Beyond good and bad: Varieties of moral judgment

William Jiménez-Leal[1], Samuel Murray[2], Santiago Amaya[3], and Sergio Barbosa[1,4]

[1]Deparment of Psychology, Universidad de los Andes

[2]Mind At Large Lab, Imagination and Modal Cognition Lab, Duke University

[3]Department of Philosophy, Universidad de los Andes

[4]Department of Psychology, Universidad del Rosario

Word Count: 15910.

Author Note

The first three authors contributed equally. Correspondence concerning this article should be addressed to William Jiménez-Leal, Universidad de los Andes. E-mail: w.jimenezleal@uniandes.edu.co.

Supplementary materials (including materials, preregistrations, raw data and code) available at: https://osf.io/kja2u/?view_only=573c1541bbd240b397b79229704dfae7.

Abstract

We argue that people regularly encounter situations involving moral conflicts among permissible options. These scenarios, which some have called *morally charged situations*, reflect perceived tensions between moral expectations and moral rights. Studying responses to such situations marks a departure from the common emphasis on sacrificial dilemmas and widespread use of single-dimension measures. In 6 experiments (n=1607), we show that people use a wide conceptual arsenal when assessing actions that can be described as suberogatory (bad but permissible) or supererogatory (good but not required). In Experiment 1 we find that people identify actions as suberogatory or supererogatory when using open descriptions to describe them. Experiment 2 shows that they differentially assess these actions in terms of how permissible, optional, and good they considered them. Experiment 3 tests the use of these evaluative dimensions with sacrificial dilemmas. We fail to find differences between these categories when people respond to dilemmas, even when controlling for trait utilitarian tendencies. By including judgments of blameworthiness and sanction, Experiments 4 and 5 provide additional evidence of the granularity and the moral significance of these evaluations. In Experiment 6 people offered their own explanations of their responses. Qualitative analyses revealed that they frequently appeal to character traits, the presence of rights, and the absence of explicit duties. Taken together these results suggest a richer spectrum of both situations and concepts relevant to characterize moral judgment than moral psychologists up to this point have generally recognized.

Keywords: moral rules, moral dilemmas, suberogatory, supererogatory, duty

Beyond good and bad: Varieties of moral judgment

The role of dilemmas in the science of morality

Dilemmas have traditionally played a prominent role in the study of moral cognition. While many have recently registered dissatisfaction with this state of affairs (Andrade, 2019; Bauman, Mcgraw, Bartels & Warren 2014; Dahl & Oftedal, 2019; Everett & Kahane, 2020; Kahane, 2020), there is considerably less said about how historically we got here. Dilemmas take on this crucial role because of the assumption that morality is a system of rules. But this assumption is problematic for reasons pertaining to both normative theories of morality and the psychology of moral judgment.

From its early days, psychologists interested in moral judgment understood morality as a structured set of rules prohibiting and prescribing certain behaviors. Piaget (1932), for instance, opened his landmark *The Moral Judgment of the Child* with the statement: "All morality consists in a system of rules, and the essence of all morality is to be sought for in the respect which the individual acquires for these rules" (1932, p. 1). Fauconnet echoes this sentiment in his *Responsibility*, stating that moral responsibility is a "quality belonging to those who must…in virtue of a rule be chosen as the passive subjects of a punishment" (1920, p. 11).

Thinking of morality as a system of rules has several theoretical pay-offs. For one, various deontic concepts can be inter-defined through rules. For example, acts are permissible if and only if those acts accord with the rules, and something counts as good only if it accords with the rules. Likewise, the impermissible is whatever constitutes violating a rule, and something is bad insofar as doing it violates a rule (see, for example, Kanger (1971) and Anderson (1958) for

two formal attempts to explain inter-definability). This, as we shall see, results in a methodological advantage. If true, asking subjects whether something is impermissible, bad, or in violation of a duty comes close to asking one and the same question.

