Cognition 217 (2021) 104886

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Original Articles

What we would (but shouldn't) do for those we love: Universalism versus partiality in responding to others' moral transgressions

Laura K. Soter [a,b,*], Martha K. Berg [a], Susan A. Gelman [a,1], Ethan Kross [a,1]

[a] Department of Psychology, University of Michigan, 1004 East Hall, 530 Church Street, Ann Arbor, MI, 48109, USA
[b] Department of Philosophy, University of Michigan, 2215 Angell Hall, 435 S State Street, Ann Arbor, MI 48109, USA

ARTICLE INFO

Keywords:
Close relationships
Moral psychology
Moral universalism
Moral partiality
Inconsistency

ABSTRACT

Recent work indicates that people are more likely to protect a close (vs. distant) other who commits a crime. But do people think it is *morally right* to treat close others differently? On the one hand, universalist moral principles dictate that people should be treated equally. On the other hand, close relationships are the source of special moral obligations, which may lead people to believe they ought to preferentially protect close others. Here we attempt to adjudicate between these competing considerations by examining what people think they *would* and *should* do when a close (vs. distant) other behaves immorally. Across four experiments ($N = 2002$), we show that people believe they *morally should* protect close others more than distant others. However, we also document a striking discrepancy: participants reported that they *would* protect close others far more than they *should* protect them. These findings demonstrate that people believe close relationships influence what they morally ought to do—but also that moral decisions about close others may be a context in which people are particularly likely to fail to do what they think is morally right.

> The moral standpoint is characterized by an attitude of impartiality, a refusal to see the life, projects, good, or interests of any particular person (oneself included) as having a greater or lesser value than those of another. (Archard, 1995, p. 129).
>
> [I]t is absurd to suggest that morality requires one to care, or act as if one cares, no more about one's own child than a stranger. (Wolf, 1992, p. 244).

1. Introduction

The quotes above reflect a deep philosophical tension regarding how to make moral decisions involving those closest to us. On the one hand, treating all people equally is a key tenet of philosophical ethical theories. On the other hand, close relationships are deeply important in people's lives, and may produce special moral obligations. These competing considerations can give rise to wrenching moral decisions when loved ones are involved.

For decades, the vast majority of empirical research in moral psychology primarily studied people's judgments regarding anonymous strangers (e.g., Greene, Sommerville, Nystrom, Darley, & Cohen, 2001, Greene et al., 2009; Haidt, 2001). Yet psychologists now recognize the importance of studying how relationships affect moral cognition, as a growing body of research reveals that people report dramatically different choices in decisions involving friends and loved ones from those involving strangers. One striking context in which this trend emerges is in deciding how to respond to the moral transgressions of others. Weidman, Sowden, Berg, and Kross (2020) asked participants to imagine witnessing either a close other (e.g., sister, best friend) or a distant other (e.g., dentist, mail carrier) committing a crime. Participants then had to decide whether they would report the perpetrator or lie to protect them when confronted by a police officer. Across several studies, a robust effect emerged: people were much more likely to protect a close (vs. distant) other, and this discrepancy increased with the severity of the crime people observed (also see Waytz, Dungan, & Young, 2013).

But do people think it is *morally right* to behave this way? Against the background of Weidman et al. (2020), we take up the currently underexplored question of whether people believe that relationships *should* influence how they respond to others' transgressions. If people think

* Corresponding author at: Department of Psychology, University of Michigan, 1004 East Hall, 530 Church Street, Ann Arbor, MI 48109, USA.
E-mail addresses: lksoter@umich.edu (L.K. Soter), bergmk@umich.edu (M.K. Berg), gelman@umich.edu (S.A. Gelman), ekross@umich.edu (E. Kross).
[1] Denotes equal contribution.

https://doi.org/10.1016/j.cognition.2021.104886
Received 11 January 2021; Received in revised form 11 August 2021; Accepted 16 August 2021
Available online 21 August 2021

they should treat close and distant others differently, it suggests that people believe moral rules are sensitive to context—an idea that would challenge a common philosophical conception of morality as applying invariantly across people. If, on the other hand, people think they should treat everyone the same regardless of relationship, then this implies a discrepancy between what people think is right and how they would behave. In addition to posing a major challenge to psychological and philosophical theories that rely on consistency, this would demonstrate the striking result that difficult moral decisions about close others may be a domain in which people are particularly likely to fail to do what they think is right. Here we address these questions across four experiments, in the context of decisions about whether to report moral transgressions of a close or distant other.

1.1. The moral universalism hypothesis

One possibility is that people think that they morally should make the same decision regardless of whether the transgressor is a close or distant other (i.e., the Moral Universalism Hypothesis). This idea that moral rules apply equally across people is a widely accepted philosophical principle. All three major philosophical ethical theories—utilitarianism, Kantian deontology, and virtue ethics—incorporate a tenet of impartiality or universalism. Utilitarianism (Mill, 1895) claims that the morally right action is that which maximizes overall happiness, and clearly states that the interests of all people matter equally, regardless of the particular relationships we have with them (see Singer, 1972 for a particularly strong version of this view). Kantian deontology posits that certain acts are always morally impermissible (Kant, 1785). Strict deontological constraints forbid someone from breaking a moral rule even for someone they love. Finally, Aristotle's twelve basic virtues include *justice*—an important component of which is treating others equally and fairly (Aristotle, 2009, Book V; Kraut, 2018).

Despite universalism's status as the dominant moral framework in philosophical ethics (Archard, 1995; Wolf, 1992), very little empirical work has directly explored whether universalism guides laypeople's moral judgments. However, several lines of research indirectly shed light on the role universalism might play in various kinds of moral decisions. For instance, both adults (Dawes, Fowler, Johnson, McElreath, & Smirnov, 2007; Fehr & Fischbacher, 2003) and children starting around age eight (Blake & McAuliffe, 2011; Shaw & Olson, 2011) prefer fair treatment for everyone in economic resource allocation games, and will reject unequal offers that advantage themselves or their ingroup (Elenbaas, Rizzo, Cooley, & Killen, 2016). Yet this research does not speak decisively in favor of universalism, for children also show ingroup and close-relation favoritism when they are in charge of resource distributions (Olson & Spelke, 2008), and often choose to preserve status quo group-based inequalities (Olson, Dweck, Spelke, & Banaji, 2011).

In further support of universalist considerations, children are sensitive to interpersonal bias in judging contexts. By fourth grade, children (like adults) prefer, and attribute greater fairness to, neutral judges over those who have a personal connection (like friendship) to a contestant, especially in subjective contests (Mills & Keil, 2008). Such work provides evidence for children's and adults' strong commitment to fairness and impartiality across contexts.

Finally, both children and adults are sensitive to a distinction between moral and conventional rules (e.g., Nucci & Turiel, 1978; Rizzo, Cooley, Elenbaas, & Killen, 2018; Smetana, 1981). One feature that distinguishes these categories is that moral rules are judged to apply invariantly to all people across all social contexts, whereas conventional rules are permitted to vary based on the relevant social standards (Smetana, 2013). Although the moral/conventional distinction does not directly address decisions regarding interpersonal relationships, it does speak to the characterization of moral rules as holding universally for all people.

These diverse lines of research shed light on some ways in which universalist principles of equality and impartiality play an important psychological role. However, none directly examines people's commitment to moral universalism in the context of high-stakes moral decisions, as the philosophical ethical theories considered above would prescribe. Although these philosophical theories are normative and not psychologically descriptive, it would be striking if their core universalist tenet were not expressed in folk moral judgments to some degree. Yet this result is far from guaranteed; rather, we will now discuss compelling evidence for a competing hypothesis.

1.2. The moral partiality hypothesis

The alternative possibility is that people believe they *should* protect close others who have committed a moral transgression more than distant others (i.e., the Moral Partiality Hypothesis). In the philosophical literature, defenses of *moral partiality* consist of arguing that people *are* morally justified in treating the people closest to them with special moral concern. This idea is not new. It can be traced back as far as Confucius's account of filial piety, which emphasized the special moral duties owed to family members (Confucius, 2005; Csikszentmihalyi, 2020), and the Ten Commandments, which tell people to "honor thy father and thy mother" (King James Bible, 1769/, 2017).

More recently, philosopher Susan Wolf has commented that it would be "absurd to suggest that morality requires one to care, or act as if one cares, no more about one's own child than a stranger" (Wolf, 1992, p. 244). She motivates this claim with a dramatic example: consider a woman in a boating accident who finds herself nearby two drowning children—one is her own child, and one is a stranger—and the ability to save only one (Wolf, 1992; see also Williams, 1981). Such cases have compelled a number of philosophers to argue that close relationships often give rise to distinctive moral obligations to display preferences towards close others (Archard, 1995; Baron, 1991; Lord, 2016; Scheffler, 2010; Williams, 1981).

The Moral Partiality Hypothesis has more direct empirical support than the Moral Universalism hypothesis. Loyalty is, for instance, a basic dimension in Moral Foundations Theory, which has received extensive support (Haidt & Graham, 2007). More specifically, people judge others who fail to help kin more negatively than those who fail to help strangers (Hughes, 2017; McManus, Kleiman-Weiner, & Young, 2020), say that it is more important to help close others (Killen & Turiel, 1998), and believe that people who fail to help those closest to them are less suitable spouses and friends (Everett, Faber, Savulescu, & Crockett, 2018).

