theory of Justice*. Cambridge, MA: Belknap Press of Harvard University Press.

Rockenbach, B., & Milinski, M. (2011). To qualify as a social partner, humans hide severe punishment, although their observed cooperativeness is decisive. *Proceedings of the*

National Academy of Sciences, *108*(45), 18307–18312. http://doi.org/10.1073/pnas.1108996108

Scanlon, T. M. (1982). Contractualism and Utilitarianism. In A. K. Sen & B. A. O. Williams (Eds.), *Utilitarianism and Beyond* (pp. 103–110). Cambridge, England: Cambridge University Press.

Scanlon, T. M. (1998). *What We Owe to Each Other* (Vol. 66). Cambridge, MA: Belknap Press of Harvard University Press.

Simonsohn, U. (2015). Small Telescopes Detectability and the Evaluation of Replication Results. *Psychological Science*, *26*(5), 559–569. http://doi.org/10.1177/0956797614567341

Smith, J. M., & Harper, D. (2003). *Animal Signals*. OUP Oxford.

Trivers, R. (1971). The Evolution of Reciprocal Altruism. *The Quarterly Review of Biology*, *46*(1), 35–57. http://doi.org/10.2307/2822435

Uhlmann, E. L., Zhu, L. (Lei), & Tannenbaum, D. (2013). When it takes a bad person to do the right thing. *Cognition*, *126*(2), 326–334. http://doi.org/10.1016/j.cognition.2012.10.005

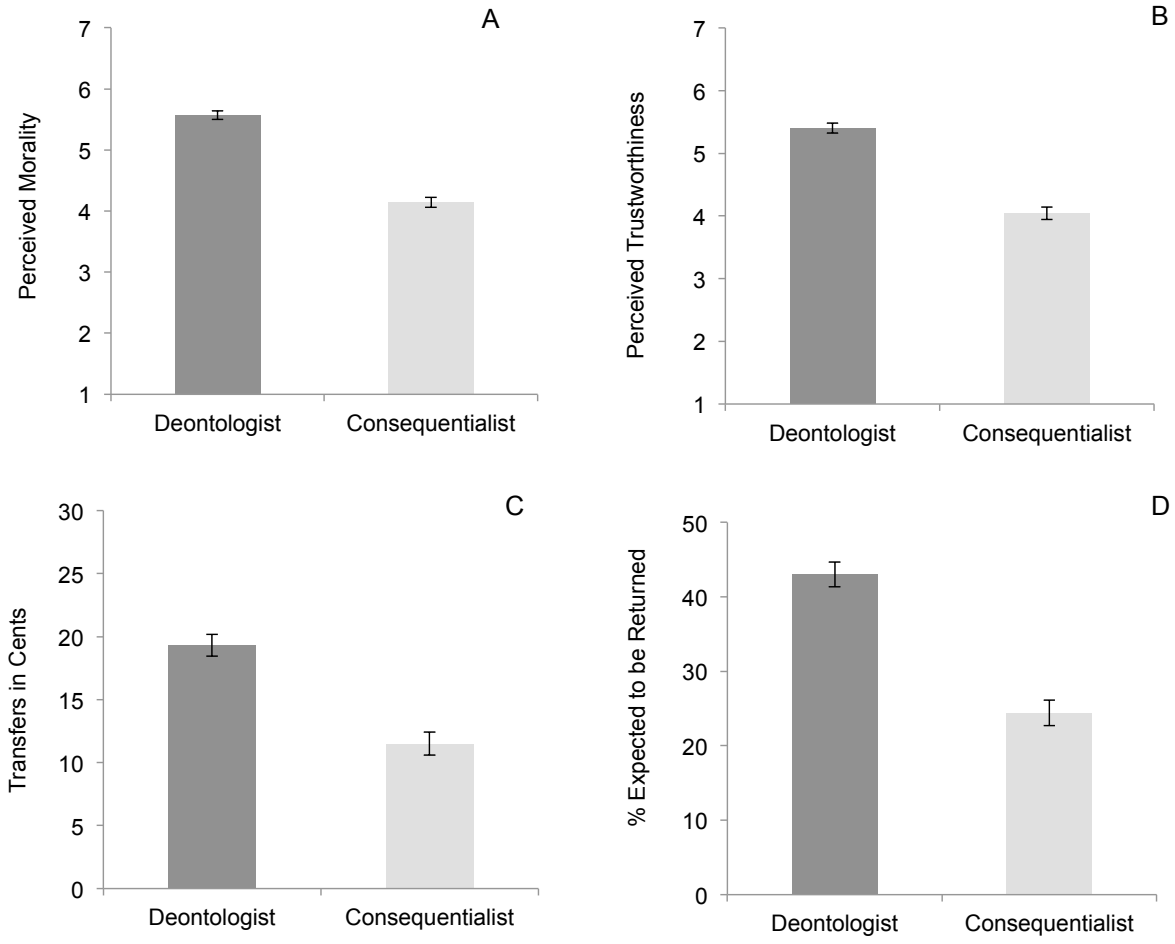

Figure 1. Agents who made deontological choices in the footbridge dilemma were rated as more moral (A) and trustworthy (B), and in a trust game received higher transfers (C), and were predicted to return more (D). Error bars represent *SE*s (Study 1b).