Focusing on rules also provides a concrete way to measure moral development. Under the assumption that maturing moral judgment consists in possessing a greater moral understanding, it is possible to define this understanding in terms of increasing aptitude in applying more sophisticated evaluative rules (Kohlberg & Hersh, 1977). This insight famously informed the stage theory of moral development (Kohlberg, Levine, & Hewer, 1983), which identified moral stages according to the kinds of rules that informed judgments: from pre-conventional rules ("Doing this is bad because I can get punished") to universal, exceptionless principles of impartial justice ("Killing is wrong").

Moral dilemmas are one specific kind of *moral encounter* (Monin, Pizarro & Beer, 2007). They characteristically present situations requiring decisions among impermissible options, in the sense that each violates some plausible moral imperative (Sinnott-Armstrong, 1988: 29-30) In this respect, they isolate different sets of rules (egocentric v. altruistic, in-group v. out-group, etc.) that are normally taken to inform moral cognition. Thus, by looking at the choices people make in these situations or by studying how they assess the decisions made by actors depicted in them, it seems possible to better understand which rules inform their judgments of goodness and badness. For similar reasons, moral dilemmas also seem to provide a good instrument to measure moral development and test hypotheses about individual developmental trajectories.

Obviously, theories of moral cognition that fall under this rule-based tradition differ from one another in important respects (see Darley & Schultz, 1990 for discussion). A sign of this are the well-known controversies that exist among them. Kohlberg's research program, for example, was criticized for the use of culturally biased materials (Simpson, 1974; Snarey, 1985), biased samples (Walker, 1984) and the idea of a linear progression in moral development (Rest, 1979). Others criticized the kind of dilemmas used to elicit moral judgment, focusing on the artificial nature or the mundane character of them (Rosen, 1980; Bauman et al., 2014). At bottom, however, many of Kohlberg's critics agreed that morality was a system of rules that could be studied by means of situations where plausible moral rules conflict with one another. They just disagreed about how to properly characterize these rules and the situations that best exemplified these conflicts.[1]

Rules and Commonsense Morality

Some contemporary psychologists have sought to take distance from this early tradition by proposing dual models of moral cognition (for discussion, see Crockett, 2013). We believe, however, that these models are in an important respect a *continuation of the rule-based tradition* that has dominated the study of moral cognition. While opening up new possibilities to

[1] One notable exception comes from feminist critiques of ethical theory. Several feminist ethicists have noted that the central preoccupation of moral theorizing prioritizes abstract principles over the particularity of ethical life (Gilligan, 1982. p. 32-38). The emphasis on abstract generalization obscures the details that are crucially important to a well-lived life (Young, 1987, p. 61-62).

understand the architecture of moral decision-making, the dual process paradigm has inherited from its early predecessors the view of morality as a system of rules.

Dual models are premised on the belief that commonsense morality reflects attitudes that lie on a continuum between full-blown utilitarian and deontological ethics. Differences on this dimension are supposed to be explained in terms of a deep architectural divide (Bartels & Pizarro, 2011; Christensen & Gomila, 2012; Conway & Gawronski, 2013; Djeriouat & Trémolière, 2014; Holyoak & Powell, 2016; Lee & Gino, 2015). Controlled processes produce characteristically utilitarian responses; intuitive processes produce characteristically deontological responses (Bialek & De Neys, 2017; Cushman, Young, & Greene, 2010; Moll & de Oliveira-Souza, 2007).

Among proponents of these models, there is a standing debate as to how deep this divide is. Some have interpreted it as showing that rules need not be represented as inputs to the decision-making processes that culminate in moral judgment (see Blair, 1995; Greene et al., 2001; Haidt, 2001). Others believe that the rules are always represented, except that sometimes they are only tacitly represented. (Mallon & Nichols, 2010; Mikhail, 2011). Still others have proposed characterizing the divide in terms of model-based and model-free processing (Crockett 2013).