These partialist tendencies emerge early: developmental researchers have found that by age eight, children judge someone who fails to help their friend (vs. stranger) to be meaner, suggesting that they believe an unhelpful friend is failing their obligations in a way that an unhelpful stranger is not. Conversely, children judge someone who helps a stranger (vs. friend) to be nicer, suggesting that helping a friend is expected, but helping a stranger surpasses one's obligations and is thus especially commendable (Marshall, Mermin-Bunnell, & Bloom, 2020).

We thus also have promising empirical and philosophical support for the partialist idea that people will say they should protect close others more than distant others. Yet the studies discussed here focus primarily on evaluations of others' helping behavior. Our primary question, in contrast, is how people weigh partialist and universalist considerations in their *own* moral decisions regarding close others (and distant) who have committed transgressions—a question that has not yet been explored in the literature.

1.3. Consistency across judgments

Thus far we have discussed competing hypotheses regarding whether people will judge that it is right to show moral preference for close others. However, the present research question of whether people think they *should* protect close others who have committed a crime arises not in isolation, but against the backdrop of Weidman et al. (2020)'s findings that people think they *would* protect close others more than distant

others. Thus, there is an additional remaining dimension not captured by the discussion thus far: *are people's judgments about what they think they would do consistent with what they think is morally right*?

Much research has established that people are motivated towards maintaining consistency (Festinger, 1957; Higgins, 1987), and a consistently positive moral self-concept (Dunning, 2007; Jordan, Mullen, & Murnighan, 2011; Mazar, Amir, & Ariely, 2008; Monin & Jordan, 2009). Yet admitting one would not do what one believes is morally right would likely challenge one's moral self-concept; thus, people may be psychologically motivated to avoid such an admission. One possible strategy for avoiding this psychological discomfort would be to bring these judgments into alignment: to change one's prediction about how one would behave to match what they judge to be morally right, or to convince themselves that what they would do is the morally right choice. Thus, in support of these psychological goals, we might expect people to report little difference between what they would and should do—especially if they are asked to make both judgments concurrently. Given Weidman et al. (2020)'s findings that people would protect close others more than distant others, this would imply that people also would say that they *should* protect close others more. Accordingly, predicting consistency across "should" and "would" judgments implies either predicting Moral Partiality, or predicting that people's universalist judgments would lead them to change their predictions about how they would act towards close vs. distant others, to bring their "should" judgments into conformity with their "would judgments."

In contrast, if the Moral Universalism Hypothesis is supported and we see the same pattern of behavioral predictions as Weidman et al. (2020), this would leave us with a notable discrepancy between what people think is right, and how they would act. Past research has found that judgments about moral norms, and choices about how to act, arise from distinct psychological processes, and factors that influence one kind of judgment may have little effect on the other (Pletti, Lotto, Buodo, & Sarlo, 2017; Sood & Forehand, 2005; Tassy, Deruelle, Mancini, Leistedt, & Wicker, 2013; Tassy et al., 2012; Yu, Siegel, & Crockett, 2019). Accordingly, it is reasonable to predict that relational closeness might impact these distinct judgments in different ways. There is even some initial evidence supporting this possibility: Kurzban, DeScioli, and Fein (2012) found that in trolley problems, more people say both that they would sacrifice one person to save five others *and also* that doing so is wrong, when the people at stake were kin or friends, than when they were strangers (also see Tassy, Oullier, Mancini, & Wicker, 2013).

Thus, discrepancies across what people should and would do in response to others' moral transgressions would not be without empirical precedent. Such discrepancies suggest that moral decisions about close others may be a domain in which people are particularly likely to fail to do what they think is right. Given how ubiquitous moral decisions involving close others are in real life, this finding would be noteworthy and indicate an important avenue for future investigation, especially insofar as we think it is important for people to successfully live up to their own moral standards.

1.4. The present studies

Across four studies, we seek to adjudicate between the Moral Universalism and Moral Partiality Hypotheses, in the context of responding to others' moral transgressions. As Weidman et al. (2020) argue, witnessing a close other transgress presents people with a unique dilemma between considerations of loyalty and a desire to protect those closest to us and considerations of justice and punishing immoral acts. In contrast, when a distant other transgresses, there is no such conflict. Weidman et al. (2020)'s findings show that when predicting how they *would* behave if faced with this dilemma, loyalty drives people's decisions. In the present work we test whether this pattern persists when people consider what the morally right response is. Thus, our driving question is: do people think that they *should* lie to the police to protect close others from punishment for a crime, in addition to thinking that they *would* do so?

We examined this question in by first asking participants to make just one kind of judgment across a series of dilemmas (Studies 1a and 1b), and then asking them to make both judgments about each dilemma (Study 2). In our final study, we undertook a more in-depth examination of participants' "should" judgments (Study 3). All studies were preregistered on As Predicted, and all preregistered methods were followed unless noted otherwise in the text. All studies were determined to be exempt by the University of Michigan IRB; informed consent was obtained from all participants prior to survey administration. All preregistrations, survey materials, data, and analyses (annotated R scripts) for each study are available at on OSF: https://osf.io/g8962/.

2. Study 1a

We started with the most basic version of our question: do participants who are asked what they *would* do in response to others' moral transgressions give different responses than those asked what they *should* do? Participants were asked to imagine a series of scenarios in which a close (or distant) other committed a high-severity moral transgression, and were asked either what they would, or should, report the transgression. We expected to replicate Weidman et al. (2020)'s findings that people believe they *would* protect close others more than distant others. Our competing hypotheses deliver distinct predictions for "should" judgments: the Moral Partiality Hypothesis predicts that people believe they should protect close others more than distant others. In contrast, the Moral Universalism Hypothesis predicts that people believe they should protect close and distant others equally.

This study was preregistered through As Predicted #31495 (https://aspredicted.org/ee49y.pdf).

2.1. Method

2.1.1. Participants

Four hundred and three English-speakers in the United States were recruited through Amazon Mechanical Turk (61% women, 39% men $M_{age} = 37.24$, $SD = 12.31$). The self-reported racial/ethnic breakdown of the sample (participants could select multiple categories) was: 7% Asian, 9% Black or African American, 5% Hispanic or Latino, 2% Native American, 79% white, and 1% other.

Participants were excluded from analysis according to the following criteria: answering "no" to a validity check question ($N = 4$); saying that English was not their native language ($N = 4$); providing the same name for multiple close/distant other nominations ($N = 16$); and failing the manipulation check ($N = 80$), for a final sample of 299. Due to oversight, manipulation check failure was not included as a pre-registered exclusion criterion; regardless, all results presented below are statistically equivalent when we include participants who failed the manipulation check.

2.1.2. Procedure

Participants were randomly assigned to either the "would" or "should" condition. They were first asked to think of two people whom they considered the closest to them (e.g., father, spouse, sister, best friend) and two of their most distant acquaintances (e.g., mailman, landlord, dentist). For each, participants provided a first name and the nature of the relationship.

Next, participants were presented with eight vignettes. In each, they were asked to imagine that they had witnessed one of the nominees committing a high-severity theft, such as stealing a laptop or a wallet (see Appendix A for full list of scenarios). Participants were then asked to imagine a police officer approaching them and asking whether they had seen anything suspicious. (Henceforth, we will call these "punish-or-protect" dilemmas, for convenience.) Depending on their assigned condition, they were asked either whether they *would* report the

transgressor ("what you really would do"), or whether they *should* report them ("the ideal, right thing to do"). Participants responded using a 6-point Likert scale (1 = "Definitely would/should not report"; 6 = "Definitely would/should report"). After the final trial, they were asked to write about their thought process as they decided how to answer the preceding question; open-ended data was collected for exploratory purposes and was not analyzed for the present research.

A manipulation check asked participants which kind of judgment they had been asked to make. Participants then reported their level of trust in the police (0 = "Very untrustworthy"; 6 = "Very Trustworthy") as an exploratory measure, and completed a set of demographic questions.

2.2. Results

To test the effects of judgment and relationship on reporting, we ran a mixed linear model using the lme4 package in R (Bates, Maechler, Bolker, & Walke, 2015). We included participant and dilemma as random intercepts, which was the maximal model that reached convergence. Likelihood of reporting the act was reverse-coded, so that higher scores indicate a greater likelihood of protecting the perpetrator. We reverse-code reporting as protecting in all studies presented here for the sake of consistency with Weidman et al. (2020)'s methods, and to facilitate comparison across these studies. We probed the interaction between relationship and judgment type with four follow-up tests, using a Tukey correction for multiple comparisons (corrected p-values reported).

Overall, participants' responses indicated more protection for close others ($M = 3.29$, $SD = 1.85$) than distant others ($M = 2.01$, $SD = 1.35$; $b = 1.23$, $t(2084.74) = 27.78$, $p < .001$, 95% CI = [1.15, 1.31]). This pattern emerged for both "would" judgments (simple effect: $b = -1.49$, $t(2084.66) = 26.19$, $p < .001$, 95% CI = [−1.71, −1.27]) and "should" judgments (simple effect: $b = -0.97$, $t(2085.15) = 14.31$, $p < .001$, 95% CI = [−1.23, −0.71]), supporting the Moral Partiality Hypothesis.

