The Strategic Moral Self:

Self-Presentation Shapes Moral Dilemma Judgments

Sarah C. Rom[1] & Paul Conway[2]

[1]University of Cologne

[2]Florida State University

This version of the paper has been accepted for publication, but is not the official version of record. For the official version, polease see https://doi.org/10.1016/j.jesp.2017.08.003

Rom, S. C., & Conway, P. (2018). The strategic moral self: Self-presentation shapes moral dilemma judgments. *Journal of Experimental Social Psychology, 74*, 24-37.

Word count: 14,393

Author Note: Correspondence concerning this article should be sent to Sarah C. Rom, Department of Psychology, University of Cologne, Richard-Strauss-Str. 2, 50931, Cologne, Germany. Phone: +49 (0) 221 470 7760, Email: sarah.rom@uni-koeln.de. Paul Conway, Florida State University Department of Psychology. Email: conway@psy.fsu.edu

Abstract

Research has focused on the cognitive and affective processes underpinning dilemma judgments where causing harm maximizes outcomes. Yet, recent work indicates that lay perceivers infer the processes behind others' judgments, raising two new questions: whether decision-makers accurately anticipate the inferences perceivers draw from their judgments (i.e., meta-insight), and, whether decision-makers strategically modify judgments to present themselves favorably. Across seven studies, a) people correctly anticipated how their dilemma judgments would influence perceivers' ratings of their warmth and competence, though self-ratings differed (Studies 1-3), b) people strategically shifted public (but not private) dilemma judgments to present themselves as warm or competent depending on which traits the situation favored (Studies 4-6), and, c) self-presentation strategies augmented perceptions of the weaker trait implied by their judgment (Study 7). These results suggest that moral dilemma judgments arise out of more than just basic cognitive and affective processes; complex social considerations causally contribute to dilemma decision-making.

Keywords: moral dilemmas, social judgment, social perception, self-perception, meta-perception

The Strategic Moral Self:

Self-Presentation Shapes Moral Dilemma Judgments

During the Second World War, Alan Turing and his team cracked the Enigma Code encrypting German war communications. Soon, British High Command discovered an impending attack on Coventry—but taking countermeasures would reveal the decryption (Winterbotham, 1974). Thus, they faced a moral dilemma: allow the deadly raid to proceed and continue intercepting German communications, or deploy lifesaving countermeasures and blind themselves to future attack. Ultimately, the Allies allowed the attack to proceed. Lives were lost, but some analysts suggest this decision expedited the war's conclusion (Copeland, 2012). The moral judgment literature suggests that such decisions reflect a tension between basic affective processes rejecting harm and cognitive evaluations of outcomes allowing harm (Greene, 2014). But is it possible that self-presentation also factored in? The British High Command may have considered how their allies would react upon learning they threw away a tool for victory to prevent one deadly, but relatively modest, raid.

Moral dilemmas typically entail considering whether to accept harm to prevent even greater catastrophe. Philosophers originally developed such dilemmas to illustrate a distinction between killing someone as the means of saving others versus as a side effect of doing so (Foot, 1967), but subsequent theorists have largely described them as illustrating a conflict between deontological and utilitarian philosophy (e.g., Greene et al., 2001). The dual process model suggests that affective reactions to harm underlie decisions to reject harm, whereas cognitive evaluations of outcomes underlie decisions to accept harm (Greene, 2014). Other theorists have described these as processes in terms of basic cognitive architecture for decision-making

(Cushman, 2013; Crockett, 2013), or heuristic adherence to moral rules (Sunstein, 2005). Notably, all such existing models focus on relatively basic, non-social processing.

Yet, Haidt (2001) argued that moral judgments are intrinsically social, and communicate important information about the speaker. Indeed, recent work indicates that lay perceivers view decision-makers who reject harm (upholding deontology) as warmer, more moral, more trustworthy, more empathic, and more emotional than decision-makers who accept harm (upholding utilitarianism), whom perceivers view as more competent and logical, with consequences for hiring decisions (Everett, Pizarro, and Crocket, 2016; Kreps & Monin, 2015; Rom, Weiss, & Conway, 2016; Uhlmann, Zhu, and Tannenbaum, 2013).[1] Moreover, social pressure can influence dilemma judgments (Bostyn & Roets, 2016, Kundu & Cummins, 2012; Lucas & Livingstone, 2014). Such findings raise the question of whether people have meta-insight into how their dilemma judgments make them appear in the eyes of others, and whether decision-makers *strategically* adjust dilemma judgments to create desired social impressions. If so, this would provide the first evidence to our knowledge that higher-order processes causally influence judgments, suggesting dilemma decisions do not merely reflect the operation of basic affective and cognitive processes.

Moral Dilemma Judgments: Basic vs. Social Processes

Moral dilemmas originated as philosophical thought experiments, including the famous trolley dilemma where decision-makers could redirect a runaway trolley so it kills one person instead of five (Foot, 1967). According to Greene and colleagues (2001), refusing to cause harm

[1] Deontological dilemma judgments appear to convey both warmth and morality (Rom et al., 2015). Although these constructs can be disentangled (e.g., Brambilla et al., 2011), in the present context they happen to covary substantially. It may be that different aspects of deontological decisions influence these perceptions (e.g., whether they accord with moral rules; whether they suggest emotional processing), but these aspects overlap in the current paradigm. We focus primarily on perceptions of warmth, which roughly corresponds to the affective processing postulated by the dual process model, and relegated findings regarding morality the supplement. Future work should disentangle such perceptions from morality character evaluations.

to save others qualifies as a 'characteristically deontological' decision, because in deontological ethics the morality of action primarily hinges on its intrinsic nature (Kant, 1785/1959). Conversely, causing harm by redirecting the trolley saves five people, thereby qualifying as a 'characteristically utilitarian' decision, because in utilitarian ethics the morality of an action primarily hinges on its outcomes (Mill, 1861/1998).[2] Note that utilitarian philosophy technically entails impartial maximization of the greater good, which represents a subset of the broader concept of consequentialism, which advocates for outcome-focused decision-making more generally. We do not wish to imply that making a judgment consistent with utilitarianism renders one a utilitarian—it need not (e.g., Kahane, 2015)—but rather we use the term 'utilitarian' in the simpler sense that such judgments a) objectively maximize overall outcomes, b) appear to often entail ordinary cost-benefit reasoning, and c) utilitarian/consequentialist philosophers generally approve of such judgments.

Although dilemmas originated in philosophy, research in psychology, neuroscience, and experimental philosophy has aimed to clarify the psychological mechanisms driving dilemma judgments. Most prominent among these is the dual process model, which postulates that basic affective and cognitive processes drive dilemma judgments (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001). Other theorists have argued judgments reflect decision-making systems focused on immediate action versus long-range goals (Cushman, 2013; Crockett, 2013), heuristic adherence to moral rules (Sunstein, 2005), or the application of innate moral grammar (Mikhail,

[2] Following Greene and colleagues (2001), we use the term 'characteristically' deontological/utilitarian, because there are many variants of each theory that do not all agree. Nonetheless, this terminology is widely employed currently, and so we follow in this terminological tradition despite its limitations. Note that we are not arguing that making a given dilemma decision implies that decision-makers ascribe to abstract philosophical commitments. Rather, we argue simply that 'utilitarian' judgments qualify as such because they tend to maximize outcomes, regardless of decision-makers' philosophical commitments. Just as one need not be Italian to cook an Italian meal, accepting outcome-maximizing harm on a dilemma does not make one a utilitarian; however, utilitarian philosophers typically agree that this decision is morally acceptable.

2007). We do not aim to adjudicate between these various claims, nor do we dispute the contribution of such processes. Rather, we simply note that these models focus on basic, nonsocial processes.

Research has largely ignored the possibility that higher-order sophisticated social processes might causally contribute to dilemma judgments. Yet, morality appears intrinsically social (Haidt, 2001), and most real-world moral judgments involve publicly communicating with others (e.g., Hofmann, Wisneski, Brandt, & Skitka, 2014). We expect the same is true of dilemma judgments. Although the best-known dilemmas are hypothetical (such as the trolley dilemma), many real-world decisions entail causing harm to improve overall outcomes (e.g., launching airstrikes in Syria to prevent ISIS from gaining momentum, punishing naughty children to improve future behavior, imposing fines to prevent speeding). As decisions in such cases align with either deontological or utilitarian ethical positions, they correspond to real world moral dilemmas. Moreover, lay decision-makers employ verbal arguments that align with deontological and utilitarian ethical positions (Kreps & Monin, 2014). Hence, social consideration of dilemma judgments is not restricted to responses to hypothetical scenarios, but forms an ordinary part of communication about common moral situations.

Kreps and Monin (2014) examined deontological and utilitarian arguments in speeches by Presidents Clinton and Bush, among other politicians. Lay perceivers viewed speakers as moralizing more when they framed arguments in terms of deontology rather than utilitarianism. These findings align with work on hypothetical dilemma decisions: perceivers rated and treated decision-makers who rejected harm (upholding deontology) as more trustworthy than decision-makers who accept harm (upholding utilitarianism, Everett et al., 2016), as well as more moral, more empathic, and less pragmatic than harm-accepting decision-makers (Uhlmann et al., 2013).

Likewise, Rom and colleagues (2016) found that lay people appear to intuit the dual process model: they rated targets who rejected harm as relatively warm, and inferred that such judgments were driven by emotion. Conversely, perceivers rated targets who accepted harm as relatively competent, and inferred that such judgments were driven by cognitive deliberation.[3] Moreover, perceivers preferred harm-rejecting decision-makers for social roles prioritizing warmth, such as social partners or their child's doctor, but preferred harm-accepting decision-makers for roles prioritizing competence, such as hospital administration (Everett et al., 2016, Rom et al., 2016). Hence, decision-makers face a warmth/competence tradeoff when presenting their decision to others. The current work examines whether decision-makers are aware of this trade-off, and whether they strategically adjust their decisions to present themselves favorably.

Meta-Perceptions Regarding Dilemma Judgments

We propose that lay perceivers hold fairly accurate meta-perceptions into how others will view them based on their dilemma decision. People care deeply about their moral reputation (Aquino & Reed, 2002; Krebs, 2011; Everett et al., 2016) and the moral reputations of others (Brambilla, Rusconi, Sacchi, & Cherubini, 2011; Goodwin, Piazza, & Rozin, 2014). Clearly, the research described above on perceptions of decision-makers indicate that dilemma decisions can

[3] If the dual-process model is correct, responses to classic moral dilemmas do not reflect the degree to which decision-makers experience affective reactions or engage in cognition in an absolute sense. If classic moral dilemmas place affect and cognition in conflict, and ultimately judges may only choose one option, then judgments reflect the *relative* strength of each process. For example, accepting harm that maximizes outcomes may occur either due to strong cognition coupled with strong but slightly weaker affect, or weak cognition coupled with weaker affect. Hence, a judgment to accept causing harm does not reveal whether the judge experienced strong or weak affect—only that cognition outweighed whatever degree of affect they experienced. Nor does such a judgment guarantee that the judge engaged in strong cognition—only that whatever cognition they engaged in outweighed their affective experience. Some people may engage in extensive affect and cognition, whereas others engage in little of either. In order to estimate each processes independently, it is necessary to use a technique such as process dissociation (see Conway & Gawronski, 2013). However, in the current work we are not interested in the actual processes underlying dilemma judgments so much as lay perceptions of these processes. To that end, lay people, like many researchers, equate harm avoidance judgments with strong affect and harm acceptance judgments with strong cognition. This intuition is effective as a rough heuristic, so long as researchers recognize that it does not accurately describe moral dilemma processing.

affect moral reputation, suggesting that people should be attuned to what messages their judgments convey. Moreover, past work suggests that people can be reasonably accurate when gauging how others perceive them. For example, narcissists appear aware that others view them less positively than they view themselves (Carlson, Vazire, & Furr, 2011, Carlson & Furr, 2009). Self- and social-ratings particularly converge when the underlying traits entail public behaviors (e.g., loquaciousness signals extraversion) rather than inner states (e.g., neurotic feelings, Vazire, 2010). Sharing one's dilemma judgment entails a clear public behavior, suggesting relative accuracy in meta-perceptions.