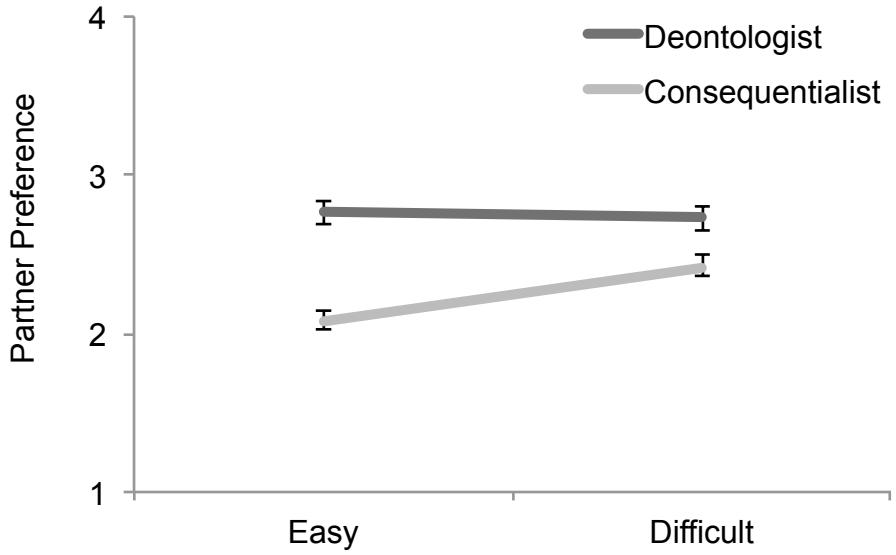

Figure 2. Expressions of choice conflict mitigate partner preferences for deontological agents. Participants strongly preferred deontological agents to consequentialist agents, but this preference was mitigated if the consequentialist agent reported difficulty in making the decision. Error bars represent *SE*s.

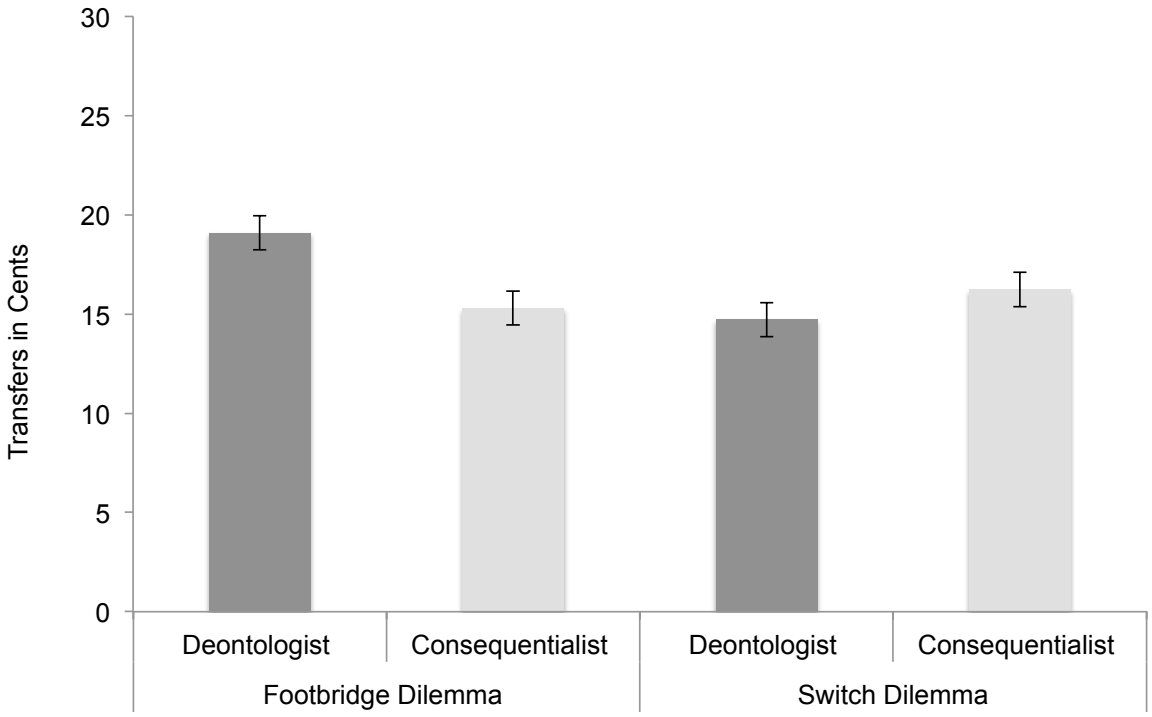

Figure 3. Preference for deontological agents is sensitive to respecting persons and not treating others as mere means. In Study 4b, an agent who made a deontological judgment in the footbridge dilemma, but not the switch, was trusted more in a trust game. Error bars represent *SE*s.

Table 1. Study 5b Design

Agent Decision and Justification	**Consent Condition**	**No-Consent Condition**
	"Please, kill me. I don't want to suffer at the hands of torturers"	"Please, don't kill me. I don't want to die out here in the field
Consequentialist	Endorses Sacrifice*	Endorses Sacrifice
	"*It is acceptable to kill someone if it reduces overall suffering*"	"*It is acceptable to kill someone if it reduces overall suffering*"
Categorical-Deontologist	Rejects Sacrifice	Rejects Sacrifice*
	"Killing people is just wrong, even if it has good consequences"	*"Killing people is just wrong, even if it has good consequences"*
Contractualist-Deontologist	Endorses Sacrifice*	Rejects Sacrifice*
	"It's right to kill the soldier if that's what they want, and it's the commander's duty to respect that"	*"It is wrong to kill the soldier if that's not what they want, and it's the commander's duty to respect that."*

Note: * indicates a conforming cases where the agent's judgment coheres with the soldier's request.