However, people also reported that they *would* protect a transgressor ($M = 2.83$, $SD = 1.82$) more than they reported that they *should* ($M = 2.39$, $SD = 1.59$; $b = -0.44$, $p < .01$, 95% CI = [−0.74, −0.14]). This difference was greater when the transgressor was a close other (simple effect: $b = 0.70$, $t(352.67) = -4.53$, $p < .01$, 95% CI = [0.15, 1.25]) versus a distant other (where the difference was not significant; simple effect: $b = 0.18$, $t(352.67) = -1.84$, $p = .24$, 95% CI = [−0.37, 0.18]; closeness x judgment interaction: $b = -0.52$, $t(2085.16) = -5.82$, $p < .001$, 95% CI = [−0.70, −0.34]). These results suggest that there is a discrepancy between what people believe they would do and what they think is morally right regarding close others—indicating that while there is partiality in "should" judgments, that partiality is weaker than for "would" judgments.

All effects of relationship and judgment were equivalent in an exploratory model that added police trust ($M = 3.96$, $SD = 1.50$ on a 0–6 scale) as a covariate; full model statistics are reported in Supplement.

3. Study 1b

Study 1b served two purposes. First, we sought to replicate the findings of Study 1a. Second, we added an explicit contrast between "would" and "should" judgments in the instructions given to the participants. Whereas in 1a, participants had read a description of only the one type of judgment they were asked to make (either "would" or "should" judgments), in 1b, participants read about both types before learning which kind of judgment they were to make. This further clarified the instructions and made an explicit contrast between the judgment types.

This study was preregistered through As Predicted #31493 (https://aspredicted.org/p2gj6.pdf).

3.1. Method

3.1.1. Participants

Three hundred and ninety-nine native English-speakers in the United States were recruited through Amazon Mechanical Turk (56% women, 43% men, 0.3% other; $M_{age} = 39.15$, $SD_{age} = 12.79$). The self-reported racial/ethnic breakdown of participants (where participants could select more than one option) was: 7% Asian, 7% Black or African American, 6% Hispanic or Latino, 1% Native American, 0.3% Native Hawaiian or Pacific Islander, 78% white, and 2% other.

Participants were excluded from analysis according to the same criteria as in 1a: non-native English speakers ($N = 5$); providing the same name twice ($N = 19$), and failing the manipulation check ($N = 59$). This gave us a final sample of $N = 316$. Results did not differ when we included participants who failed the manipulation check.

3.1.2. Procedure

The design of Study 1b was nearly identical to Study 1a, again using a 2 (relationship: close, distant; within-subjects) x 2 (judgment: should, would; between-subjects) design. The only difference was in the instructions: while in 1a, participants had simply been asked to make either "would" or "should" judgments, in 1b the instructions explicitly contrasted the two kinds of judgments. The instructions read, "In such situations, people may think about what they should do (the ideal, right thing to do) or they may think about what they would actually do (how they would behave in the real world)," and then told participants which kind of judgment they would be asked to make.

3.2. Results

We used the same mixed linear model as in 1a, again including participant and dilemma as random intercepts, which was the maximal model that reached convergence. Likelihood of reporting was reverse-coded; higher scores here indicate greater likelihood of protecting the transgressor. We probed the interaction between relationship and judgment with four follow-up tests, using a Tukey correction for multiple comparisons (corrected p-values reported).

Overall, we replicated our results from 1a. Participants were significantly more inclined to protect close others ($M = 3.78$, $SD = 1.94$) than distant others ($M = 2.23$, $SD = 1.57$; $b = 1.49$, $t(2203.21) = 33.10$, $p < .001$, 95% CI = [1.41, 1.57]). This pattern again emerged for both "would" judgments (simple effect: $b = -1.84$, $t(2205.49) = 31.54$, $p < .001$, 95% CI = [−2.06, −1.62]) and "should" judgments (simple effect: $b = -1.14$, $t(22.04.52) = 16.62$, $p < .001$, 95% CI = [−1.40, −0.88]).

As in 1a, people also revealed that they *would* protect a transgressor ($M = 3.39$, $SD = 1.98$) more than they *should* ($M = 2.47$, $SD = 1.72$; $b = -0.93$, $t(314.00) = -6.07$, $p < .001$, 95% CI = [−1.23, −0.63]). The difference was again greater when the transgressor was a close other (simple effect: $b = 1.27$, $t(370.68) = -8.00$, $p < .001$, 95% CI = [0.69, 1.85]), versus a distant other others (simple effect: $b = 0.58$, $t(370.68) = -3.64$, $p = .05$, 95% CI = [−0.004, 1.16]; relationship x judgment interaction: $b = -0.69$, $t(2206.41) = -7.71$, $p < .001$, 95% CI = [−0.87, −0.51]; see Fig. 1).

Results were equivalent when we controlled for police trust ($M = 3.83$; $SD = 1.58$); see Supplement for full model.

Finally, we note that across both Studies 1a and 1b we observed that more people failed the manipulation check—which asked them what kind of judgment they had been asked to make—in the "should" condition (after other exclusions: 89% of failures in 1a; 85% of failures in 1b). Of those "should" failures, most said that they had been asked to say what they *would* do (96% in 1a; 92% in 1b). We interpret this pattern as resulting from the fact that "what should you do" is sometimes used in everyday language to ask a predictive question (about how one is likely to act) rather than a strictly normative one (about how one ought to act). This ambiguity between readings is absent for "would," which asks a clearly predictive question. Thus, although our opening instructions

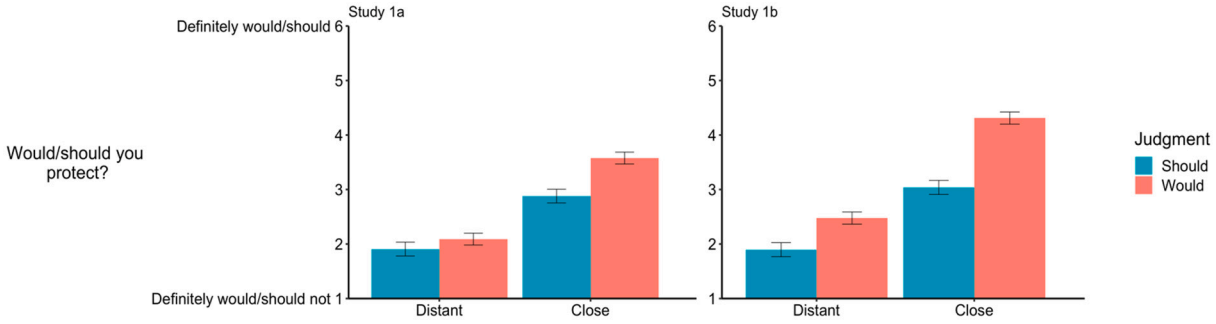

Fig. 1. Effect of Relationship on "Would" and "Should" Judgment: Studies 1a and 1b.
Note. Error bars show standard error of the mean.

asked people what they *ideally should do*—a normative, deontic question—some participants may have strayed towards the predictive reading as they completed the survey. Excluding these participants did not alter the results.

3.3. Studies 1a and 1b discussion

In Studies 1a and 1b, we tested whether participants who were asked what they would do in response to a close (vs. distant) others' moral transgression gave different responses than those who were asked what they should do. In both studies, we replicated Weidman et al. (2020)'s findings that participants would protect close others more than distant others. Participants also said they *should* protect close others more – providing initial support for the Moral Partiality Hypothesis.

However, we also saw a notable discrepancy emerge across judgments: people thought they *would* protect close others more than they *should*, due to the fact that relationship affected "would" judgments more strongly than "should" judgments. These findings thus suggest that people are relatively less partialist when thinking about what is morally right, as opposed to how they would actually act. This resulting discrepancy between judgments also provides initial evidence that moral decisions involving close others may indeed be a context in which people are especially unlikely to do what they think is right.

4. Study 2

Studies 1a and 1b provide initial evidence for two important claims. First, that people believe they should protect close others more than distant others. And second, that relationships influence what people think they should do more weakly than what they would do. These findings suggest that while people believe that relationships influence what the morally right decision is, there is also discrepancy between what people think is right and what they would actually do when it comes to close others.

In Study 2, we sought to test the strength of this discrepancy between "would" and "should" judgments. In Studies 1a and 1b, participants made only one type of judgment, meaning there may not have been much psychological pressure for people to avoid a discrepancy between what they would do and what they say is right. If participants were asked to make both "would" and "should" judgments together, it is possible that they would feel an increased pressure to make those judgments consistent, in order to avoid admitting that they would fail to act morally (by their own lights).

Participants in Study 2 were thus asked both what they would and should do for every dilemma. (See Supplement for two studies in which participants made only one judgment about each dilemma, but where judgment type is varied across dilemma within subjects.) For the discrepancy between judgments to emerge in this within-subjects design, participants would have to be both *self-aware* of this discrepancy, and also *willing to admit* that they would not do what they think is morally right.

Based on our previous results, we predicted that we would still see higher protection of close than distant others, for both judgment types. We further predicted that relationship would influence "would" judgments more strongly than "should" judgments, leading to a discrepancy between what people think they would and should do regarding close others.