However, other research casts doubt on the possibility of accurate dilemma meta-perceptions in dilemma research. Besides public expression, dilemma judgments entail intrapsychic aspects such as emotional reactions, perceptions of conflict, and so on (e.g., Kruger & Gilovich, 2004; Anderson & Ross, 1984; Pronin, 2008; Winkielman & Schwarz, 2001). Decision-makers hold privileged knowledge of their experience of these inner states. People often fail to consider that others have access to less information than they do (Chambers, Epley, Savitsky, & Windschitl, 2008). Whereas egocentric perspectives come to mind easily, adjusting away from egocentricity is difficult (Epley, Keysar, Van Boven, & Gilovich, 2004). Thus, meta-perceptions are often biased by self-understanding (Chambers et al., 2008; Kaplan, Santuzzi, & Ruscher, 2009; Kenny & DePaulo, 1993). Moreover, people are motivated to view themselves positively in the moral domain (Epley & Dunning, 2000) much like non-moral domains (e.g., Dunning & McElwee, 1995), and can rationalize either dilemma decision in self-flattering ways (Uhlmann, Pizarro, Tannenbaum, & Ditto, 2009; Liu & Ditto, 2014). Thus, people may well judge themselves as high in both warmth and competence regardless of their dilemma decision—and may expect others to agree with this flattering self-assessment.

If decision-makers erroneously base meta-perceptions on self-perceptions, meta-perception ratings should converge with self-ratings and diverge from ratings of others following the same judgment—that is, people may believe they come across as both warm and competent regardless of their dilemma decision, whereas they view others' decisions as reflecting a warmth/competence trade-off. Conversely, if people have accurate meta-insight into how others perceive them, meta-perception ratings should converge with other ratings and diverge from self-ratings—that is, people may privately believe they are warm and competent regardless of dilemma decision, yet expect others to rate them according to the same warmth/competence tradeoff implied by others' judgments. We contrasted these predictions empirically.

Strategic Self-Presentation in Dilemma Judgments

If people evince accurate meta-insight into what their dilemma decision conveys, this raises the possibility that they strategically adjust such decisions to present themselves favorably. There are potential upsides and downsides to selecting each dilemma judgment, as the precise cause of others' dilemma decisions appear ambiguous. Upholding utilitarianism by accepting outcome-maximizing harm amounts to bloodying one's hands for the sake of the community. Such bold and brutal action may convey either competent leadership (Lucas & Galinksy, 2015) or a callous disregard for causing harm—as in psychopathy (Bartels & Pizarro, 2011) or low empathy (Gleichgerrcht & Young, 2013). Conversely, rejecting harm (upholding deontology) may convey either a warm concern for others and/or principled respect for life and/or trustworthiness (Everett et al., 2016; Kreps & Monin, 2015; Rom et al., 2016), or suggest incompetent paralysis when the situation demands bold action (Gold et al 2015; Gawronksi et al., 2015). Hence, in some circumstances it may be preferable to risk appearing incompetent in order

to convey warmth, trustworthiness, and respect for life; in other situations, it may be preferable to risk appearing cold and callous in order to convey decisive competence and leadership.

People care deeply about presenting themselves favorably. They tailor their public images in various domains to the perceived values and preferences of important others (Reis & Gruzen, 1976; von Baeyer, Sherk, & Zanna, 1981; Leary & Kowalski, 1990; Leary, 1995). People change social roles over time, and social roles carry expectations regarding how individuals who occupy those roles ought to behave (Sarbin & Allen, 1968). Hence, people often flexibly present themselves to conform to different social role expectations (Leary, Robertson, Barnes, & Miller, 1986; Leary, 1989). Indeed, Everett and colleagues (2016) argued that deontological dilemma judgments may operate as a reputation-management mechanism to present oneself as a trustworthy social interaction partner by demonstrating respect for others autonomy and wishes (see also Bostyn & Roets, 2016).

Accordingly, previous work demonstrates that social situations influence dilemma responses. In a modification of the Asch conformity paradigm, Kundu and Cummins (2012) asked participants whether they would accept or reject outcome-maximizing harm after a series of confederates gave a particular answer. They found evidence for conformity pressure: participants were more likely to give answers consistent with those of the confederates. Bostyn and Roets (2016) conducted a similar study, and argued that conformity pressure was stronger for harm rejection (upholding deontology) than harm acceptance (upholding utilitarianism). However, Lucas and Livingstone (2014) found that participants who socially connected with others before completing dilemmas after were more willing to accept harm (upholding utilitarianism). It may be that resolving dilemmas in front of strangers motivated participants to skew towards deontological answers so as to avoid appearing immoral—after all, research

suggests that moral traits appear especially important when forming first impressions (Brambilla et al., 2011; Goodwin et al., 2014), and that warmth may also be important when forming first impressions (Fiske, Cuddy, & Glick, 2007). Conversely, when participants have an opportunity to establish warmth or morality through social interactions, they may have felt free to demonstrate other qualities, such as competence. These findings suggest that context may shift whether accepting or rejecting harm seems to be the optimal answer. If participants strategically adjust dilemma judgments, their perception of expectations should vary depending on whether the circumstances appear to prioritize warmth over competence, and their public (but not private) dilemma answers should track such expectations.

Overview

Across seven studies, we investigated whether people hold accurate meta-perceptions regarding how others view them based on their dilemma judgments, and whether they strategically modify such judgments to present themselves favorably. First, we examined whether people have accurate meta-insight into the warmth and competence ratings others infer from their dilemma judgments by comparing warmth and competence ratings of others, the self, and meta-perceptions of the self (Studies 1-3). Second, we tested whether people shift public (but not private) dilemma judgments depending on whether warmth or competence is favored in a given situation (Studies 4-6). Third, we investigated whether people can use communication strategies to offset the weaker trait implied by their judgment—whether people who accept harm can come across as warm, and people who reject harm can come across as competent (Study 7). Across all studies, we disclose all measures, manipulations, and exclusions, as well as the method of determining the final sample size. In none of the studies data collection was continued after data analysis.

Study 1

Study 1 examined the accuracy of participants' meta-perceptions (i.e., meta-accuracy, Anderson, Ames, & Gosling, 2008) following moral dilemma judgments. We randomly assigned participants to one of three conditions: participants either made a dilemma judgment themselves (self and meta-perception condition) or read about another persons' dilemma judgment (other condition). Then, participants in the self-condition rated their own warmth and competence, those in the other condition rated the others' warmth and competence, and those in the meta-perception condition rated how they believed others would view their warmth and competence. Hence, the design was a 3 (target: self vs. other vs. meta-perceptions) × 2 (decision: harm inappropriate vs. appropriate) × 2 (personality dimension: warmth vs. competence) quasi-experimental design (as participants were free to make either dilemma judgment themselves) with the first two factors between-subjects and the third within-subjects.

Given that people tend to view themselves positively in the moral domain (Epley & Dunning, 2000), and have access to internal perceptions of conflict between response options, we expected participants in the self condition would rate themselves high on both warmth and competence, regardless of their dilemma decision. We expected participants in the other condition to replicate the patterns demonstrated by Rom and colleagues (2016): they should rate targets who rejected causing harm as warmer but less competent than targets who accepted causing outcome-maximizing harm. Most importantly, we predicted that participants' meta-perception condition would exhibit meta-accuracy, by anticipating that others would rate them using the same warmth/competence tradeoff (depending on dilemma decision) as participants in the other condition, rather than the uniformly high warmth and competence ratings participants privately make about themselves.

Method

Participants. We decided to recruit 200 American participants (134 males, 66 females, M_{age} = 30.63, *SD* = 8.92) via Mechanical Turk, who received $0.25, using a heuristic of aiming for ~50 per between-subjects condition, although actual responses varied substantially ($N_{self_harm_rejection}$ = 14; $N_{self\_harm\_acceptance}$ = 30; $N_{other\_harm\_rejection}$ = 54; $N_{other\_harm\_acceptance}$ = 46; $N_{meta\_harm\_rejection}$ = 14; $N_{meta\_harm\_acceptance}$ = 42). We excluded no one. Although we did not conduct a priori power analyses, we felt confident that this design provided reasonable power based on past work (Rom et al., 2016). Indeed, a post hoc power analysis using GPower (Faul, Erdfelder, Lang, & Buchner, 2007) for a fixed-effects between-within design where η_p^2 = .10, *N* = 200, α = .05, and the correlation between repeated measures was *r* = .33 suggested that we had ~99% power to detect the obtained interaction.

Procedure. All participants read the widely-employed crying baby dilemma (e.g., Conway & Gawronski, 2013), where the actor must decide whether to smother a baby to prevent its cries from alerting murderous soldiers hunting for other townspeople in hiding. Participants in the self and meta-perception conditions then selected either *yes, this action is appropriate* or *no, this action is not appropriate* (following Greene et al., 2001). Participants in the other condition viewed a photo of a university student named Brad, then learned that Brad had selected either one or the other of these responses (following Rom et al., 2016). Then, participants completed measures of warmth and competence using items adapted from Fiske, Cuddy, Glick and Xu (2002).

Depending on condition, participants either rated themselves, Brad, or indicated how they thought others would rate them following their decision (meta-perception). Specifically, those in the meta-perception condition read:

Now take a moment to imagine that another person saw the judgment you made. Based on that information, what would they think about you? From their perspective how well do you think they would say each trait describes you? THEY would think you are...

Participants indicated how well four warmth traits (*warm, good-natured, tolerant, sincere*) and five competence traits (*competent, confident, independent, competitive, intelligent*) described the target on 7-point scales anchored at 1 (*not at all*) and 7 (*very much*). We averaged judgments into composites of warmth ($\alpha = .91$) and competence ($\alpha = .87$), which were modestly correlated ($r = .33$). Item order was randomized for each participant. For exploratory reasons, we also included the single item *moral*, consistent with Rom and colleagues (2016).

Some researchers have argued that morality and warmth are distinguishable constructs (Brambilla, et al. 2011; Goodwin et al., 2014). We find these arguments persuasive—used car salesmen that evince warm sociability should not be trusted, whereas a cold and dispassionate judge who sentences criminals may nonetheless appear moral. Nonetheless, it may be that these constructs align more in some contexts than others. Hence, we empirically examined how well these constructs dissociated in the current studies using five strategies.

First, we noted that the item *moral* consistently correlated highly with the warmth composite measure, $\sim r = .75$, consistent with Rom and colleagues (2016). Second, we noted that the item *moral* varied across conditions in the same manner as the warmth composite on all studies (see Supplementary analysis). Now, it remains possible that these findings simply reflect the fact that some items in the warmth composite—such as *sincerity*—assess perceptions morality instead of warmth. Therefore, third, we conducted factor analyses (principle axis factoring with oblimin rotation) for all studies assessing warmth and morality (see supplementary material). In each case, all warmth items loaded together with the item *moral* onto a single

factor, whereas the competence items loaded onto a separate factor. Fourth, we conducted follow-up analyses for each study using only the single items warmth and competent instead of the composite measures; findings were very similar (find an example for Study 1 in the supplementary material). Fifth, we conducted follow-up analyses for each study using an alternative warmth score based on two items (*warm, good-natured*), and an alternative morality score based on three items (*sincere, tolerant, morality*),[4] as well as a combined warmth/morality score including all warmth items plus the item morality. In each case, the pattern of findings remained very similar to the patterns presented below.

These findings suggest that in the context of dilemma perceptions, participants may find it difficult to disentangle warmth and morality. After all, perceivers may find it ambiguous whether a given deontological judgment reflects affective processing or adherence to moral rules. Alternatively, it may be that the particular items presented in this scale underestimate the difference between these constructs. Either way, the current paradigm was not designed to distinguish between warmth and morality. Indeed, these analyses suggest it may even be warranted to include the item moral in the warmth composite measure. Nonetheless, in recognition of the important theoretical distinction between warmth and morality (Brambilla & Leach, 2011; Goodwin et al., 2014) and to remain consistent with Rom and colleagues (2016), we decided to treat the item morality as a separate construct. Given that the current focus was on contrasting perceptions of warmth and competence, and the similarity between the patterns of warmth and morality, we decided to relegate the morality findings to the supplementary material.