These discussions, however, operate within *the rule-based tradition* that we wish to challenge here. They concern the architectural or algorithmic processes underlying how moral rules are instantiated in judgments about specific actions. Precisely because of this, they leave untouched the more basic assumption regarding the subject matter of morality itself. After all,

the styles of thinking that are supposed to dominate commonsense morality are still conceptualized in terms of systems of moral rules: either expressions of duty or prescriptions to maximize or minimize some valued quantity.

In general, moral rules can be defined as functions from relevant inputs to behavioral imperatives. For example, a rule against murder is a general function from some action's being an instance of murder to an imperative against so acting. As Sidgwick (1981, p. 228) explains, "…rules of duty ought to admit of precise definition in a universal form." Under this definition, many generalizations (but not all of them) count as moral rules: the prohibition not to harm innocents, the injunction to maximize saved lives, etc.

Thus, it might be an open question whether the algorithmic processes that result in moral judgments take as inputs explicit representations of these functions. It is possible that not all moral decision-making is based on models shaped by a moral grammar. Be that as it may, the fact is that researchers working under this new paradigm still classify the *outputs* of these processes by conformity to the prescriptions of certain kinds of moral rules. The styles of thinking modeled by them, deontology and utilitarianism, which are supposedly characteristic of commonsense morality, are still styles of thinking in *accordance with some distinctive moral rules*.

It is not surprising, then, that contemporary work on the psychology of moral judgment continues to be dominated by moral dilemmas, in particular sacrificial dilemmas. Observing people's choices or their evaluation of the available options when each choice is made impermissible by deontological or utilitarian rules seems a natural way to measure people's

attitudes in terms of these moral frameworks. Whatever the underlying computational processes are, dilemmas seem recommended by the goal of understanding whether commonsense morality embodies deontological or utilitarian rules.

In sum, there is a long tradition, spanning from Piaget to contemporary dual models, that views morality as a rule-based system. Dilemmas appear useful for studying moral judgment in virtue of this underlying assumption. Hence, moving beyond dilemmas requires moving beyond the assumption that morality is a rule-based system. And moving beyond the assumption of morality as a rule-based system requires moving beyond the study of moral judgment in situations where rules conflict.

There are, as we shall see, numerous dimensions of the moral life that are not governed by rules, whether these refer to duties or codify maxims for the maximization of some valuable result. Hence, a science of morality that focuses only on dilemmas and rule-based judgments risks painting a picture of moral cognition that is overly narrow and stilted (Bauman et al., 2014). We want to challenge the use of dilemmas to study moral judgment because we reject the assumption that morality consists in a structured system of rules. Further, seeing how moral cognition operates beyond the rules provides a fresh perspective for the study of moral judgment. To this end, we examine in what follows moral judgment in cases of suberogatory behavior.

The Supererogatory and the Suberogatory

From time to time, people are faced with the option of doing more than what they are required to do, for example, spending some of their free time volunteering at the local animal shelter or donating a large portion of money to charity. These actions are admirable, despite the

fact that failing to do them does not seem to merit condemnation. Some classify these actions as 'supererogatory', going above and beyond the call of duty when failing to go above and beyond is perfectly morally permissible (Archer, 2018).

More controversially, people seem at times to underperform relative to some ideal in a way that is permissible, for instance, not offering to proctor the exam of a sick colleague despite the fact that one is available and the colleague has helped one in the past. Doing this tends to be regarded as bad despite the fact that there is no rule that requires one to pick up the duties of sick colleagues or that one no explicit agreement to help each other was made. Some use the label 'suberogatory' to describe this kind of behavior (see Driver, 1992; Hurd, 1998). The behavior is morally objectionable but there is no well-defined duty that it violates.

In failing to do a supererogatory action, one usually does not do anything bad. It's an admirable thing to donate money to charity, but failing to do so is not reprehensible. However, in some situations, failing to do a supererogatory action constitutes suberogatory behavior. If a tourist asks you for directions, you are completely within your rights to walk away without saying anything. Doing it, though permissible, is bad, whereas giving directions is good despite not being required. This possibility suggests a different kind of conflict that people encounter in their day-to-day experiences of morality: conflicts between equally permissible good and bad options. Here, we refer to these moral encounters that do not constitute real dilemmas as *morally charged situations*. We claim that these situations, along with the concepts used to evaluate them, offer a distinctive opportunity to study moral judgment.