This study was preregistered through As Predicted #38203 (https://aspredicted.org/mj4uk.pdf).

4.1. Method

4.1.1. Participants

Four hundred and two native English-speakers in the United States were recruited through Amazon Mechanical Turk (51% women, 48% men, 0.7% other; $M_{age} = 36.38$, $SD_{age} = 11.58$). The self-reported racial/ethnic breakdown of participants (where they could select more than one option) was: 7% Asian, 11% Black or African American, 4% Hispanic or Latino, 0.5% Native American, 77% white, and 1% other.

Participants were excluded for being non-native English speakers ($N = 8$), responding that their data was not valid ($N = 3$), and giving the same name more than once ($N = 35$; all pre-registered exclusion criteria); for a final sample of $N = 356$.

4.1.2. Procedure

Participants again nominated two close and two distant others, and considered the same eight punish-or-protect dilemmas as in the previous studies. Instructions included the explicit contrast between "should" and "would" judgments. After each dilemma, participants were asked both what they should and what they would do. Both questions were presented on the same screen; question order was held constant across dilemmas for each participant, but counterbalanced across participants.

4.2. Results

We used the same linear mixed model method as in previous studies, including participant and dilemma as random intercepts as the maximal model that reached convergence. Likelihood of reporting was again reverse-coded; higher scores here indicate greater likelihood of protecting the transgressor. There was no significant effect of question order across participants; accordingly, we collapsed across order for the analyses reported here. We probed the interaction between relationship and judgment with four follow-up tests using a Tukey correction for multiple comparisons (corrected p-values reported).

We saw largely the same pattern of results as in Studies 1a and 1b. Again, participants showed higher protection for close others ($M = 3.13$, $SD = 1.93$) compared to distant others ($M = 2.18$, $SD = 1.55$; $b = 0.95$, $t(5328.38) = 30.75$, $p < .001$, 95% CI = [0.89, 1.01]). This pattern emerged for both "would" judgments (simple effect: $b = -1.36$, $t(5328.190) = 31.31$, $p < .001$, 95% CI = [−1.51, −1.21]) and "should"

judgments (simple effect: $b = -0.53$, $t(5328.19) = 12.17$, $p < .001$, 95% CI = [−0.68, −0.38]), further supporting the Moral Partiality Hypothesis.

We again saw a difference between what people thought they would do ($M = 3.22$, $SD = 1.89$) and should do ($M = 2.09$, $SD = 1.54$; $b = -1.13$, $t(5327.99) = -36.62$, $p < .001$, 95% CI = [−1.19, −1.07]). This difference emerged for both close (simple effect: $b = 1.54$, $t(5327.99) = -35.47$, $p < .001$, 95% CI = [1.39, 1.69]) and distant others (simple effect: $b = 0.71$, $t(5327.99) = -16.32$, $p < .001$, 95% CI = [0.56, 0.86]), but was greater for close others (interaction: $b = -0.83$, $t(5327.99) = -13.54$, $p < .001$, 95% CI = [−0.95, −0.71]; see Fig. 2).

Results were equivalent when we controlled for police trust ($M = 3.91$, $SD = 1.61$); full model statistics reported in Supplement.

To further explore the distribution of would/should discrepancies, we calculated each participant's average would-should difference for close and distant others; this distribution is plotted in Fig. 3. This visualization reveals a group of participants who showed no difference between what they reportedly would and should do (though, consistent with our primary analyses, a far larger number of participants showed no average would/should discrepancy for trials involving distant others), and a group who indicated a difference, though to varying degree. Future research might explore the psychological predictors and effects of these individual differences.

4.3. Discussion

Study 2 revealed largely the same pattern of results as Studies 1a and 1b. In further support of the Moral Partiality Hypothesis, participants again said that they both would, *and should*, protect close others more than distant others. Relationship again influenced "should" judgments more weakly than "would" judgments, resulting in the key finding of Study 2: that the discrepancy between what participants said they would and should do persisted even when they were asked to make both judgments about each dilemma. Given what we know about people's motivation to maintain a positive and consistent moral self-concept, this discrepancy across judgments is striking: it suggests that people recognize—and are willing to admit—that they would not do what they think is right by their own lights. This finding further supports the hypothesis that difficult moral decisions about close others may be a domain in which people are particularly likely to fail to act as (they believe) they should.

5. Study 3

Thus far, our evidence supports the Moral Partiality Hypothesis: people believe that they should be more lenient regarding close others' moral transgressions. However, we have also seen that relationship influences what people think is right *less* than it influences what they think they would do. This finding suggests delivers a key upshot: participants have admitted to a striking *discrepancy* between what they say they would and should do, especially when it comes to close others.

There remain, however, two available—and importantly different—interpretations of this discrepancy, both of which are consistent with the evidence thus far. The first is that people genuinely think that they *would not* do what they think they *should do*; that is, they are revealing an inconsistency between their predicted actions and their evaluative judgments. Natural extensions of such a finding would include discussions of whether, in displaying this inconsistency, people are being irrational, hypocritical, or akratic (acting against their own considered best judgments).

An alternative view appeals to a "Many Reasons" picture of judgment, which is a widespread (though not entirely uncontroversial) view in philosophy. On this view, one's overall set of reasons to rationally choose to do something include moral reasons (e.g., whether it would harm someone) as well as non-moral "pragmatic" reasons (e.g., whether it would advance one's own personal goals). It follows that a person can rationally and consistently choose to do something even if the moral reasons oppose it, because they are outweighed by pragmatic reasons. That is, all things considered, they "overall" should do that thing, even though they *morally* should not.

To give a concrete example, consider someone at the airport who notices a passport on the ground. Under normal circumstances, they probably ought (morally and overall) to turn it in to security so it can be safely returned to its owner. But on this day, doing so would come at excessive cost: if they take the time to turn in the passport, they will miss their flight. So while they still might have moral reason to turn the wallet in, it is reasonable to say—taking all the relevant reasons into account (moral and pragmatic) —that they ought to leave it and go catch their flight. If, like the passport-finder, people distinguish what they *morally* should do from what they *overall* should do, then perhaps there is no inconsistency at all in our observed difference between "would" and "should" judgments in PP dilemmas. In some circumstances, it may be that failing to act on our moral reasons over practical ones is selfish or otherwise reveals bad character, but there may be nothing *inconsistent* or *irrational* about it.

In Study 3, we sought to adjudicate between these two possible interpretations. We also addressed a related question arising from wording of the "should" questions used thus far: asking people about the "ideal, right thing to do" is arguably ambiguous between the "overall should" and "morally should" interpretations. Accordingly, in this study we asked participants to make one of four judgments about the set of punish-or-protect dilemmas: participants were either asked about what they *actually would* do, *ideally should* do (the language used throughout Studies 1a, 1b, and 2), *morally should* do, or *overall should* do.

We propose two competing hypotheses for this study, based on the conceptual possibilities outlined above. The Genuine Inconsistency Hypothesis predicts that participants will say they *would* protect close others more than they morally *and overall* should protect them. In contrast, the Many Reasons Hypothesis predicts no discrepancy between what people think they would and overall should do, instead revealing that people say they *overall should* protect close others more than they *morally should.*

Finally, it is possible that we will see something in between these two clear hypothesized alternatives: that "overall should" judgments will fall somewhere between "would" and "morally should" judgments. This would suggest that people can distinguish between "should" judgments that encompass a broad range of reasons and those that include merely moral ones, but also that there is an inconsistency between what people think they ought to do and actually would do. This study will serve the additional purpose of revealing whether people were interpreting our "ideally should" question in the prior studies as a specifically moral question (as we intended), or whether they were taking it to ask about a broader, all-things-considered "should" judgment.

This study was preregistered through As Predicted #43538 (https://aspredicted.org/56iu2.pdf).

5.1. Method

5.1.1. Participants

Seven hundred and ninety-eight native English-speakers from the United States were recruited through Prolific (51% women, 47% men, 2% other; $M_{age} = 33.34$, $SD_{age} = 11.91$). This sample size entailed 200 participants per cell, before exclusions, as in all previous studies. The self-reported racial/ethnic makeup of participants was: 9% Asian, 8% Black or African American, 6% Hispanic or Latino, 0.4% Native American, 0.4% Native Hawaiian or Pacific Islander, 72% white, and 2% other.

Participants were excluded from analysis according to the following criteria: duplicate worker IP addresses (cutting just the second response, $N = 4$); non-native English speakers ($N = 3$); not responding to all dilemmas ($N = 1$); saying their data were not valid ($N = 9$); providing the same name more than once ($N = 15$); gibberish or nonsensical answers

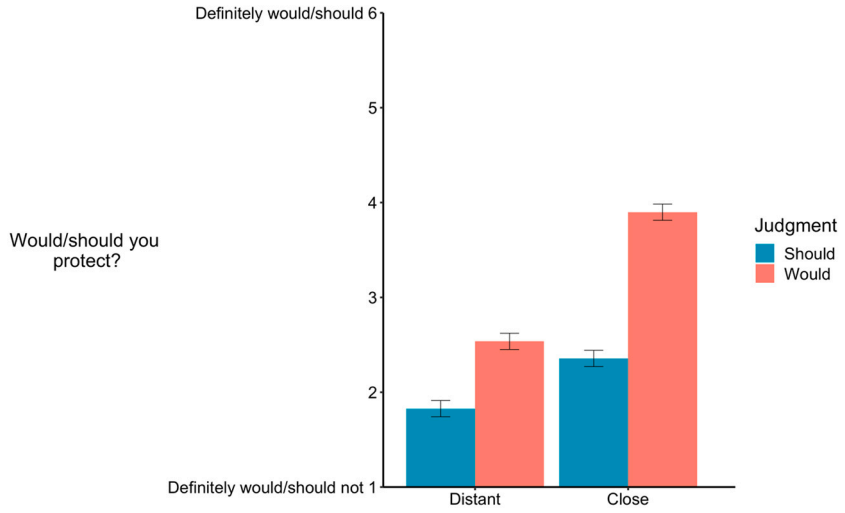

Fig. 2. Study 2: Would/Should Judgments in a within-subjects design.
Note. Error bars show standard error of the mean.