Results

[4] We thank an anonymous reviewer for this suggestion.

We submitted ratings to a 3 (target: self vs. other vs. meta-perceptions) × 2 (decision: harm inappropriate vs. appropriate) × 2 (personality dimension: warmth vs. competence) repeated measures ANOVA with the first two factors between and the last factor within subjects (see Figure 1). There was a main effect of target: participants gave higher ratings overall in the self (M = 5.18, SD = 1.12) than other (M = 4.64, SD = .90), or meta-perception conditions (M = 4.28, SD = 1.04), $F(2, 194) = 8.47$, $p < .001$, $\eta_p^2 = .08$. There was also a main effect of decision: participants rated targets who rejected harm, upholding deontology, higher overall (M = 4.86, SD = 1.10), than targets who accepted harm, upholding utilitarianism (M = 4.51, SD = 1.03), $F(2, 194) = 8.32$, $p = .004$, $\eta_p^2 = .04$. There was no main effect of personality dimension, $F(2, 194) = 1.75$, $p = .18$, $\eta_p^2 = .01$. These main effects were qualified by a significant two-way interaction between target decision and personality measure, $F(1, 194) = 45.65$, $p < .001$, $\eta_p^2 = .19$, and a marginal interaction between target and personality measure, $F(2, 194) = 3.03$, $p = .050$, $\eta_p^2 = .03$, 95%, whereas the interaction between target and decision was not significant, $F(2, 194) = 1.55$, $p = .214$, $\eta_p^2 = .02$. Moreover, the three-way interaction was significant, $F(2, 194) = 11.14$, $p < .001$, $\eta_p^2 = .10$.

We decomposed these interactions by examining post-hoc tests within each condition. As predicted, participants in the self-condition rated themselves equally high on warmth when they rejected (M = 5.69, SD = 1.24) or accepted (M = 5.12, SD = 1.40) causing harm, $F(1,194) = 2.70$, $p = .102$, $\eta_p^2 = .01$, and equally competent when they rejected (M = 5.50, SD = 1.17) versus accepted causing harm (M = 4.85, SD = 1.94), $F(1,194) = 3.60$, $p = .059$, $\eta_p^2 = .03$. However, participants in the other-condition replicated the predicted warmth/competence tradeoff found previously: Participants rated Brad higher on warmth when he rejected (M = 5.00, SD = 1.19), than when he accepted causing outcome-maximizing harm (M = 4.03, SD = .99), $F(1, 194) =$

15.57, $p < .001$, $\eta_p^2 = .07$. Conversely, they rated Brad as higher in competence when he accepted ($M = 5.16$, $SD = 1.16$), rather than rejected causing outcome-maximizing harm ($M = 3.36$, $SD = 1.31$),-$F(1, 194) = 11.67$, $p < .001$, $\eta_p^2 = .06$.

Crucially, participants in the meta-perception-condition evinced the same warmth/competence tradeoff as participants in the other-condition: When participants rejected harm they inferred others would perceive them as warmer ($M = 5.16$, $SD = 1.59$) than when they accepted causing outcome-maximizing harm ($M = 3.36$, $SD = 1.31$), $F(1, 194) = 22.95$, $p < .001$, $\eta_p^2 = .10$. In contrast, when they accepted such harm, they inferred that others would perceive them as (slightly) more competent ($M = 4.89$, $SD = 1.10$) than when they rejected such harm ($M = 4.38$, $SD = 1.46$), although results did not reach conventional levels of significance, $F(1, 194) = 2.32$, $p = .129$, $\eta_p^2 = .01$.

Discussion

These findings suggest that participants have accurate meta-insight regarding the inferences others will draw about their personality from their dilemma judgments. Privately, participants rated themselves equally high on warmth and competence regardless of their dilemma decision. However, in the meta-perception condition they expected others to rate them similar to how they rated others: just as participants viewed targets who rejected causing harm as warmer and less competent than targets who accepted causing harm, they expected that others would rate them as warmer and less competent when they rejected vs. accepted causing harm themselves. To our knowledge, this is the first evidence that participants are aware of the impression their dilemma judgments convey to others.

However, our quasi-experimental design suffered from the limitation of nonrandom assignment: participants in the self and meta-perception conditions freely choose which dilemma

decision to make. Hence, it remains possible that our meta-perception results reflect the general psychology of people who made a specific decision, rather than inferences regarding that decision per se. Even though this interpretation seems unlikely give the null effect in the private self-rating condition, we aimed to resolve this confound in Study 2.

Study 2

Study 2 replicated the meta-perception condition from Study 1, together with a *communication error* condition where participants imagined that others erroneously learned they made the dilemma judgment opposite to the one they truly made. This design allowed us to test whether meta-perceptions in Study 1 would hold for decisions that participants personally disagreed with. We expected that warmth and competence meta-perceptions would track the decision others believed participants made (harm rejection: higher warmth than competence, harm acceptance: higher competence than warmth), rather than the decision participants actually made.

Method

Participants. To increase confidence in the effects, we decided to approximately double the sample size from Study 1, recruiting 397 American participants via Mechanical Turk, who received $0.25. We excluded 24 participants who completed less than 50% of dependent measures, leaving a final sample of 373 (244 males, 123 females, 6 unreported, M_{age} = 30.49, *SD* = 9.89, $N_{correct\_harm\_rejection}$= 32; $N_{correct\_harm\_acceptance}$ = 157; $N_{error\_harm\_rejection}$ = 45; $N_{error\_harm\_acceptance}$ = 139). Gpower suggested we had ~99% post-hoc power to detect the obtained interaction with this sample size.

Procedure. Each participant read the crying baby dilemma from Study 1, and selected one of the two dilemma responses. Then we randomly assigned them to the *correct*

communication or *communication error* condition. Participants in the correct communication condition imagined that others correctly learned which dilemma decision they made, as in Study 1. Participants in the communication error condition imagined that others erroneously learned they made the dilemma decision opposite to their real decision. Specifically, they read:

> *Now take a moment to imagine that another person learned about the judgment you made. As often happens, misinformation got out and this other person thinks you chose: Yes, harm is appropriate [No harm is not appropriate]. Based on the information that you would [not] SMOTHER the baby, what would this person think of you? From their perspective how well do you think they would say each trait describes you? THEY would think you are...*

Participants indicated how they believed others would perceive them on the same warmth ($\alpha = .89$), competence ($\alpha = .87$), and morality items as Study 1. This resulted in a 2 (communication: correct vs. error) × 2 (decision: harm inappropriate vs. appropriate) × 2 (dimension: warmth vs. competence) quasi-experimental design, as participants could not be randomly assigned to make a particular judgment. Consistent with Study 1 and past work (Rom et al., 2016), warmth and competence correlated moderately ($r = .40$), whereas morality correlated highly with warmth ($r = .87$) and less with competence ($r = .38$). Morality yielded results similar to warmth across condition (replicating previous work, Rom et al., 2016) but was not focus of the current manuscript, so we again relegated it to the supplement.

Results

We submitted warmth and competence ratings to a 2 (communication: correct vs. error) × 2 (decision: harm unacceptable vs. acceptable) × 2 (dimension: warmth vs. competence) repeated measures ANOVA (see Figure 2) with the first two factors between-subjects and the last factor

within-subjects. There was a main effect of communication: participants gave higher personality ratings overall in the correct communication (M = 4.25, SD = 1.26) than communication error condition (M = 3.98, SD = 1.35), $F(1, 369) = 17.51$, $p < .001$, $\eta_p^2 = .05$. There was no main effect of decision, $F(1, 369) = 0.46$, $p = .499$, $\eta_p^2 = .001$, but there was a main effect of personality dimension: participants gave lower warmth (M = 3.89, SD = 1.81) than competence ratings overall (M = 4.35, SD = 1.47), $F(1, 369) = 7.30$, $p = .007$, $\eta_p^2 = .02$. In addition, there were significant 2-way interactions between communication and personality dimension, $F(1, 369) = 19.26$, $p < .001$, $\eta_p^2 = .05$, and between decision and personality dimension, $F(1, 369) = 4.43$, $p = .036$, $\eta_p^2 = .01$, though not between communication and personality dimension, $F(1, 369) = 2.25$, $p = .134$, $\eta_p^2 = .01$. More importantly, we obtained the expected three-way interaction, $F(1, 369) = 49.02$, $p < .001$, $\eta_p^2 = .12$.

Post-hoc contrasts largely replicated Study 1 in the correct communication condition: Participants expected that others would rate them as warmer when they rejected harm, upholding deontology (M = 5.06, SD = 1.49) than accepted causing harm, upholding utilitarianism (M = 3.40, SD = 1.68), $F(1, 182) = 19.70$, $p < .001$, $\eta_p^2 = .10$. Results for competence trended in the expected direction, but did not reach significance: participants expected that others would rate them as similarly competent when they rejected (M = 4.51, SD = 1.19), rather than accepted, causing harm (M = 4.82, SD = 1.19), $F(1, 187) = 1.91$, $p = .168$, $\eta_p^2 = .01$. Participants in the error communications condition showed the opposite pattern. Participants expected that others would rate them as less warm when they rejected (M = 2.94, SD = 1.93) rather than accepted causing harm (M = 4.42, SD = 1.68), $F(1, 369) = 22.28$, $p < .001$, $\eta_p^2 = .11$. Again, ratings for competence trended nonsignificantly in the expected direction: participants expected others to

rate them similarly on competence when they rejected (M = 3.66, SD = 1.40), versus accepted causing harm (M = 4.00, SD = 1.29), $F(1, 369) = 2.25$, $p = .135$, $\eta_p^2 = .01$.

Discussion

Study 2 replicated the findings from Study 1 in the correct communication condition: participants who rejected harm (upholding deontology) accurately inferred that others would perceive them as relatively warmer but (nonsignificantly) less competent, whereas participants who accepted harm (upholding utilitarianism) accurately inferred that others would perceive them as (nonsignificantly) more competent but less warm. Moreover, these meta-perception ratings flipped when participants imagined that a communication error occurred, and others erroneously believed they made the judgment opposite to the judgment they actually made. Hence, meta-perceptions tracked the information available to others, rather than reflecting the judgments participants actually made. This finding rules out the possibility that the Study 1 meta-perception findings were driven by individual differences in meta-perceptions among people who rejected versus accepted harm, thereby overcoming the limitation of employing quasi-experimental designs. However, thus far we have examined meta-perceptions using only the crying baby dilemma in American MTurk samples. To improve generalizability, we examined whether these effects would replicate using a whole battery of dilemmas and an in-lab sample of German-speaking student participants.

Study 3

Study 3 examined whether the meta-perception findings in Studies 1 and 2 would generalize to other dilemmas and samples. Thus, we recruited a laboratory sample of German-speaking university students and broadened the stimulus set by translating a standardized battery of 10 dilemmas into German, and randomly presenting participants with one of the ten dilemmas

from this battery (Conway & Gawronski, 2013).

Method

Participants. We obtained 131 German university students (55 males, 75 females, 1 other, 2 no gender indication, M_{age} = 30.49, *SD* = 9.90) who received $0.25. Again, we aimed for ~50 participants per cell, and excluded no one ($N_{harm_rejection}$ = 66; $N_{harm\_acceptance}$ = 65). Again, we had ~99% power to detect the obtained interaction.

Procedure. The design was similar to the meta-perceptions condition in Study 1. Each participant read one dilemma at random from a battery of 10 dilemmas, selected either accept or reject outcome-maximizing harm as in Study 1, and completed the same meta-perception measures of warmth (α = .89), competence (α =.87) as in Study 1 (this time we did not measure morality). The battery consisted of the 10 incongruent dilemmas from Conway and Gawronski (2013), where causing harm always maximized outcomes. The crying baby and vaccine dilemmas from Study 1 are examples of incongruent dilemmas from this set. Other examples include the *torture dilemma* (is it appropriate to torture a man in order to stop a bomb that will kill people?), and the *car accident dilemma* (is it appropriate to run over a grandmother in order to avoid running over a mother and child?). As before, this resulted in a 2 (decision: harm inappropriate vs. appropriate) × 2 (dimension: warmth vs. competence) quasi-experimental design. Again, warmth and competence correlated moderately (*r* = .28).