Sacrificial dilemmas represent an interesting albeit limited subset of what people encounter in their everyday life. Using ecological momentary assessment, Hofmann, Wisneski, Brandt and Skitka (2014) asked a large sample of people to report whether they had witnessed, committed, or heard about a moral situation during the last hour, five times a day for three days. While some situations resemble the kind of conflict expressed in sacrificial dilemmas (e.g., "Reminded waitress I did not pay for my bill when she thought I did"), many of the situations reported seemed more similar to the situations depicted above. The choice between the competing options didn't seem a matter of aligning oneself with some well-defined rule.

Over-relying on dilemmas risks papering over these distinctions and, more generally, the differences between moral categories that inform varieties of moral judgment. O'Hara, Sinnott-Armstrong, and Sinnott-Armstrong (2010) compared responses to 15 dilemmas where they asked people to rate how wrong, forbidden, inappropriate and blameworthy an action was. They found that "the influence of wording variations on moral judgments was negligible" (p. 552) and they analyzed the small differences found as a matter of magnitude. Likewise, many researchers treat terms like 'forbidden' or 'blameworthy' as linguistic variations of some homogenous moral judgment (Bjorklund, 2003; Cushman, Young, & Hauser, 2006; Greene et al., 2001b; Koenigs et al., 2012). The assumption is that common sense moral judgment is not granular enough to reflect differences between being forbidden, blameworthy, bad, and so on. It is, instead, monolithic cognitive product, to which different labels provide different access points.

While there have been calls to more carefully use these measures (Christensen & Gomila, 2012; Monin, Pizarro, & Beer, 2007) it is unclear what the rationale for using different terms could be (cf. Cushman, 2008; Barbosa & Jiménez-Leal, 2017). But, once the repertoire of moral

encounters is expanded to include non-dilemma situations, it is possible that nuances and variability in moral cognition will emerge. In sum, by expanding the kind of moral encounters used when empirically probing people's intuitions and by enlarging the dimensions along which these encounters are assessed we can more adequately study the variation and granularity of moral judgment.

Here we present 6 experiments that study moral judgments in non-dilemma situations. One key feature of these situations is that the correct choice—if there is one—is not obviously settled by appealing to rules. Experiment 1 is an exploratory study that maps out the descriptions people offer of different situations. We find that people describe situations as suberogatory ('bad but permissible') or supererogatory ('good but not required') when supplying open descriptions of them. Experiment 2 shows that judgments of good/bad, permissible/impermissible, and optional/obligatory dissociate when evaluating suberogatory and supererogatory situations. We also find that people's beliefs about duties negligibly correlate with judgments of goodness, permissibility, obligation, and blame. In Experiment 3, we compare judgments in sacrificial dilemmas to see whether the same distinctions appear. We do not find the same dissociations, which suggests that eliciting these patterns of judgment requires more than giving participants the options to judge along various dimensions. In Experiment 4, we included a measure of praise/blame to see whether it correlates with the main "erogatory" measures. In Experiment 5, we replicated previous findings by using different vignettes that describe more characteristically moral situations adapted from classic philosophical thought experiments about abortion and property rights. We also measured whether beliefs about rights predict any kind of judgment. We find the same pattern of dissociations in these different vignettes and again find a negligible

correlation between judgments and individual beliefs about rights. In Experiment 6, we conducted a qualitative study to begin exploring the variety of factors that differentially drive different judgments. Using independent coders for a qualitative analysis, we find that people describe these situations using the language of character traits and rights rather than duties. In fact, in many cases people explicitly state the *absence* of duties to do anything in our scenarios

A methodological coda: Even though this research is mostly exploratory, its ideas are developed against a backdrop of well-established findings in moral psychology. We decided to preregister Experiments 2 to 6 because we believe that clearly establishing design and analysis plans can help distinguish the confirmatory and exploratory aspects of our research by clearly specifying our intent. The procedure, therefore, reduces needless post hoc interpretations (Nosek et al, 2019).[2] Materials, data, preregistrations and code for all experiments are available on the OSF page of the project (https://osf.io/kja2u/?view_only=573c1541bbd240b397b79229704dfae7). The IRB of the University (blinded for review) approved this study.