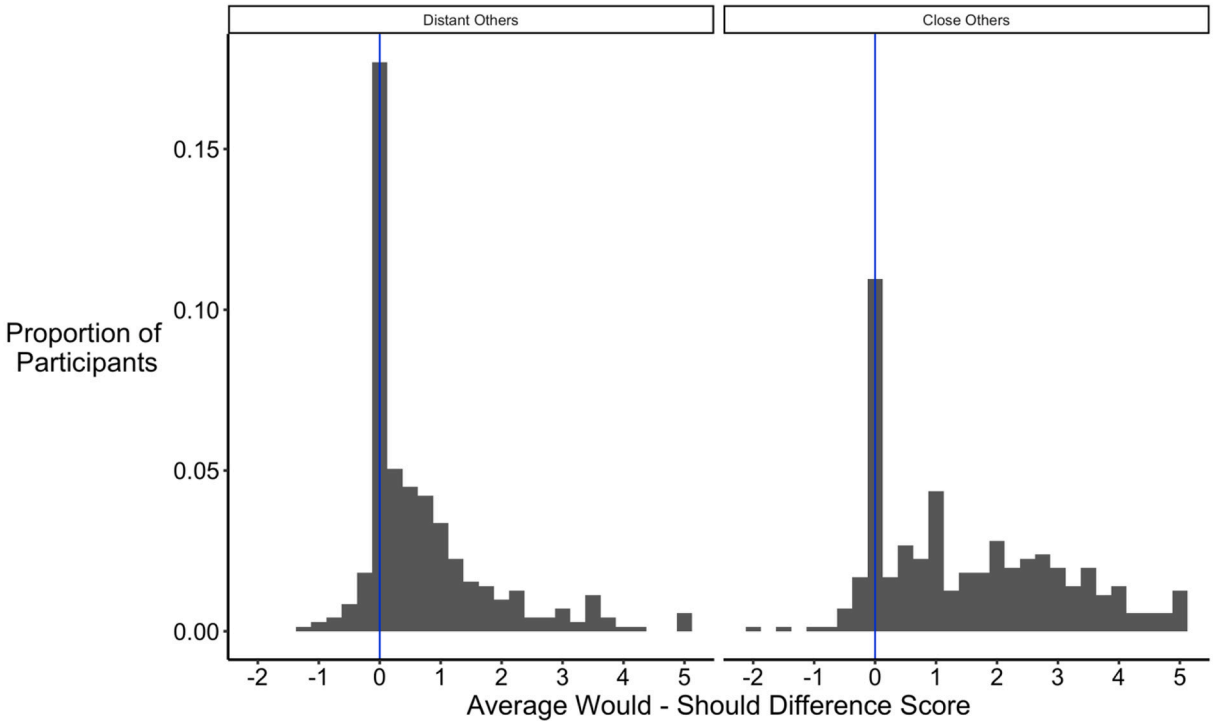

Fig. 3. Would-Should Difference Scores for Close vs. Distant Others.
Note. Difference scores by participant, calculated by subtracting "should" judgments from "would" judgments for each vignette and averaging across vignettes, separately for close vs. distant others.

to the open-ended study probe ($N = 10$); being in the wrong "ballpark" for the judgment manipulation check (e.g., selecting "would" when they were in one of the "should" conditions; $N = 147$).[2] This gave us a final sample of $N = 609$.

5.1.2. Procedure

The basic paradigm was the same as in all previous studies, but with two additional levels in the judgment factor, giving us a 2 (relationship: close, distant; within-subjects) x 4 (judgment: ideally should, morally should, overall should, would; between-subjects) design. Before the dilemmas, participants were presented with one of four sets of instructions telling them what kind of judgment to make: what you ideally should do; what you morally should do; what you overall should do; or what you actually would do; see Table 1 for full instructions. Because this study

[2] We preregistered that we would exclude participants who failed an attention check question, which was disguised as an additional police trust question but asked participants to type a particular response. However, only 304 participants responded correctly; it is possible that it was "too well hidden" as a closing demographic question that came after all the main experimental survey questions were completed. Accordingly, we did not use this question as an exclusion criterion for the analyses reported here.

Table 1
Instructions for each judgment.

Judgment	Instructions
Would	We will be asking you to think about what you **actually would do** – how you would behave in the real world, if you really found yourself in this situation.
Ideally Should	We will be asking you to think about what you **ideally should do** – the ideal, right thing to do.
Morally Should	We will be asking you to think about what you **morally should do** – what the most morally right choice is in this situation.
Overall Should	We will be asking you to think about what you **overall should do** – what the all-things-considered best decision is, when you account for all the factors and complexities of the decision, including both moral and practical considerations.

depended on participants clearly understanding which judgment they were supposed to be making, additional language was included in the presentation of each dilemma reminding participants of which judgment they were being asked for. After participants responded to the eight dilemmas, they completed an exploratory open-ended response question, an exploratory police trust measure, a judgment manipulation check, an attention check, demographic questions, and an open-ended study probe.

5.2. Results

Likelihood of reporting was again reverse-coded; higher scores here indicate greater likelihood of protecting the transgressor. To test our hypotheses, we fit a multilevel model predicting protecting from the type of judgment people made, which was coded as a set of three orthogonal contrasts to test our three primary research questions: do "would" judgments differ from all kinds of "should judgments; do "overall should" judgments differ from "morally should" and "ideally should" judgments; and do "ideally should" and "morally should" judgments differ from each other. The maximal model that reached convergence included participant and dilemma as random intercepts, and relationship as a random slope across participants. To interpret interactions, we ran four follow-up simple effects tests, using a Tukey correction for multiple comparisons (corrected p-values reported). Descriptive statistics for cells are reported in Appendix B.

We replicated the main effect of relationship across judgment types, with overall higher protecting for close others ($M = 3.04$, $SD = 1.82$) than distant others ($M = 2.01$, $SD = 1.42$; $b = 1.01$ $t(604.79) = 19.50$, $p < .001$, 95% CI = [0.91, 1.11]).

The first contrast tested whether "actually would" judgments differed from the three kinds of "should" judgments. "Would" judgments ($M = 3.17$, $SD = 1.84$) elicited higher protecting responses than all "should" judgments ($M = 2.26$, $SD = 1.58$; $b = 0.89$, $t(605.01) = 7.97$, $p < .001$, 95% CI = [0.67, 1.11]), consistent with the Genuine Inconsistency Hypothesis. As in previous studies, the discrepancy between what people said they would and should do was greater for close others (simple effect: $b = 1.20$, $t(702.85) = 10.37$, $p < .001$, 95% CI = [0.96, 1.44]) than distant others (simple effect: $b = 0.56$, $t(702.86) = 4.98$, $p < .05$, 95% CI = [0.32, 0.80]; contrast 1 x relationship interaction: $b = 0.63$, $t(605.29) = 5.60$, $p < .001$, 95% CI = [0.41, 0.85]). (See Supplement for an exploratory analysis comparing only "overall should" and "would" judgments.)

The second contrast tested whether participants' judgments of what they *overall should* do differed from judgments of what they *morally should* do. "Overall should" judgments ($M = 2.57$, $SD = 1.65$) elicited higher protecting responses than "morally should" and "ideally should" judgments ($M = 2.13$, $SD = 1.53$; $b = 0.43$, $t(605.01) = 3.23$, 95% CI = [0.18, 0.68], $p < .01$). The difference between what people overall and morally/ideally should do was greater for decisions about close others (simple effect: $b = 0.63$, $t(702.65) = 1.72$, 95% CI = [0.36, 0.90], $p < .01$) than distant others (simple effect: $b = 0.24$, $t(702.65) = 1.72$, $p = .93$, 95% CI = [−0.03, 0.51]; contrast 2 x relationship interaction: $b = 0.39$, $t(604.78) = 2.90$, $p < .01$, 95% CI = [0.14, 0.64]).

Finally, the third contrast tested for differences between "morally should" ($M = 2.12$, $SD = 1.57$) and "ideally should" judgments ($M = 2.15$, $SD = 1.48$). No significant difference emerged ($b = -0.03$, $t(605.01) = -0.21$, $p = .84$, 95% CI = [−0.32, 0.26]). Results for Study 3 are shown in Fig. 4.