Results

We submitted warmth and competence ratings to a 2 (decision: harm inappropriate vs. appropriate) × 2 (dimension: warmth vs. competence) repeated measures ANOVA with the first factor between-subjects and the second factor within-subjects. There was a main effect of target decision: Participants who rejected harm, upholding deontology, reported higher meta-perception

ratings overall (M = 4.76, SD = 1.14), than participants who accepted harm, upholding utilitarianism (M = 4.16, SD = 1.26), $F(1, 187) = 6.44$, $p = .012$, $\eta_p^2 = .03$. There was also main effect of personality dimension: participants reported lower overall meta-perception ratings for warmth (M = 3.74, SD = 1.76) than competence (M = 4.78, SD = 1.16), $F(1, 187) = 9.06$, $p = .003$, $\eta_p^2 = .05$. More importantly, these results were qualified by the predicted interaction, $F(1, 187) = 42.58$, $p < .001$, $\eta_p^2 = .19$.

Post-hoc tests revealed the same warmth/competence tradeoff as in Studies 1 and 2: Participants expected that others would rate them as warmer when they rejected (M = 5.02, SD = 1.49) rather than accepted causing harm (M = 3.48, SD = 1.71), $F(1, 129) = 25.75$, $p < .001$, $\eta_p^2 = .17$. Conversely, participants expected that others would rate them as less competent when they rejected (M = 4.51, SD = 1.19) rather than accepted causing harm (M = 4.82, SD = 1.19), $F(1, 129) = 13.98$, $p < .001$, $\eta_p^2 = .01$.

Discussion

Study 3 replicated and generalized the meta-perception findings from Studies 1 and 2 to a different sample and broader array of dilemma stimuli. These findings increase confidence in the claim that participants in both Germany and the United States hold accurate meta-perceptions regarding how their judgments on many dilemmas make them appear to others—namely, participants are aware of the warmth/competence perception tradeoff implied by dilemma judgments. Next, we turn to the possibility that people use this meta-perception information to adjust their dilemma decisions to strategically present themselves as relatively warm or competent depending on which trait is most valued in a given context.

Study 4

In Study 4, we examined whether people sometimes strategically adjust their private dilemma judgments to mesh with social expectations. We randomly assigned participants to learn that the study was ostensibly comparing either the intellectual or emotional abilities of people in different university degree programs. Part of the study involved responding to moral dilemmas. If people consider self-presentation when answering dilemmas, they should be more likely to reject harm when they think the study assessed emotional competency (i.e., warmth), and more likely to accept outcome-maximizing harm when they think the study is about intellectual ability (i.e., competence).

Method

Participants. We obtained 120 German participants (57 males, 63 females, M_{age} = 22.99, *SD* = 4.41) from a large University in Western Germany, who received € 2.00. Participants were randomly assigned to a condition prioritizing logical reasoning (associated with competence) or emotional competency (associated with warmth, Rom et al., 2016). Sample size was predetermined at ~50 participants per cell, though we ended up collecting a few extra people. No participants were excluded. ($N_{emotion}$ = 60; N_{logic} = 60). Again, Gpower suggested we had ~99% power to detect the obtained interaction.

Procedure. To manipulate the perceived importance or warmth and competence, we randomly assigned participants to read the following instructions: *This is a study to measure the logical reasoning ability (or emotional competency) between people in different degree programs. Please imagine the following situation and tell us your solution.* Then we presented them with three dilemmas, presented on individual screens, in a fixed random order. We again employed the same ten dilemmas by randomly presenting dilemmas from a standardized battery as in Study 3 (Conway & Gawronski, 2013). Participants indicated how much they accepted vs.

rejected such outcome-maximizing harm on scales from *harm is not acceptable* (1) to *harm is acceptable* (7). We averaged ratings to form an aggregate score of relative harm acceptability.

Results and Discussion

As predicted, participants indicated that causing harm to maximize outcomes was relatively more appropriate in the condition emphasizing logic (M = 4.74, SD = 1.50), versus emotional ability (M = 4.06, SD = 1.45), $t(118) = -2.53$, $p = .013$, $d = 1.70$. This finding provides initial evidence suggesting that participants may modify dilemma answers to present themselves as relatively warm or competent, depending on which trait is prioritized in the current context. However, it remains unclear whether this effect reflects strategic self-presentation or whether the instructions simply primed participants to focus more on emotion or logic when forming judgments (similar to Valdesolo & DeSteno, 2007). Therefore, in Study 5 we assessed not only which dilemma decision participants report, but also which decision they believed others expected them to make. If strategic self-presentation plays a role in dilemma judgments, then both actual and expected decisions should reflect the influence of the context manipulation. We also employed a new manipulation to increase generalizability.

Study 5

In Study 5 we examined whether both expected and reported dilemma judgments conform to social role expectations. Participants imagined they were applying for a job as a military physician, and one interview question involved a moral dilemma. We manipulated whether warmth or competence was valued most by emphasizing either the military (competence) or medical treatment (warmth) aspects of the position. Then participants reported which dilemma answer they thought interviewers expected, as well as their actual decision. If people adjust their dilemma judgments to conform to social role expectations, then participants

in the military condition should be more likely to infer and report accepting harm to cause outcome-maximizing harm than in the physician condition.

Method

Participants and design. Again, to increase confidence in the findings of this conceptual replication, we roughly doubled sample size to 200 American Mechanical Turk participants (aiming for ~100 per cell), who received $0.25 (128 males, 72 females, M_{age} = 33.23, *SD* = 10.89). ($N_{military}$ = 100; $N_{physician}$ = 100). This time, we predicted a main effect rather than interaction; Gpower indicates this design again provided ~99% power to detect the predicted main effect. We randomly assigned participants to either the military or physician emphasis conditions. No data were excluded.

Procedure. We asked participants to imagine they were interviewing for a job they really wanted, and gave them one of two job descriptions adapted from past work on masculine and feminine job descriptions (Rudman & Glick, 1999). In the military condition participants read: *As a **military** physician you will be responsible for the health and well-being of personnel in your military unit. On the battlefield, soldiers are in harm's way. There will be casualties. The ideal **military** doctor is technically **skilled**, **ambitious, strongly independent**, and able to perform well under pressure.* In the physician condition participants read: *As a military **physician** you will be responsible for the health and well-being of personnel in your military unit. On the battlefield, soldiers are in harm's way. There will be casualties. The ideal military **doctor** is technically skilled and able to work well under pressure, but also **helpful**, **sensitive** to the needs of each individual patient, and able to **listen carefully** to their patients' concerns.*

Next, participants imagined they must answer a moral dilemma as part of the interview

process. We presented a version of the transplant dilemma where a surgeon could allow one ill patient to die to use their organs to save five other patients (Greene at el., 2001). Participants reported their perception of interviewer expectations on two scales from 1 (*not at all*) to 7 (*very much*): *How much does the interviewer want you to say **YES (NO)**; that withholding the medical care from Patient 6 in order to save the other five patients is **(NOT)** appropriate?*

We measured expectations twice using different framings in case participants viewed these as independent questions. However, they strongly negatively correlated (r = -.79), so we reverse-coded the second question and combined them into a single measure reflecting increased acceptance of outcome-maximizing harm. Finally, participants indicated their actual judgment on the same scale as Study 4.

Results

We conducted a 2 (condition: physician vs. military emphasis) × 2 (decision: expectation vs. answer) repeated measures ANOVA (see Figure 3) with the first factor between-subjects and second factor within-subjects. This analysis yielded the expected main effect of condition: participants gave higher harm acceptance ratings (upholding utilitarianism/rejecting deontology) in the military (M = 4.10, SD = 2.33) than physician emphasis condition (M = 3.09, SD = 1.80), $F(1, 198) = 11.63$, $p = .001$, $\eta_p^2 = .06$. We also found an unexpected main effect for decision: overall, participants reported lower harm acceptance ratings for expectations (M = 3.41, SD = 2.03) than their real answers (M = 3.80, SD = 2.62), $F(1, 198) = 8.13$, $p = .005$, $\eta_p^2 = .04$. The emphasis condition × decision type interaction was not significant, $F(1, 198) = 1.07$, $p = .301$, $\eta_p^2 = .01$. Post-hoc tests confirmed that participants thought the interviewers expected them to accept harm more in the military (M = 3.84, SD = 2.19) than physician emphasis condition (M = 3.00, SD = 1.76), $F(1, 198) = 9.49$, $p = .002$, $\eta_p^2 = .05$; likewise, participants were more likely to

actually accept harm in the military (M = 4.36, SD = 2.77) than physician emphasis condition (M = 3.21, SD = 2.34), $F(1, 198) = 10.01$, $p = .002$, $\eta_p^2 = .05$.

Discussion

When a military physician job description emphasized sensitive caring, participants expected and indicated that causing harm to maximize outcomes was less acceptable to job interviews; when a description of the same job emphasized ambitious independent skill, participants expected and indicated that causing harm to maximize outcomes was more acceptable to job interviews. These findings replicate Study 4 using a different manipulation, providing further support for the argument that participants modify dilemma judgments to present themselves favorably. However, it remains possible that features of the job description simply primed participants to consider emotion or logic more carefully when forming both expectation judgments and actual dilemma judgments. In order to demonstrate *strategic* self-presentation, it is necessary to demonstrate that only participants' public dilemma decisions—not private decisions—accord with expectations. We examined this possibility in Study 6.

Study 6

Study 6 employed a similar design to Study 5, except for two changes. This time we assessed private dilemma decisions in addition to perceived expectations and public dilemma decisions. Second, we employed yet another method of manipulating whether warmth or competence traits were situationally valued: participants imagined applying for a prestigious scholarship that emphasized either warmth or competence.[5] If people employ strategic self-presentation when forming dilemma judgments, participants who learn the scholarship foundation values warmth should both expect and publicly select harm rejection judgments more

[5] This manipulation was derived from a real life experience of the first author: she had to complete a moral dilemma while applying for a prestigious fellowship, and she was trying to guess which answer was expected.

often, whereas participants who learn the foundation values competence should both expect and publicly select harm acceptance judgments more often—however, private judgments should remain unaffected. Conversely, if the manipulation simply primed participants to differentially consider emotion or logic when forming their answer, then private judgments should evince the same pattern as expectations and public judgments.

Method

Participants. We obtained 200 American (117 males, 83 females, M_{age} = 32.42, *SD* = 11.30) participants via Mechanical Turk, who received $0.25. Again, this design provided ~99% power to detect the obtained interaction. Participants were randomly assigned to one of two conditions: warmth or competence scholarship emphasis. ($N_{warmth_emphasis}$ = 100; $N_{competence\_emphasis}$ = 100).

Procedure. Participants imagined they were interviewing for a prestigious scholarship they *really wanted.* They read: *You are interviewing with the National Merit Foundation for a prestigious fellowship. It is extremely important for you to get the fellowship and you have been training a long time to get it. During the interview, you remember what kind of person they are looking for*. The competence emphasis condition continued, *"The ideal scholar of our foundation is* ***skilled****,* ***ambitious****, and a good* ***thinker****,"* whereas the warmth emphasis condition continued: *"The ideal scholar of this foundation is* ***good-natured****,* ***helpful****, and a good* ***listener****"* (bold in original).

Participants then imagined they had answered numerous questions about their background and the interview was going well—but one final question remained that would impact whether or not they would obtain the scholarship: the vaccine dilemma employed previously. Participants read that dilemma, then indicated a) which answer they thought the

interviewers *expected*, b) which answer they would *privately make*, and c) which answer they would *publicly make* in front of the interviewers. Participants indicated each answer on 7-point scales from *harm is not appropriate* (1) to *harm is appropriate* (7).