[2] The only important deviation from preregistration plans occurred in Experiment 2, where the main statistical analysis proposed (a repeated measures ANOVA) was replaced by a mixed linear model, since it is better suited to model our data.

Experiment 1

The objective of this first study is exploratory. We presented participants with vignettes describing either suberogatory or supererogatory behavior. They were instructed to select words from a list to describe these scenarios and to offer a description in their own words. The goal of this is twofold. The first is to see what language people use to spontaneously describe a moral encounter. The second is to see whether people recognize a distinction between different judgment categories that maps onto the complex category of suberogatory. This requires that people have distinct concepts of permissibility and goodness such that they can describe some behavior as bad but permissible. Hence, we decided to run an exploratory study using word selection and open response to see whether participants utilize the moral categories we aim to study without being prompted.

We expected people to always select always more than one word (e.g., "good") and to give descriptions that characterize both the action and the person.

Method

Participants

95 participants (60 women and 35 men, mean age = 32.42, SD =10.43), based in the United States and recruited through Prolific Academic, took part in the study in exchange for 40 pence. Participants were aware that their answers would be anonymous and were monetarily compensated for their participation. The average completion time was 5.2 minutes.

Materials and procedure

We constructed four scenarios, some of which were based on thought experiments by Driver (1992). Each scenario described an individual faced with a choice between a suberogatory and a supererogatory option. Additionally, in order to account for possible asymmetries between actions and omissions (Haidt & Baron 1996), we created two versions of each scenario that describe either an action or an omission. This generated eight vignettes, described below (suberogatory versions of these vignettes are in brackets):

Newlyweds Two newlyweds are boarding a plane to go on their honeymoon. Because of a booking error by the airline, the couple does not have seats together. They ask someone, already seated, if they would switch seats so the couple could sit together. The passenger switches seats, and the newlyweds can sit together. [*The passenger does not switch seats, and the newlyweds have to sit separately*.]

Kidney Alex is suffering from severe kidney failure and Alex's only hope is to obtain a transplanted kidney. Alex's cousin, Jamie, is the only known compatible donor. Jamie offers to donate the kidney to Alex. [*Jamie does not offer to donate the kidney to Alex*.]

Mowing Early one Sunday morning when the neighbors are usually sleeping, Sam notices that the lawn needs to be mowed. Although it is his property and it would be inconvenient to do it later, he decides to not mow the lawn. He knows that starting the lawn mower will probably wake up the neighbors. [*Even though he knows that starting the lawn mower will probably wake up the neighbors, he does it anyway. It's his property and it will be inconvenient to mow the lawn later.*]