Results were equivalent when we controlled for police trust ($M = 3.07$, $SD = 1.78$); full model statistics reported in Supplement. Results were also equivalent when we applied a more stringent manipulation check exclusion criterion, excluding any participant who truly failed the manipulation check, rather than just those who were in the wrong "ballpark" category ($N = 530$).

5.3. Study 3 discussion

Study 3 examined whether the discrepancy between "should" and "would" judgments documented in Studies 1a, 1b, and 2 reflected a genuine inconsistency between what people thought they should and would do in punish-or-protect dilemmas (the Genuine Inconsistency Hypothesis), or whether participants were *not* showing any inconsistency, but instead thought that what they morally should do is different from what they all-things-considered should do (the Many Reasons Hypothesis).

Consistent with the Genuine Inconsistency Hypothesis, "would" judgments elicited higher protecting responses than all three kinds of "should" judgments, especially for close others. Additionally, "overall should" judgments elicited higher protecting than "morally should" and "ideally should" judgments, particularly for close others. This suggests that people distinguish between specifically moral "oughts" and all-things-considered normative judgments. The fact that people said they *would* protect more than they overall *should* suggests that there is a genuine inconsistency across people's judgments—and that this inconsistency cannot merely be accounted for by appealing to a broader kind of "should" judgment.

Finally, our analyses revealed no difference between "morally should" and "ideally should" judgments, suggesting that the judgments participants have been making across all our studies have indeed been judgments about what they morally should do.

6. General discussion

Across four studies, we examined whether people believe they not only would, but also *should*, preferentially protect close others who have committed severe moral transgressions. In doing so, we tested two competing hypotheses—the Moral Universalism and Moral Partiality Hypotheses—each of which carries an important, but distinct, insight into the nature of moral decision-making.

Our findings provide strong evidence for the Moral Partiality Hypothesis: in each study, participants said they should protect close others more than distant others. However, the effect of relationship was consistently weaker for "should" judgments than "would" judgments, revealing that people show *relatively less* partiality in their judgments of what is morally right, compared to judgments of how they would act. These findings suggest that when it comes to difficult moral decisions about close others, people think that they would fail to do what is by their own standards "right."

These studies provide further evidence that judgments about moral rules versus decisions about how to act may be sensitive to different considerations (Pletti et al., 2017; Tassy et al., 2012; Tassy, Deruelle, et al., 2013; Yu et al., 2019). This fits with the growing appreciation of moral cognition as a multifaceted area that involves many distinct kinds of judgments, including whether actions are right or wrong (e.g., Cipolletti, McFarlane, & Weissglass, 2016; Greene et al., 2009), what to do in hypothetical, lab-based situations (e.g., Bostyn, Sevenhant, & Roets, 2018; FeldmanHall et al., 2012; Francis et al., 2016; Patil, Cogoni,

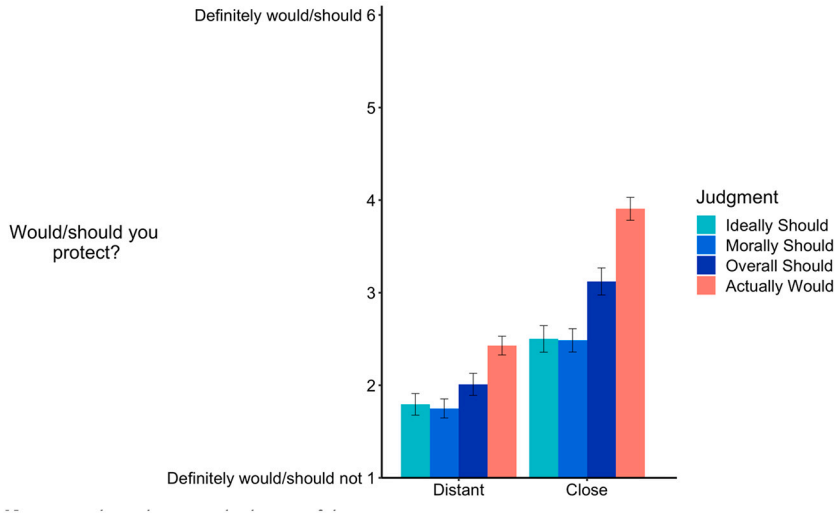

Fig. 4. Study 3: protecting decisions across four judgments.
Note. Error bars show standard error of the mean.

Zangrando, Chittaro, & Silani, 2014), what to do when faced with real-world moral choices (Hofmann, Wisneski, Brandt, & Skitka, 2014; Johnson & Goldstein, 2003; Rosenbaum, 2009), and how to evaluate people's moral character (Siegel, Crockett, & Dolan, 2017; Siegel, Mathys, Rutledge, & Crockett, 2018).

Our findings also further inform discussions about how relationships and other elements of identity and context influence moral judgments (Hester & Gray, 2020; Schein, 2020). The present work provides additional evidence that relationships influence moral decision-making, and that people think close relationships carry moral weight and generate moral obligations. However, it also shows that different kinds of moral judgments are influenced by relational considerations to different degrees. Further exploring how relational and contextual factors influence a diverse set of moral judgments will help us gain a fuller understanding of the real-world functioning of moral cognition. On a practical level, our findings reinforce an important methodological lesson for moral psychology researchers: it matters both conceptually and empirically what kind of judgment we ask participants to make, and we cannot assume that there will be no difference between what people think they would and should do (see also Barbosa & Jimenez Leal, 2017).

Our findings also are relevant to a number of topics of interest to researchers working at the intersection of psychology and philosophy. Researchers studying hypocrisy—particularly as it relates to inconsistency, self-deception, and akratic thinking—may be particularly interested in our finding that people willingly admit they do not think they would do what they think they should (Alicke, Gordon, & Rose, 2013; Bartel, 2019; Batson, Thompson, Seuferling, Whitney, & Strongman, 1999; Jordan, Sommers, Bloom, & Rand, 2017; Laurent & Clark, 2019; Lönnqvist, Irlenbusch, & Walkowitz, 2014). Our findings may shed light on how laypeople think about the relation between moral and practical reasons for action—a topic that ties into many debates in philosophical ethics and action theory. The present work is also relevant for researchers engaged in the philosophical moral partiality debate, and holds particular promise for philosophers who take a more naturalistic and empirically-informed approach to their theorizing (even if the empirical documentation of folk intuitions about moral dilemmas cannot on its own settle normative philosophical questions).

6.1. Future directions

An important question that we did not address concerns the mechanisms underlying the differences between "would" and "should" judgments, and between responses for close and distant others (especially for "should" judgments; see Weidman et al., 2020, for mechanisms underlying differences in what people would do).

First, future work should explore what leads people to say they *morally should* protect close others more. One possibility is that when close others are involved, one has to consider competing virtues of *justice* and *loyalty*. Thus, people's normative endorsement of moral partiality could reflect the belief that loyalty is an important moral virtue, and that close relationships give rise to special moral obligations. People might also feel more empathy towards close others than distant others (as they do for ingroup members; see e.g., Gutsell & Inzlicht, 2012; Tarrant, Dazeley, & Cottom, 2009). This could lead people to incorporate an expanded set of moral considerations beyond just justice (Batson, Klein, Highberger, & Shaw, 1995), could make them more inclined to believe in the underlying goodness of close others, and/or could lead them to assume that the person had a good reason for acting as they did—all of which might lead people to judge the need for (legal) punishment to be less pressing. Finally, attentional mechanisms may contribute to this effect: Berg, Kitayama, and Kross (2021) show that people tend to focus on the *actor* when a close other transgresses, but the *crime* when a distant other does so. These differences in focus could affect not only the decisions people make (Berg et al.'s focus) but also their conclusions about what is the right thing to do.

These empathic processes, character attributions, and attentional mechanisms might also be mechanisms that underlie the would/should discrepancy. For example, people's empathy towards a close other might make it more challenging for them to do what they think justice requires in these situations (Batson et al., 1995). Another potential explanation for the would/should discrepancy involves reputational concerns and the relational stakes of acting impartially towards close others. People sometimes say that making the morally right decision makes one a worse friend or spouse (Everett et al., 2018), and that leaders who do the morally right thing are judged as cold and lacking empathy (Uhlmann, Lei Zhu, & Tannenbaum, 2013). Thus, people who say they would not do what they should, might be worried about reputational or relational costs of doing what they believe to be right. However, this does not seem to explain why people think they would protect more than they *all things considered* should protect (Study 3)—unless they think that these practical reasons would influence their actions in a manner that is not justified. Our findings also suggest that people think they would protect *distant* others more than they should (though the effect was smaller than for close others). This could be driven by fear of retaliation or worries about the consequences of engaging with law enforcement. These numerous possibilities emphasize that there is much future research to

be done to disentangle all of these possible mechanisms and their implications.

Another important future question is whether people *actually* act in real-life scenarios how they *think* they will in hypothetical ones. On the one hand, some work suggests that people may actually behave more morally than they think they will (Teper, Inzlicht, & Page-Gould, 2011), positing that affective forecasting failures can lead to poor behavioral predictions. Thus, perhaps people are actually more likely to report a close other (i.e., act as they say they should) than they anticipate. On the other hand, it might be that people are actually rather successful affective forecasters in this domain, if it turns out that thinking about a close other's transgression is highly emotional. Though the general emotionality of these situations has not been directly tested, Weidman et al. (2020) show that psychological distancing—a common strategy for regulating emotions—makes people less likely to say they would report a close other. Given these competing predictions, the relation of "would" and "should" judgments to actual behavior marks an exciting area for future research.