Results

We conducted a 2 (emphasis: warmth vs. competence) × 3 (decision type: expectation vs. private judgment vs. public judgment) repeated measures ANOVA with the first factor between and the second factor within participants (see Figure 4). This analysis yielded a main effect of condition: participants gave lower overall harm acceptability ratings in the warmth (M = 5.40, SD = 1.91) than competence emphasis condition (M = 4.75, SD = 1.92), F(2, 197) = 3.21, p = .042, η_p^2 = .03. There was no main effect for decision type, F(2, 197) = 2.64, p = .072, η_p^2 = .013. However, results were qualified by the expected interaction, F(2, 197) = 6.65, p < .001, η_p^2 = .05.

Post-hoc comparisons indicated that both expectations and public judgments replicated Study 5: participants in the warmth emphasis condition reported that interviewers expected less harm acceptance (M = 4.60, SD = 2.31), than participants in the competence emphasis condition (M = 5.64, SD = 2.24), F(1, 198) = 10.29, p =.002, η_p^2 = .05. Likewise, participants in the warmth emphasis condition were less likely to publicly indicate acceptance of outcome-maximizing harm (M = 4.37, SD = 2.26), than participants in the competence emphasis condition (M = 5.40, SD = 2.36), F(1, 198) = 10.10, p = .002, η_p^2 = .05. However, private judgments remained unaffected by the manipulation: participants in the warmth emphasis condition (M = 5.28, SD = 2.25) were not significantly different from participants in the competence emphasis condition regarding harm acceptability (M = 5.16, SD = 2.28), F(1, 198) = .13, p = .719, η_p^2 = .00.

Discussion

Study 6 replicated and extends the findings of Studies 4 and 5, providing increased support for our argument that participants strategically adjust dilemmas judgments in order to present situationally favorable impressions. Using a different manipulation, we again found that participants both expected and publicly made fewer harm-acceptance judgments when the situation emphasized warmth than competence. However, private judgments remained unaffected by the manipulation. These findings rule out the possibility that the differences in expectation and public judgment reflect priming, as private judgments remained unaffected by the manipulation. Instead, these findings suggest that participants employed strategic self-presentation to publicly provide dilemma answers that accorded with expectations rather than private considerations.

Study 7

Together with past findings (Rom et al., 2016), the current work suggests that dilemma answers entail a warmth/competence trade-off: rejecting harm makes one appear warm but less competent, whereas accepting outcome-maximizing harm makes one appear cold but more competent. Although people appear to strategically modify their dilemma judgments to emphasize either warmth or competence, doing so entails the trade-off of appearing weaker on the converse trait. Yet, on many occasions it may be optimal to present oneself as high on both traits. For example, politicians may wish to appear both warm and competent to increase chances of re-election, yet face dilemmas that pit individual well-being against public interest, such as authorizing forceful interrogations to obtain life-saving information, or directing medical funding away from rare but deadly disorders towards widespread problems. Is it possible to frame one's dilemma decision so as to reduce the warmth/competence trade-off previously obtained? Everett and colleagues (2016) provided initial evidence consistent with this possibility: when an injured

soldier begged to death to avoid capture and torture by the enemy, and decision-makers rejected this request, perceivers viewed them as more moral and trustworthy when they offered categorical deontological justifications (i.e., "killing is wrong even if it has good consequences") compared to utilitarian or contractual reasons. These arguments do not speak directly to perceptions of warmth or competence, but they suggest that perceivers draw inferences from the justifications decision-makers provide, beyond the decisions they make.

We hypothesized that decision-makers can augment perception of their weaker trait by supplementing their dilemma decision itself with justifications that appeal to emotions or to logic. If such appeals impact dilemma perceptions, then people who accept causing outcome-maximizing harm may appear less cold by expressing emotional concern for the victim of harm. Conversely, people who reject harm may appear less incompetent by expressing consideration of logical reasoning. To examine this possibility, we assessed warmth and competence perceptions of dilemma decision-makers who either accepted or rejected harm, and who framed their decision either in terms of emotion or cognition.

Method

Participants and design. We obtained 401 American (251 males, 150 females, M_{age} = 32.42, *SD* = 11.30) participants via Mechanical Turk, who received $0.30. Participants were randomly assigned to learn that a previous participant either accepted or rejected harm for either emotional or logical reasons, and rated them on warmth and competence, for a 2 (explanation: emotional vs. logical) × 2 (target decision: harm inappropriate vs. appropriate) × 2 (personality dimension: warmth vs. competence) design, where the first two factors varied between-subjects and the final factor varied within-subjects. Despite the large sample, this study had only ~82% power to detect the three-way interaction. ($N_{emotional}$ = 100; $N_{logical}$ = 101).

Procedure. The procedure was similar to the other-perception condition in Study 1: Participants viewed a photo of a university student named Brad who ostensibly previously participated. They read the crying baby dilemma and learned that Brad ostensibly either accepted or rejected the specified harm, accompanied by a brief written explanation emphasizing either emotional or logical justifications for this decision.

Specifically, in the emotional harm rejection condition, participants read, "*No, it is completely unacceptable to kill the baby! It doesn't matter what the reasons are; I just could not live with myself if I hurt an innocent little baby. Killing is forbidden for any reason and never justified.*" In the in logical harm rejection condition participants read, "*No, it is unacceptable to kill the baby! I understand that doing so makes logical sense, but killing some people to protect others creates an immoral society. It is better to live in a society that forbids killing for any reason than one where killing some people is justified to help others*." In the emotional harm acceptance condition participants read, "*Yes, it is acceptable to kill the baby. It is true that it would break my heart to kill an innocent baby, but it just makes sense to perform the action that saves everybody. It upsets me very much, but it's the only logical thing to do.*" Finally, in the rational harm acceptance condition participants read, "*Yes, it is completely acceptable to kill the baby! It just makes sense to perform the action that saves everybody. It's the only logical thing to do.*" After reading Brad's dilemma decision and justification, participants rated Brad's warmth ($\alpha = .87$) and competence ($\alpha = .82$) as before.

Results

Target warmth and competence. We submitted ratings to a 2 (justification type: emotional vs. logical) × 2 (target decision: harm inappropriate vs. appropriate) × 2 (personality dimension: warmth vs. competence) repeated-measures ANOVA with the first two factors

between-subjects and the last factor within-subjects (see Figure 5). There was no main effect for justification type, $F(1, 397) = 0.00$, $p = .990$, $\eta_p^2 = .00$, or personality dimension, $F(1, 397) = .86$, $p = .355$, $\eta_p^2 = .00$. However, there was a main effect of target decision: participants gave higher ratings overall when Brad rejected ($M = 5.10$, $SD = 1.10$) versus accepted causing harm ($M = 4.70$, $SD = .99$), $F(1, 397) = 15.13$, $p < .001$, $\eta_p^2 = .04$. These results were qualified by significant two-way interactions between justification type and personality dimension, $F(1, 397) = 31.91$, $p < .001$, $\eta_p^2 = .07$, and between target decision and personality dimension, $F(1, 397) = 260.04$, $p < .001$, $\eta_p^2 = .40$. The three-way interaction did not approach conventional levels of significance, $F(1, 397) = 2.27$, $p = .136$, $\eta_p^2 = .01$, so we examined the two-way interactions.

The first interaction indicated that justifications impacted warmth and competence decisions: perceivers rated Brad as higher on warmth ($M = 5.05$, $SD = 1.30$) than competence ($M = 4.74$, $SD = 1.21$), when he provided emotional justifications, $F(1, 399) = 12.83$, $p < .001$, $\eta_p^2 = .03$, whereas they rated him higher on competence ($M = 4.78$, $SD = 1.34$) than warmth ($M = 5.01$, $SD = 1.06$), when he provided rational justifications, $F(1, 399) = 7.25$, $p = .007$, $\eta_p^2 = .02$.

The second interaction replicated previous findings by showing that dilemma decisions impacted warmth and competence ratings: perceivers rated Brad as higher on warmth ($M = 5.51$, $SD = 1.15$) than competence ($M = 4.69$, $SD = 1.27$), when he rejected outcome-maximizing harm (upholding deontology), $F(1, 399) = 96.16$, $p < .001$, $\eta_p^2 = .19$, whereas they rated him higher on competence ($M = 5.06$, $SD = .98$) than warmth ($M = 4.34$, $SD = 1.24$), when he accepted outcome-maximizing harm (upholding utilitarianism), $F(1, 399) = 10.47$, $p = .001$, $\eta_p^2 = .03$.

Discussion

As predicted, results indicated that decision-makers can bolster their weaker trait implied by their judgment with argumentation that indicates trait-relevant processing. We replicated the

warmth/competence trade-off demonstrated in previous work (Rom et al., 2016), but found that justifications referring to emotional struggles or logical considerations independently increased perceptions of warmth and competence, respectively. Hence, decision-makers who accept harm can offset concerns about their warmth by framing their decision in emotional terms, whereas those who reject harm can offset concerns about their competence by framing their decision in logical terms.

General Discussion

Across seven studies, we garnered evidence that people hold accurate meta-perceptions regarding whether their dilemma decisions convey warmth or competence, and strategically adjust dilemma judgments to present themselves favorably. Study 1 replicated past work (Rom et al., 2016) by demonstrating participants rated decision-makers who rejected harm (upholding deontology) as warmer but less competent than decision-makers who accepted outcome-maximizing harm (upholding utilitarianism), together with the novel finding that participants anticipated that others would rate them according to the same warmth/competence tradeoff following the same respective decisions—even though, privately, participants rated themselves as high on both warmth and competence regardless of their decision. Moreover, participants anticipated that others' warmth and competence ratings would reflect whichever judgment those others learned participants made, even when this belief was erroneous (Study 2). Importantly, meta-perceptions of this warmth/competence trade-off generalized to a battery of various dilemma stimuli and a different sample (Study 3). Thus, it seems clear that people hold accurate meta-perceptions regarding how others perceive them based on their dilemma judgments, that these meta-perceptions differ from self-perceptions, track information available to others, and do not merely reflect individual differences in which judgments people prefer.

Next, we examined whether people use meta-perception information to strategically adjust their judgments. First, we demonstrated that dilemma decisions are sensitive to context: When we framed the Study 4 as focusing differences in emotional competency, participants were more likely to reject causing harm (upholding deontology), thereby emphasizing their warmth and emotional processing, compared to when the study was framed as examining differences in logical reasoning, when participants were more likely to accept harm (upholding utilitarianism), thereby emphasizing their competence and logical skills. Study 5 replicated this finding using a different manipulation, where participants simulated interviewing for a job as a military physician, where the description emphasized either military competency or physician care. Participants were more likely to reject harm in the care than competency condition. Study 6 replicated both of these effects using yet another manipulation—a scholarship application that emphasized either academic competency or interpersonal skills. Importantly, this manipulation influenced both expectations and public judgments—but failed to impact private judgments, suggesting that participants were *strategically* adjusting dilemma judgments rather than merely responding to primes in the stimulus materials.

Finally, Study 7 demonstrated that decision-makers can use communication strategies to augment the weaker trait implied by the warmth/competence meta-perception trade-off when making dilemma decisions. Specifically, decision-makers can provide either emotional or logical justifications for either dilemma judgment, and these justifications independently impact perceptions of warmth and competence. Hence, decision-makers who accept harm (upholding deontology) can offset perceptions of incompetence by describing logical reasons for their decision, whereas decision-makers who accept outcome-maximizing harm (upholding utilitarianism) can offset perceptions of coldness by describing emotional experiences.

In each case, the impacts of dilemma decisions on warmth perceptions was mirrored by similar patterns on ratings of decision-maker *morality*. Indeed, warmth and morality correlated highly in all studies. Such findings could be taken as evidence that warmth and morality reflect a single core construct, but recent work suggests that lay people draw important distinctions between warmth/sociability (i.e., interpersonal friendliness) and morality (e.g., trustworthiness—see Brambilla et al., 2011; Goodwin et al., 2014). Such findings could also reflect the possibility that the current measure of warmth actually reflects moral character evaluations instead of genuine perceptions of warmth/sociability, by including items such as *sincere*. However, as noted above, re-analyses employing revised composites excluding such terms, or indeed using only the single item *warm* demonstrate the same pattern as the warmth composite using all warmth items. Therefore, we suggest that perceivers draw inferences of both warmth and morality from others' deontological dilemma judgments, and these inferences happen to covary substantially in the current paradigm. It may be that these inferences stem from different aspects of deontological judgments—perhaps warmth perceptions reflect inferences of emotional processing, whereas morality inferences reflect perceptions of rule-following—which covary in the current paradigm. Consistent with this possibility, Rom and colleagues (2016) found that perceptions of emotional processing mediated the effect of dilemma judgment on perceptions of warmth—but not on perceptions of morality. Future work might profit from disentangling which aspects of deontological decision-making imply warmth and which imply morality.

Implications for Models of Moral Judgment

The dual-process model of moral judgment (Greene et al., 2001) and other popular models (Cushman, 2013; Crockett, 2013; Sunstein, 2005; Mikhail, 2007) describe the impact of basic psychological processes on moral dilemma judgments, such as affective reactions to harm,

cognitive evaluations of outcomes, or heuristic application of moral rules. Importantly, all of these processes should apply similarly whether participants respond to moral dilemmas alone on a desert island or during a live television broadcast watched by millions. We do not dispute the importance of basic processes for influencing dilemma judgments, but we suggest that existing theories are incomplete if they treat public versus private circumstances as identical. We suggest that answering dilemmas while on television—or in any social situation—evokes concern over others' perceptions of ones' dilemma judgment, and how that judgment reflects on oneself. People appear aware of the warmth/competence trade-off others infer from their decision, and strategically modify judgments to present themselves favorably. Hence, higher-order social processes causally contribute to dilemma responses, in addition to basic processes.

The finding that strategic self-presentation drives variance in dilemma judgments suggests that researchers should revisit earlier findings to consider whether self-presentation may account for some of the variance ascribed to basic processes. For example, various researchers have documented gender differences in dilemma judgments (e.g., Fumagalli et al., 2010; Arutyunova, Alexandrov, & Hauser, 2016), such that women evince stronger inclinations to reject harm than men, but similar inclinations to maximize outcomes, leading to higher reports of conflict (Friesdorf, Conway, & Gawronski, 2015). Typically, researchers explain such gender differences in terms of biologically-based constructs such as empathy (Eisenberg & Lennon, 1983) and testosterone (Carney & Mason, 2010), or differences in socialization practices (Eagly & Wood, 1999). However, the current findings raise an alternative possibility: it may be that women experience stronger social expectations to avoid causing harm than do men, even as they appreciate the logic of doing so. After all, women often face pressure to appear both warm and competent, whereas often competence alone often meets male role expectations (Rudman &

Glick, 1999). Moreover, women often feel more obliged to engage in self-presentation than do men (Deaux & Major, 1987). Such expectations could lead women to reject harm (upholding deontology) more frequently, despite experiencing similar basic processing as do men.

In other work, Lucas & Livingstone (2014) found that participants who socially connected with others made more utilitarian judgments, presumably because social connection reduced aversive affect associated with deontological judgments. However, our results suggest an alternative process: participants who had already connected with others may have felt they established sufficient evidence of warmth or morality that they could afford to display other qualities, such as competence. Indeed, such alternative explanations may occur even in studies where there is no direct social contact between participants and others (e.g., online studies). From a Griceian (1989) perspective, every research study is effectively an act of social communication between the participant and the experimenter. Cues in the framing, instructions, or manipulations of any dilemma study may hint at whether warmth or competence is contextually prioritized, leading participants to infer that one or another dilemma answer is preferred.

Indeed, self-presentation of this sort may even account for some of the response variance between the original trolley dilemma, where approximately 80% of people accept causing harm to save lives, and the footbridge dilemma, where about 80% of people reject causing harm (Greene et al., 2001). In the footbridge dilemma, harm acceptance means being willing to push and thereby kill with one's own hands, whereas in the trolley dilemma harm, acceptance means simply pressing a button. Research suggests that employing the personal force of one's physical being to kill another is more aversive than employing a mechanical mediator (Greene et al., 2009). Accordingly, lay perceivers may view harm caused through personal force as more likely

evidence of cold-heartedness than harm caused through intermediaries—thereby creating greater social pressure to avoid causing harm on the footbridge than trolley dilemma. Consistent with this possibility, Everett and colleagues (2016) found that perceivers drew important distinctions between the trustworthiness of decision-makers who accepted vs. rejected harm on the footbridge dilemma, but less of a distinction between those who accepted vs. rejected harm on the trolley dilemma. Future work should directly examine social expectations of appropriate answers in such cases.

Limitations

This research shares limitations with nearly all dilemma research: of necessity, participants make decisions about hypothetical scenarios rather than actual situations. Hence, it remains possible that perceptions and meta-perceptions of real-life dilemma decisions (such as Turing's decision from the beginning of the paper) evince different or even stronger effects. In addition, like most dilemma research, the dilemmas employed here vary on a number of factors that may influence judgments, such as whether the victim of harm is guilty of causing danger or not, or is fated to die or not (Christensen, Flexas, Calabrese, Gut, & Gomila, 2014). Future work should systematically vary each of these factors to determine whether they impact perceptions and meta-perceptions of dilemma judgments. Moreover, the dilemmas employed here examine only violations of moral proscriptions—causing harm to maximize outcomes—whereas it is possible to conceptualize dilemmas involving prescription violations—saving one person at a risk to many—that may entail different perceptions and meta-perceptions (Gawronski et al., 2015). Future work may profit by comparing such dilemmas.

In addition, although the current work employed participants from different countries in several languages, all participants hailed from broader 'Western' culture. Recent work has

documented that East Asian participants are less likely to endorse outcome-maximizing harm than Western participants (e.g., Gold, Colman, & Pulford, 2014). One reason for this difference may be increased fatalism in Asian culture—the belief that one should not interfere with destiny (Chih-Long, 2013). It remains unclear whether perceptions of dilemma judgments also reflect such cultural variation—the cultural background of both perceivers and decision-makers may matter. Future research might profitably investigate these possibilities.

Conclusion

Building on work examining the role of basic psychological processes in driving dilemma judgments, the current work provides evidence that higher-order social processes also play a role. Participants demonstrated accurate meta-insight into how warm and competent their dilemma judgments would make them appear to others, and strategically shifted public (but not private) dilemma judgments to accord with such expectations depending on whether situations prioritized warmth or competence. These findings suggest that classic models of dilemma decision-making (e.g., Greene et al., 2001) underestimate the influence of social considerations. When Allied forces allowed the Axis raid on Coventry to proceed so as to protect the Enigma Code decryption, they likely engaged in not only basic emotional and logical processing, but also considered how their allies would have reacted to this decision. In the midst of a desperate war, they selected a decision that made them appear competent at the cost of warmth—had circumstances been different, perhaps they would have selected an entirely different answer.

References

Amit, E., & Greene, J. D. (2012). You see, the ends don't justify the means: Visual imagery and moral judgment. *Psychological Science, 23*, 861-868. doi:10.1177/0956797611434965

Anderson, C., Ames, D. R., & Gosling, S. D. (2008). Punishing hubris: The perils of overestimating one's status in a group. *Personality and Social Psychology Bulletin, 34*, 90-101. doi: 10.1177/0146167207307489

Andersen, S. M., & Ross, L. (1984). Self-knowledge and social inference: I. The impact of cognitive/ affective and behavioral data. *Journal of Personality and Social Psychology, 46*, 280-293.

Asch, S. E. (1948). The doctrine of suggestion, prestige and imitation in social psychology. *Psychological Review*, 55, 250–276.

Aquino, K., & Reed, A. II. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology, 83*, 1423-1440. doi:10.1037//0022-3514.83.6.1423

Arutyunova, K. R., Alexandrov, Y. I., & Hauser, M. D. (2016). Sociocultural Influences on Moral Judgments: East–West, Male–Female, and Young–Old. *Frontiers in Psychology.*

Barish, K., (Producer), & Pakula, A. J. (Director). (1982). *Sophie's Choice* [Motion picture]. United States: Incorporated Television Company.

Bartels, D. (2008). Principled moral sentiment and the flexibility of moral judgment and decision making. *Cognition, 108*, 381–417. doi:10.1016/j.cognition.2008.03.001

Bartels, D. M., & Pizarro, D. A. (2011). The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition, 121*, 154-161. doi:10.1016/j.cognition.2011.05.010

Brambilla, M., Rusconi, P., Sacchi, S., & Cherubini, P. (2011). Looking for honesty: The primary role of morality (vs. sociability and competence) in information gathering. *European Journal of Social Psychology, 41,* 135-143. Brambilla, M., Rusconi, P., Sacchi, S., & Cherubini, P. (2011). Looking for honesty: The primary role of morality (vs. sociability and competence) in information gathering. *European Journal of Social Psychology, 41,* 135-143.

Bloom, P. (2011). Family, community, trolley problems, and the crisis in moral psychology. *The Yale Review*, *99*(2), 26-43.

Carlson, E. N., & Furr, R. M. (2009). Evidence of differential meta-accuracy: People understand the different impressions they make. *Psychological Science, 20,* 1033-1039. doi: 10.1111/j.1467-9280.2009.02409.x

Carlson, E. N., Vazire, S., & Furr, R. M. (2011). Meta-insight: Do people really know how others see them? *Journal of Personality and Social Psychology, 101*, 831-846. doi: 10.1037/a0024297

Carney, D. R., & Mason, M. F. (2010). Decision making and testosterone: When the ends justify the means. *Journal of Experimental Social Psychology*, *46*(4), 668–671. http://doi.org/10.1016/j.jesp.2010.02.003

Chambers, J. R., Epley, N., Savitsky, K., & Windschitl**,** P. D. (2008). Knowing too much: Using private knowledge to predict how one is viewed by others. *Psychological Science, 19,* 542-548. doi: 10.1111/j.1467-9280.2008.02121.x

Chih-Long, Y. (2013). It is our destiny to die: The effects of mortality salience and culture-priming on fatalism and karma belief. *International Journal of Psychology*, *48,* 818-828.

doi: 10.1080/00207594.2012.678363

Christensen, J. F., Flexas, A., Calabrese, M., Gut, N. K., & Gomila, A. (2014). Moral judgment reloaded: a moral dilemma validation study. *Frontiers in Psychology, 5*, 1-18. doi:10.3389/fpsyg.2014.00607

Conway, P., & Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision-making: A process dissociation approach. *Journal of Personality and Social Psychology*, *104*, 216-235. doi:10.1037/a0031021

Copeland, B. J. (2014). *Turing: pioneer of the information age*. Oxford University Press.

Crockett, M. J. (2013). Models of morality. Trends in cognitive sciences, 17, 363-366.

Cushman, F. (2013). Action, outcome, and value a dual-system framework for morality. *Personality and social psychology review, 17*, 273-292.

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*, 1082–1089. doi: 10.1111/j.1467-9280.2006.01834.x

Deaux, K., & Major, B. (1987). Putting gender into context: An interactive model of gender-related behavior. *Psychological review*, *94*, 369.

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological review*, *109*(3), 573.

Eagly, A. H., & Wood, W. (1999). The origins of sex differences in human behavior: Evolved dispositions versus social roles. *American Psychologist*. doi:10.1037//0003-066x.54.6.408

Eisenberg, N., & Lennon, R. (1983). Sex differences in empathy and related capacities. *Psychological Bulletin, 94*, 100–131.

Epley, N., & Dunning, D. (2000). Feeling “Holier than thou”: Are self-serving assessments produced by errors in self or social prediction? *Journal of Personality and Social Psychology, 79,* 861-875. doi: 10.1037/0022-3514.79.6.861

Epley, N., Keysar, B., Van Boven, L., & Gilovich**,** T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology, 87,* 327-339. doi: 10.1037/0022-3514.87.3.327

Everett, J. A., Pizarro, D. A., & Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology: General, 145*, 772.