6.2. Concluding comment

There are several key takeaways from the present findings. First, we have demonstrated that people believe it is *morally right* to treat close others differently than distant others. This suggests that relationship is a powerful factor in folk ethical theory. Second, we have shown that—despite their preference for moral partiality—people also believe that they *would* protect close others more than they *should*, suggesting that people believe they are likely to fail to do what they think is right in moral decisions involving close others. Third, our work has relevance to a number of topics of interest to philosophers, including the relation between moral and practical reasons in folk psychology, and hypocrisy, inconsistency, and akratic thinking. Our results carry both theoretical implications for understanding how relationships influence moral judgments and how different kinds of moral judgments relate to each other, and practical methodological implications for moral psychologists investigating different kinds of moral judgments. Overall, these findings reinforce the claim that decisions involving close others remains a lively domain and fruitful area for moral psychological research.

Acknowledgements

We thank Chandra Sripada for comments on an earlier version of the manuscript. We also thank the members of the Kross Emotion and Self-Control Lab, the University of Michigan Language and Cognition Lab Group, and University of Michigan Weinberg Institute for Cognitive Science Seminar Series—particularly Rick Lewis—for their feedback on this work. This work was supported by the University of Michigan, Ann Arbor, and the National Science Foundation Graduate Research Fellowship.

Declarations of Competing Interest

None.

Funding

This work was supported by the University of Michigan, Ann Arbor, and the National Science Foundation Graduate Research Fellowship.

Credit statement

Laura Soter: Conceptualization, Methodology, Software, Formal Analysis, Investigation, Writing (original draft; review & editing) **Martha Berg:** Software, Formal Analysis, Resources, Writing (review & editing) **Susan Gelman:** Conceptualization, Methodology, Writing (review & editing), Supervision; **Ethan Kross:** Conceptualization, Methodology, Writing (review & editing), Supervision, Funding Acquisition.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2021.104886.

Appendix A

Below is the set of theft dilemmas used in the studies presented here.

1. Stealing a wallet from the seat of a parked car.
2. Stealing an unattended laptop in a coffee shop.
3. Taking a dog tied up outside a shop.
4. Stealing jewelry from a locked store display.
5. Blackmailing someone for money by threatening to post unflattering pictures of them online.
6. Breaking into a house and stealing a TV.
7. Taking money out of a charity donations basket.
8. Taking a credit card left on a restaurant table.

Appendix B

Table 2 displays the cell descriptive statistics for Study 3.

Table 2
Mean Protecting across Judgment Conditions (1–6 Scale).

	Ideally Should	Morally Should	Overall Should	Would
Close	2.49 (1.61)	2.49 (1.74)	3.12 (1.71)	3.91 (1.77)
Distant	1.78 (1.24)	1.75 (1.30)	2.00 (1.35)	2.42 (1.60)
Total	2.14 (1.48)	2.12 (1.58)	2.55 (1.64)	3.17 (1.84)

Note. Standard deviations in parentheses.

References

Alicke, M., Gordon, E., & Rose, D. (2013). Hypocrisy: What counts? *Philosophical Psychology, 26*(5), 673–701. https://doi.org/10.1080/09515089.2012.677397.

Archard, D. (1995). Moral partiality. *Midwest Studies In Philosophy, 20*(1), 129–141. https://doi.org/10.1111/j.1475-4975.1995.tb00308.x.

Aristotle. (2009). The Nicomachean Ethics (L. Brown, Ed.; D. Ross, Trans.). Oxford University Press.

Barbosa, S., & Jimenez Leal, W. (2017). It's not right but it's permitted: Wording effects in moral judgement. *Judgment and Decision making, 12*, 308–313.

Baron, M. (1991). Impartiality and friendship. *Ethics, 101*(4), 836–857. https://doi.org/10.1086/293346.

Bartel, C. (2019). Hypocrisy as either deception or Akrasia. *The Philosophical Forum, 50*(2), 269–281. https://doi.org/10.1111/phil.12220.

Bates, D., Maechler, M., Bolker, B., & Walke, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of statistical software, 67*(1), 1–48. Doi: 10.18637/jss.v067.i01.

Batson, C. D., Klein, T. R., Highberger, L., & Shaw, L. L. (1995). Immorality from empathy-induced altruism: When compassion and justice conflict. *Journal of Personality and Social Psychology, 68*(6), 1042–1054. https://doi.org/10.1037/0022-3514.68.6.1042.

Batson, C. D., Thompson, E. R., Seuferling, G., Whitney, H., & Strongman, J. A. (1999). Moral hypocrisy: Appearing moral to oneself without being so. *Journal of Personality and Social Psychology, 77*(3), 525–537. https://doi.org/10.1037/0022-3514.77.3.525.

Berg, M. K., Kitayama, S., & Kross, E. (2021). How relationships bias moral reasoning: Neural and self-report evidence. *Journal of Experimental Social Psychology, 95*, 104156. https://doi.org/10.1016/j.jesp.2021.104156.

Blake, P. R., & McAuliffe, K. (2011). "I had so much it didn't seem fair": Eight-year-olds reject two forms of inequity. *Cognition, 120*(2), 215–224. https://doi.org/10.1016/j.cognition.2011.04.006.

Bostyn, D. H., Sevenhant, S., & Roets, A. (2018). Of mice, men, and trolleys: Hypothetical judgment versus real-life behavior in trolley-style moral dilemmas. *Psychological Science, 29*(7), 1084–1093. https://doi.org/10.1177/0956797617752640.

Cipolletti, H., McFarlane, S., & Weissglass, C. (2016). The moral foreign-language effect. *Philosophical Psychology, 29*(1), 23–40. https://doi.org/10.1080/09515089.2014.993063.

Confucius. (2005). *The Analects of Confucius (A. Waley, Trans.)*. Psychology Press.

Csikszentmihalyi, M. (2020). Confucius. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy (summer 2020)*. Metaphysics Research Lab: Stanford University. https://plato.stanford.edu/archives/sum2020/entries/confucius/.

Dawes, C. T., Fowler, J. H., Johnson, T., McElreath, R., & Smirnov, O. (2007). Egalitarian motives in humans. *Nature, 446*(7137), 794–796. https://doi.org/10.1038/nature05651.

Dunning, D. (2007). Self-image motives and consumer behavior: How sacrosanct self-beliefs sway preferences in the marketplace. *Journal of Consumer Psychology, 17*(4), 237–249. https://doi.org/10.1016/S1057-7408(07)70033-5.

Elenbaas, L., Rizzo, M. T., Cooley, S., & Killen, M. (2016). Rectifying social inequalities in a resource allocation task. *Cognition, 155*, 176–187. https://doi.org/10.1016/j.cognition.2016.07.002.

Everett, J. A. C., Faber, N. S., Savulescu, J., & Crockett, M. J. (2018). The costs of being consequentialist: Social inference from instrumental harm and impartial beneficence. *Journal of Experimental Social Psychology, 79*, 200–216. https://doi.org/10.1016/j.jesp.2018.07.004.

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature, 425*(6960), 785–791. https://doi.org/10.1038/nature02043.

FeldmanHall, O., Mobbs, D., Evans, D., Hiscox, L., Navrady, L., & Dalgleish, T. (2012). What we say and what we do: The relationship between real and hypothetical moral choices. *Cognition, 123*(3), 434–441. https://doi.org/10.1016/j.cognition.2012.02.001.

Festinger, L. (1957). *A theory of cognitive dissonance (Vol. 2)*. Stanford University Press.

Francis, K. B., Howard, C., Howard, I. S., Gummerum, M., Ganis, G., Anderson, G., & Terbeck, S. (2016). Virtual morality: Transitioning from moral judgment to moral action? *PLoS One, 11*(10), Article e0164374. https://doi.org/10.1371/journal.pone.0164374.

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition, 111*(3), 364–371. https://doi.org/10.1016/j.cognition.2009.02.001.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*(5537), 2105–2108. https://doi.org/10.1126/science.1062872.

Gutsell, J. N., & Inzlicht, M. (2012). Intergroup differences in the sharing of emotive states: Neural evidence of an empathy gap. *Social Cognitive and Affective Neuroscience, 7*(5), 596–603. https://doi.org/10.1093/scan/nsr035.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814–834.

Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research, 20*(1), 98–116. https://doi.org/10.1007/s11211-007-0034-z.

Hester, N., & Gray, K. (2020). The moral psychology of Raceless, Genderless Strangers. *Perspectives on Psychological Science, 174569161988584*. https://doi.org/10.1177/1745691619885840.

Higgins, E. T. (1987). Self-discrepancy: A theory relating self and affect. *Psychological Review, 54*(9), 319–340.

Hofmann, W., Wisneski, D. C., Brandt, M. J., & Skitka, L. J. (2014). Morality in everyday life. *Science, 345*(6202), 1340–1343. https://doi.org/10.1126/science.1251560.

Hughes, J. S. (2017). In a moral dilemma, choose the one you love: Impartial actors are seen as less moral than partial ones. *British Journal of Social Psychology, 56*(3), 561–577. https://doi.org/10.1111/bjso.12199.