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2006). Universal dimensions of social cognition: Warmth and Competence. *Trends in Cognitive Sciences, 11*, 77-83. doi:10.1016/j.tics.2006.11.005

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology, 82*, 878-902. doi: 10.1037/0022-3514.82.6.878

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review, 5,* 5-15. doi: 10.1093/0199252866.003.0002

Friesdorf, R., Conway, P., & Gawronski, B. (2015). Gender Differences in Responses to Moral Dilemmas A Process Dissociation Analysis. *Personality and Social Psychology Bulletin*, 0146167215575731.

Fumagalli, M., Ferrucci, R., Mameli, F., Marceglia, S., Mrakic-Sposta, S., Zago, S., … Priori, A. (2010). Gender-related differences in moral judgments. *Cognitive Processing, 11*, 219–

226. doi:10.1007/s10339-009-0335-2

Faul, F., Erdfelder, E., Lang, A.-G., &; Buchner, A. (2007). GPower 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191. doi:10.3758/BF03193146

Gawronski, B., Conway, P., Armstrong, J., Friesdorf, R., & Hütter, M. (2015). Moral dilemma judgments: Disentangling deontological inclinations, utilitarian inclinations, and general action tendencies. In J. P. Forgas, P. A. M. Van Lange, & L. Jussim (Eds.), *Social psychology of morality*. New York: Psychology Press.

Gold, N., Colman, A. M., & Pulford, B. D. (2014). Cultural differences in responses to real-life and hypothetical trolley problems. *Judgment and Decision Making, 9*, 65-76.

Gold, N., Pulford, B. D., & Colman, A. M. (2015). Do as I say, don't do as I do: Differences in moral judgments do not translate into differences in decisions in real-life trolley problems. *Journal of economic psychology*, *47*, 50-61.

Gleichgerrcht, E., & Young, L. (2013). Low levels of empathic concern predict utilitarian moral judgment. *PLOS ONE*, *8,* 1-9. doi:10.1371/journal.pone.0060418

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology, 106*, 148.

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*, 364-371. doi:10.1016/j.cognition.2009.02.001

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44*, 389-400. doi: 10.1016/j.neuron.2004.09.027

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*, 2105-2108. doi: 10.1126/science.1062872

Grice, H. P. (1989). Studies in the Way of Words. Cambridge, MA: Harvard University Press.

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

Hofmann, W., Wisneski, D. C., Brandt, M. J., & Skitka, L. J. (2014). Morality in everyday life. *Science, 345*(6202), 1340-1343.

Kahane, G. (2015). Sidetracked by trolleys: Why sacrificial moral dilemmas tell us little (or nothing) about utilitarian judgment. *Social neuroscience, 10*(5), 551-560.

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. doi:10.1016/j.cognition.2014.10.005

Kant, I. (1785/1959). *Foundation of the metaphysics of morals* (L. W. Beck, Trans.). Indianapolis: Bobbs-Merrill.

Kaplan, S. A., Santuzzi, A. M., & Ruscher, J. B. (2009). Elaborative metaperceptions in outcome-dependent situations: The diluted relationship between default self-perceptions and metaperceptions. *Social Cognition, 27*, 601-614. doi: 10.1521/soco.2009.27.4.601

Kenny, D. A., & DePaulo, B. M. (1993). Do people know how others view them? An empirical and theoretical account. *Psychological Bulletin, 114*, 145-161. doi: 10.1037/0033-2909.114.1.145

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*, 908-911. doi:10.1038/nature05631

Kohlberg, L. (1969). Stage and sequence: The cognitive–developmental approach to socialization. In D. A. Goslin (Ed.), *Handbook of socialization theory and research*. (pp. 347–480). Chicago, IL: Rand McNally.

Krebs, D. L. (2011). The evolution of a sense of morality. *Creating consilience*, 299-317.

Kreps, T. A., & Monin, B. (2014). Core values versus common sense consequentialist views appear less rooted in morality. *Personality and Social Psychology Bulletin*, 0146167214551154.

Kruger, J., & Gilovich, T. (2004). Actions, intentions, and self-assessment: The road to self-enhancement is paved with good intentions. *Personality and Social Psychology Bulletin*, *30*(3), 328-339.

Kundu, P., & Cummins, D. D. (2012). Morality and conformity: The Asch paradigm applied to moral decisions. *Social Influence*, *8*, 268-279.

Leary, M. R. (1989). Self-presentational processes in leadership emergence and effectiveness.

Leary, M. R. (1995). *Self-presentation: Impression management and interpersonal behavior*. Brown & Benchmark Publishers.

Leary, M. R., & Kowalski, R. M. (1990). Impression management: A literature review and two-component model. *Psychological bulletin*, *107*(1), 34.

Liu, B. S., & Ditto, P. H. (2013). What dilemma? Moral evaluation shapes factual belief. *Social Psychological and Personality Science, 4*, 316–323. http://dx.doi.org/10.1177/1948550612456045.

Lucas, B. J., & Galinsky, A. D. (2015). Is utilitarianism risky? How the same antecedents and mechanism produce both utilitarian and risky choices. *Perspectives on Psychological Science, 10*(4), 541-548.

Lucas, J. L., & Livingstone, R. W. (2014). Feeling socially connected increases utilitarian choices in moral dilemmas. *Journal of Experimental Social Psychology, 53*, 1–4. doi: 10.1016/j.jesp.2014.01.011

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in cognitive sciences, 11*(4), 143-152.

Milgram, S. (1963). Behavioral Study of obedience. *The Journal of abnormal and social psychology*, *67*(4), 371.

Mikhail, J. (2007). Universal moral grammar: theory, evidence and the future. *TRENDS in Cognitive Sciences, 11*, 143-152. doi:10.1016/j.tics.2006.12.007

Mill, J. S. (1861/1998). *Utilitarianism*. In R. Crisp (Ed.), New York: Oxford University Press.

Moore, A. B., Clark, B. A., & Kane, M. J. (2008). Who shalt not kill? Individual differences in working memory capacity, executive control, and moral judgment. *Psychological Science, 19*, 549-57. doi:10.1111/j.1467-9280.2008.02122.x

Peeters, G. (1983). Relational and informational pattern in social cognition. In W. Doise & S. Moscovici (Eds.), *Current issues in European social psychology* (pp. 201-237). Cambridge, England: Cambridge University Press.”.

Pronin, E. (2008). How we see ourselves and how we see others. *Science, 320*, 1177-1180.

doi: 10.1126/science.1154199

Reis, H. T., & Gruzen, J. (1976). On mediating equity, equality, and self-interest: The role of self-presentation in social exchange. *Journal of Experimental Social Psychology*, *12*(5), 487-503.

Rom, S. C., Weiss, A., & Conway, P. (2017). Judging those who judge: Perceivers infer the roles of affect and cognition underpinning others' moral dilemma responses. *Journal of Experimental Social Psychology*, *69*, 44-58.

Sarbin, T. R., & Allen, V. L. (1968). Role theory.

Sherif, M. A. (1935). A study of some social factors in perception, *Archives of Psychology*, 27, 1– 60.

Uhlmann, E. L., Pizarro, D. A., Tannenbaum, D., & Ditto, P. H. (2009). The motivated use of moral principles. *Judgment and Decision Making*, *4*(6), 479.

Valdesolo, P. & DeSteno, D. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science, 17*, 476–477. doi:10.1111/j.1467-9280.2006.01731.x

Vazire, S. (2010). Who knows what about a person? The self-other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology*, *98*, 281-300. doi: 10.1037/a0017908

Von Baeyer, C. L., Sherk, D. L., & Zanna, M. P. (1981). Impression management in the job interview: When the female applicant meets the male (chauvinist) interviewer. *Personality and Social Psychology Bulletin*, *7*(1), 45-51.

Wiggins, J. S. (1979). A psychological taxonomy of trait-descriptive terms: The interpersonal domain. *Journal of Personality and Social Psychology, 37*, 395-412. doi: 10.1037/0022-3514.37.3.395

Winkielman, P., & Schwarz, N. (2001). How pleasant was your childhood? Beliefs about memory shape inferences from experienced difficulty of recall. *Psychological Science*, *12*(2), 176-179.

Winterbotham, F. W. (1974). The Ultra Secret (New York, 1974). *Ronald Lewin, Ultra Goes to War (New York, 1978).*

Wood, W., & Eagly, A. H. (2012). Biosocial construction of sex differences and similarities in behavior. *Advances in experimental social psychology*, *46*, 55-123.

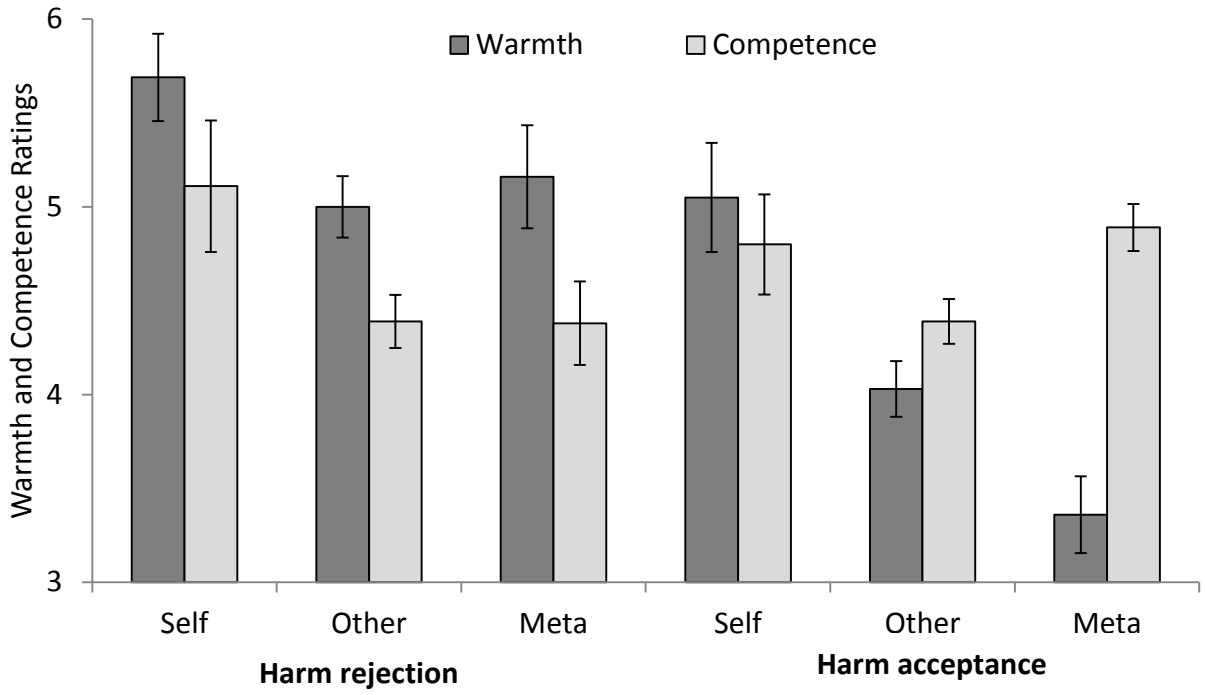

Figure 1.

Participants' self, target, and meta-perception warmth and competence ratings when they or the target rejected causing harm to maximize outcomes (upholding deontology), or accepted such harm (upholding utilitarianism), Study 1. Error bars reflect standard errors.

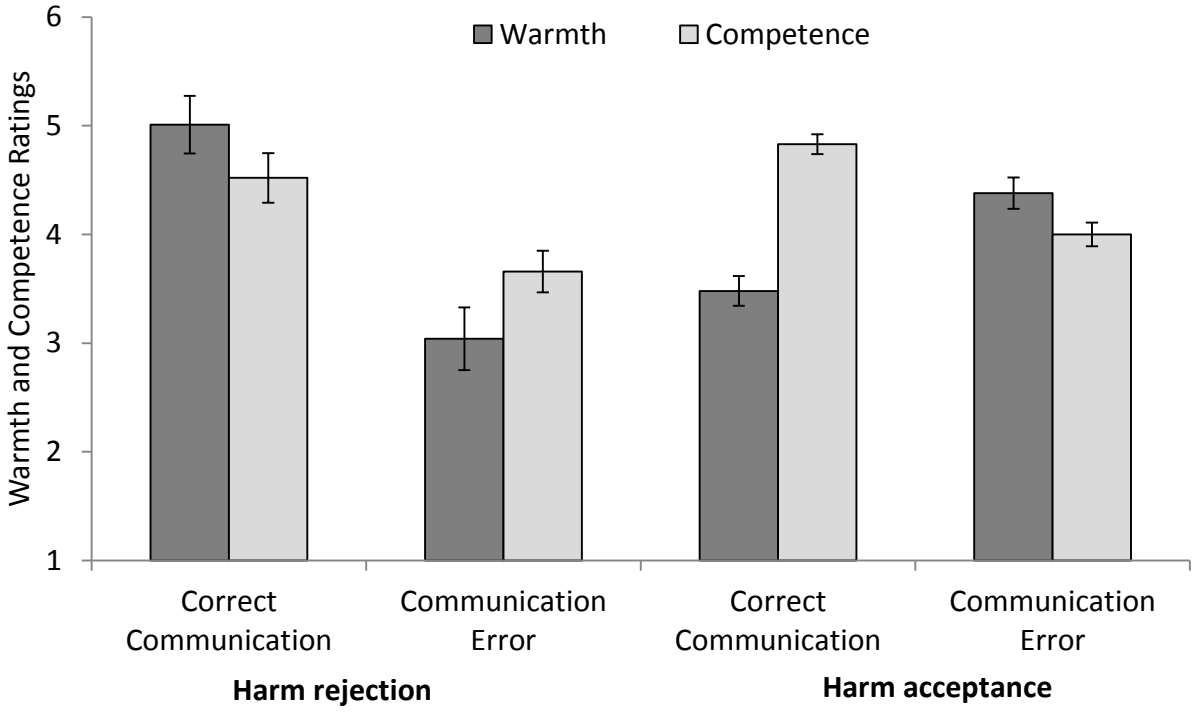

Figure 2.

Warmth and competence meta-perceptions when participants rejected outcome-maximizing harm (upholding deontology), or accepted such harm (upholding utilitarianism), and imagined others correctly learned their judgment (correct communication condition) or erroneously believed they made the opposite judgement (communication error condition), Study 2. Error bars reflect standard errors.

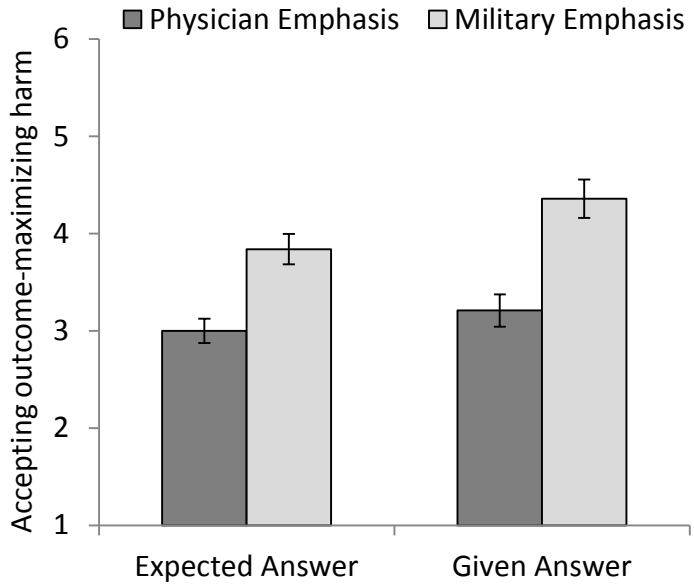

Figure 3.

Mean expected and actual dilemma decisions (accepting outcome-maximizing harm, upholding utilitarianism) during military physician job interview emphasizing either military or physician skills, Study 5. Error bars reflect standard errors.

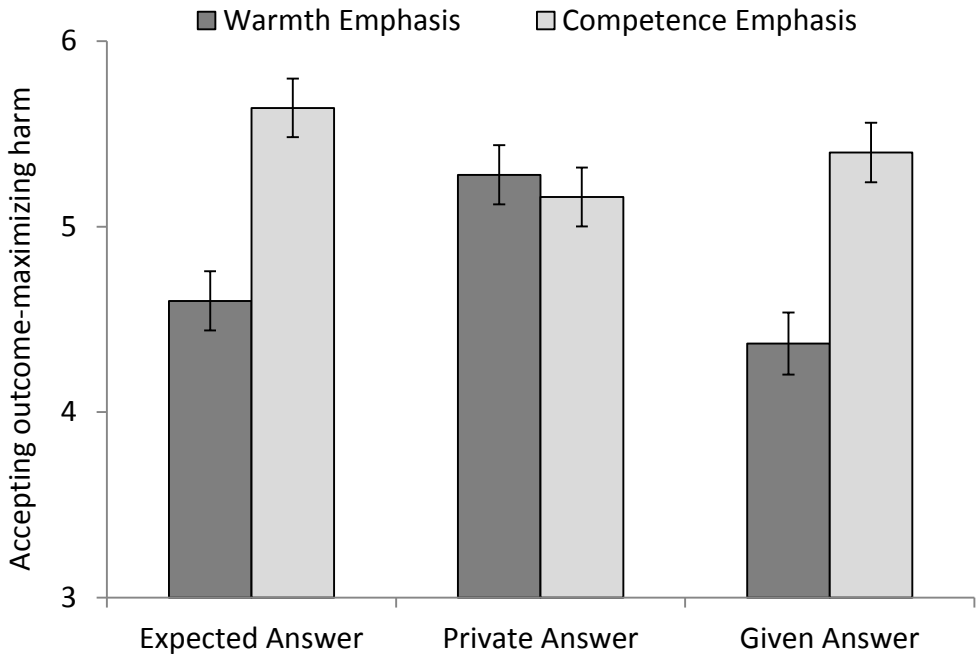

Figure 4.

Mean expected, private, and public dilemma decisions (accepting outcome-maximizing harm, upholding utilitarianism) when applying for a scholarship emphasizing either warmth or competence, Study 6. Error bars reflect standard errors.

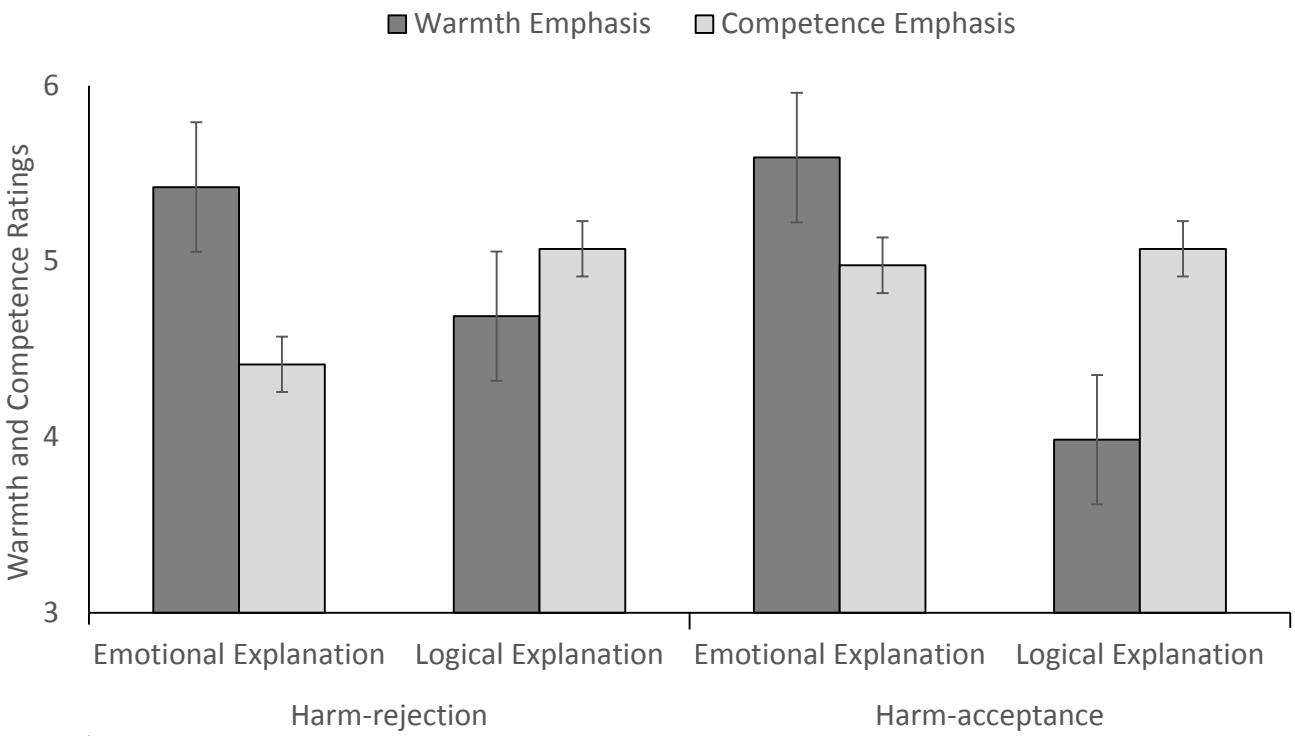

Figure 5.

Warmth and competence ratings when Brad rejected or accepted causing harm and either gave an emotional or logical explanation, Study 7. Error bars reflect standard errors.

Supplementary Material

Study 1: Alternative Contrasts

Participants rated Brad higher on warmth when he rejected (M = 5.00, SD = 1.19), than when he accepted causing outcome-maximizing harm (M = 4.03, SD = .99), $F(1, 194) = 15.57$, $p < .001$. Conversely, they rated Brad as higher in competence when he accepted (M = 5.16, SD = .16), rather than rejected causing outcome-maximizing harm (M = 3.36, SD = 1.31),-$F(1, 194) = 11.67$, $p < .001$.

When participants rejected harm they inferred others would perceive them as warmer (M = 5.16, SD = 1.59) than when they accepted causing outcome-maximizing harm (M = 3.36, SD = 1.31), $F(1, 194) = 22.95$, $p < .001$. In contrast, when they accepted such harm, they inferred that others would perceive them as (slightly) more competent (M = 4.89, SD = .80) than when they rejected such harm (M = 4.38, SD = 1.46), although results did not reach conventional levels of significance, $F(1, 194) = 2.37$, $p = .129$.

Study 2: Alternative Contrasts

Participants expected that others would rate them as warmer when they rejected (M = 5.06, SD = 1.49) rather than accepted causing harm (M = 3.40, SD = 1.68), $F(1, 182) = 19.70$, $p < .001$, $\eta_p^2 = .10$.

Regarding competence, participants expected that others would rate them as equally competent when they rejected (M = 4.51, SD = 1.19), rather than accepted causing harm (M = 4.82, SD = 1.19). Although results were trending in the expected direction, they did not reach significance, $F(1, 187) = 1.91$, $p = .168$, $\eta_p^2 = .01$.

Participants expected that others would rate them as less warm when they rejected (M = 2.94, SD = 1.93) rather than accepted causing harm (M = 4.42, SD = 1.68), $F(1, 369) = 22.28$, $p < .001$, $\eta_p^2 = .11$. Participants in the error meta-perceptions condition expected that others would rate them as equally competent when they rejected (M = 3.66, SD = 1.40) rather than accepted causing harm (M = 4.00, SD = 1.29), $F(1, 369) = 2.25$, $p = .135$, $\eta_p^2 = .01$.

Study 3: Alternative Contrasts

Participants expected that others would rate them as warmer when they rejected (M = 5.02, SD = 1.49) rather than accepted causing harm (M = 3.48, SD = 1.71), $F(1, 129) = 25.75$, $p < .001$, $\eta_p^2 = .17$. Regarding competence participants expected that others would rate them as less competent when they rejected (M = 4.51, SD = 1.19) rather than accepted causing harm (M = 4.82, SD = 1.19), $F(1, 129) = 13.98$, $p < .001$, $\eta_p^2 = .01$.