Johnson, E. J., & Goldstein, D. (2003). Do defaults save lives? *Science, 302*(5649), 1338–1339. https://doi.org/10.1126/science.1091721.

Jordan, J., Mullen, E., & Murnighan, J. K. (2011). Striving for the moral self: The effects of recalling past moral actions on future moral behavior. *Personality and Social Psychology Bulletin, 37*(5), 701–713. https://doi.org/10.1177/0146167211400208.

Jordan, J. J., Sommers, R., Bloom, P., & Rand, D. G. (2017). Why do we hate hypocrites? Evidence for a theory of false signaling. *Psychological Science, 28*(3), 356–368. https://doi.org/10.1177/0956797616685771.

Kant, I. (1785). *Grounding for the metaphysics of morals (J. W. Ellington, trans.)*. Hackett Publishing Company.

Killen, M., & Turiel, E. (1998). Adolescents' and Young Adults' evaluations of helping and sacrificing for others. *Journal of Research on Adolescence, 8*(3), 355–375. https://doi.org/10.1207/s15327795jra0803_4.

King James Bible. (2017). Cambridge University Press.

Kraut, R. (2018). Aristotle's ethics. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy (summer 2018)*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2018/entries/aristotle-ethics/.

Kurzban, R., DeScioli, P., & Fein, D. (2012). Hamilton vs. Kant: Pitting adaptations for altruism against adaptations for moral judgment. *Evolution and Human Behavior, 33*(4), 323–333. https://doi.org/10.1016/j.evolhumbehav.2011.11.002.

Laurent, S. M., & Clark, B. A. M. (2019). What makes hypocrisy? Folk definitions, attitude/behavior combinations, attitude strength, and private/public distinctions. *Basic and Applied Social Psychology, 41*(2), 104–121. https://doi.org/10.1080/01973533.2018.1556160.

Lönnqvist, J.-E., Irlenbusch, B., & Walkowitz, G. (2014). Moral hypocrisy: Impression management or self-deception? *Journal of Experimental Social Psychology, 55*, 53–62. https://doi.org/10.1016/j.jesp.2014.06.004.

Lord, E. (2016). Justifying partiality. *Ethical Theory and Moral Practice, 19*(3), 569–590.

Marshall, J., Mermin-Bunnell, K., & Bloom, P. (2020). Developing judgments about peers' obligation to intervene. *Cognition, 201*, 104215. https://doi.org/10.1016/j.cognition.2020.104215.

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research, 45*(6), 633–644.

McManus, R. M., Kleiman-Weiner, M., & Young, L. (2020). What we owe to family: The impact of special obligations on moral judgment. *Psychological Science, 0956797619900321*. https://doi.org/10.1177/0956797619900321.

Mill, J. S. (1895). *Utilitarianism*. Green and Company: Longmans.

Mills, C. M., & Keil, F. C. (2008). Children's developing notions of (im)partiality. *Cognition, 107*(2), 528–551. https://doi.org/10.1016/j.cognition.2007.11.003.

Monin, B., & Jordan, A. H. (2009). The dynamic moral self: A social psychological perspective. In D. Narvaez, & D. K. Lapsley (Eds.), *Personality, identity, and character (pp. 341–354)*. Cambridge University Press. https://doi.org/10.1017/CBO9780511627125.016.

Nucci, L. P., & Turiel, E. (1978). Social interactions and the development of social concepts in preschool children. *Child Development, 49*(2), 400–407. JSTOR https://doi.org/10.2307/1128704.

Olson, K. R., Dweck, C. S., Spelke, E. S., & Banaji, M. R. (2011). Children's responses to group-based inequalities: Perpetuation and rectification. *Social Cognition, 29*(3), 270–287. https://doi.org/10.1521/soco.2011.29.3.270.

Olson, K. R., & Spelke, E. S. (2008). Foundations of cooperation in young children. *Cognition, 108*(1), 222–231. https://doi.org/10.1016/j.cognition.2007.12.003.

Patil, I., Cogoni, C., Zangrando, N., Chittaro, L., & Silani, G. (2014). Affective basis of judgment-behavior discrepancy in virtual experiences of moral dilemmas. *Social Neuroscience, 9*(1), 94–107. https://doi.org/10.1080/17470919.2013.870091.

Pletti, C., Lotto, L., Buodo, G., & Sarlo, M. (2017). It's immoral, but I'd do it! Psychopathy traits affect decision-making in sacrificial dilemmas and in everyday moral situations. *British Journal of Psychology, 108*(2), 351–368. https://doi.org/10.1111/bjop.12205.

Rizzo, M. T., Cooley, S., Elenbaas, L., & Killen, M. (2018). Young children's inclusion decisions in moral and social–conventional group norm contexts. *Journal of Experimental Child Psychology, 165*, 19–36. https://doi.org/10.1016/j.jecp.2017.05.006.

Rosenbaum, J. E. (2009). Patient teenagers? A comparison of the sexual behavior of virginity pledgers and matched nonpledgers. *Pediatrics, 123*(1), e110–e120. https://doi.org/10.1542/peds.2008-0407.

Scheffler, S. (2010). Partiality and impartiality. *Oxford University Press.*. https://doi.org/10.1093/acprof:oso/9780199579952.001.0001.

Schein, C. (2020). The importance of context in moral judgments. *Perspectives on Psychological Science, 15*(2), 207–215. https://doi.org/10.1177/1745691620904083.

Shaw, A., & Olson, K. (2011). Children discard a resource to avoid inequity. *Journal of Experimental Psychology. General, 141*, 382–395. https://doi.org/10.1037/a0025907.

Siegel, J. Z., Crockett, M. J., & Dolan, R. J. (2017). Inferences about moral character moderate the impact of consequences on blame and praise. *Cognition, 167*, 201–211. https://doi.org/10.1016/j.cognition.2017.05.004.

Siegel, J. Z., Mathys, C., Rutledge, R. B., & Crockett, M. J. (2018). Beliefs about bad people are volatile. *Nature Human Behaviour, 2*(10), 750–756. https://doi.org/10.1038/s41562-018-0425-1.

Singer, P. (1972). Famine, affluence, and morality. *Philosophy & Public Affairs, 1*(3), 229–243. JSTOR.

Smetana, J. (2013). Young Children's moral and social-conventional understanding. In M. Banaji, & S. Gelman (Eds.), *Navigating the social world: What infants, children, and other species can teach us (pp. 352–35)*. Oxford University Press.

Smetana, J. G. (1981). Preschool Children's conceptions of moral and social rules. *Child Development, 52*(4), 1333–1336. JSTOR https://doi.org/10.2307/1129527.

Sood, S., & Forehand, M. (2005). On self-referencing differences in judgment and choice. *Organizational Behavior and Human Decision Processes, 98*(2), 144–154. https://doi.org/10.1016/j.obhdp.2005.05.005.

Tarrant, M., Dazeley, S., & Cottom, T. (2009). Social categorization and empathy for outgroup members. *British Journal of Social Psychology, 48*(3), 427–446. https://doi.org/10.1348/014466608X373589.

Tassy, S., Deruelle, C., Mancini, J., Leistedt, S., & Wicker, B. (2013). High levels of psychopathic traits alters moral choice but not moral judgment. *Frontiers in Human Neuroscience, 7*. https://doi.org/10.3389/fnhum.2013.00229.

Tassy, S., Oullier, O., Duclos, Y., Coulon, O., Mancini, J., Deruelle, C., Attarian, S., Felician, O., & Wicker, B. (2012). Disrupting the right prefrontal cortex alters moral judgement. *Social Cognitive and Affective Neuroscience, 7*(3), 282–288. https://doi.org/10.1093/scan/nsr008.

Tassy, S., Oullier, O., Mancini, J., & Wicker, B. (2013). Discrepancies between judgment and choice of action in moral dilemmas. *Frontiers in Psychology, 4*. https://doi.org/10.3389/fpsyg.2013.00250.

Teper, R., Inzlicht, M., & Page-Gould, E. (2011). Are we more moral than we think?: Exploring the role of affect in moral behavior and moral forecasting. *Psychological Science, 22*(4), 553–558. https://doi.org/10.1177/0956797611402513.

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

Waytz, A., Dungan, J., & Young, L. (2013). The whistleblower's dilemma and the fairness–loyalty tradeoff. *Journal of Experimental Social Psychology, 49*(6), 1027–1033. https://doi.org/10.1016/j.jesp.2013.07.002.

Weidman, A. C., Sowden, W. J., Berg, M. K., & Kross, E. (2020). Punish or protect? How close relationships shape responses to moral violations. *Personality and Social Psychology Bulletin, 46*(5), 693–708. https://doi.org/10.1177/0146167219873485.

Williams, B. (1981). *Moral luck: Philosophical papers 1973–1980*. Cambridge University Press.

Wolf, S. (1992). Morality and partiality. *Philosophical Perspectives, 6*, 243–259. JSTOR https://doi.org/10.2307/2214247.

Yu, H., Siegel, J. Z., & Crockett, M. J. (2019). Modeling morality in 3-D: Decision-making, judgment, and inference. *Topics in Cognitive Science, 11*(2), 409–432. https://doi.org/10.1111/tops.12382.