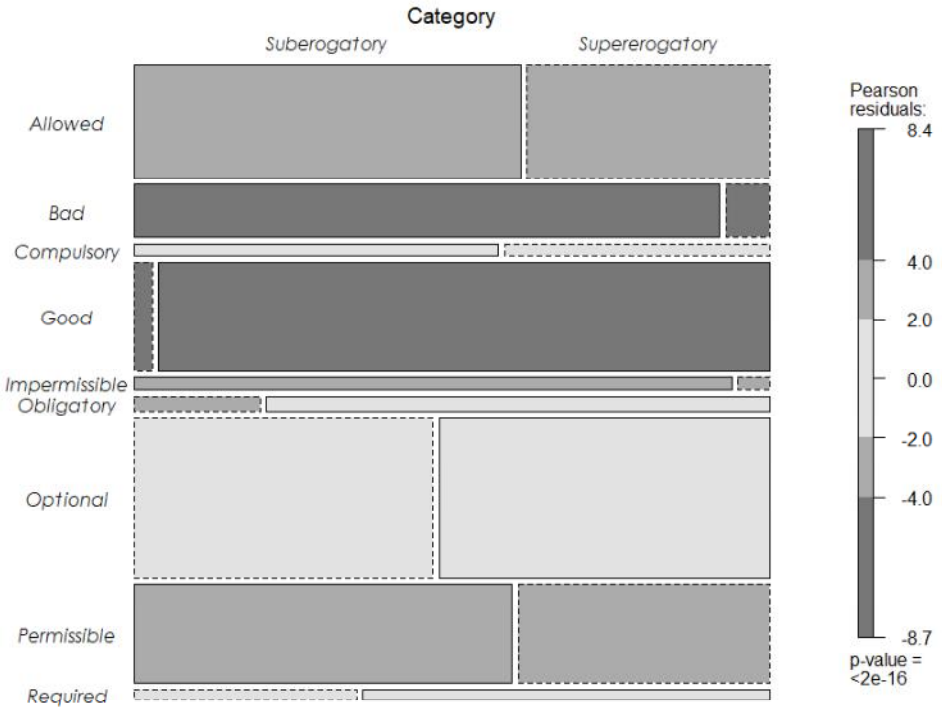
Raffle During the Christmas party, the secretary publicly announced the results of the office raffle: "Congratulations to Alex, who has won the trip for two to Disney World. She can come up front to claim her prize or she can let a cash equivalent go to a hurricane relief fund". After hearing the news, Alex looked excited: ". Even though I have the winning ticket and Disney World sounds fun, I am going to donate the prize to one of the

charities." [*After hearing the news, Alex looked excited: "I have the winning ticket! Even though I don't really care much about Disney World, I am going to claim the prize anyway".*]

Participants were presented with four out of the eight possible variations, so each participant was presented with a suberogatory omission, a suberogatory action, a supererogatory omission, and a supererogatory action. After reading the vignettes, participants completed a word selection task by selecting "the word(s) that you think best describe the situation." The word choices were: "permissible", "impermissible", "required", "good", "obligatory", "allowed", "bad", "optional" and "compulsory". Participants also completed an open description task by offering a description of the situation in their own words. The vignette presentation order was counterbalanced and the order of tasks and words was randomized.

Results

For the word selection task, participants selected 2.6 words on average (see Figure 1). For suberogatory behaviors, people most often chose the words "optional" (27%) and "allowed" (25%), followed by "permissible" (21%). For supererogatory behaviors, "good" (35%), "optional" (28%), and "allowed" (14%) were the most common choices. "Allowed" and "optional" are the most common pair of words used across all vignettes. "Allowed" and "permissible" are more strongly associated with evaluating suberogatory behaviors. Figure 1 summarizes these results.



281

282 *Figure 1. Mosaic plot for words selection*. Height is associated with overall selection frequency
283 and width is associated with prevalence for suberogatory and superogatory behaviors (Hornik,
284 Zeileis & Meyer, 2006). Darker shades and solid lines represent positive associations, whereas
285 lighter shades and dotted lines represent a negative association within the sub/supererogatory
286 categories. The plot represents a model contrasting observed and expected frequencies of word
287 choices.

288

In the open descriptions, participants predominantly used character trait descriptors, such as *selfish* for the suberogatory vignettes and *thoughtful* for the supererogatory situation (see Figure 2). They rarely used words like “permissible”. Also, more nuanced descriptions of the suberogatory situations generally highlighted the optionality of the response, in line with the abstract concept of the suberogatory: “he is allowed to do it” (mowing), “Jamie has a right to her own decision” (Kidney), “it is understandable” (kidney). Lastly, people were generally sensitive to the different moral aspects that structure the situations depicted as morally charged. They were, for instance, quick to describe the action/person negatively (mean, rude, etc), while also recognizing that rights and expectations were at play in the scenarios evaluated. They also recognized the possibility that some behaviors could be bad but permissible.:

- “It is a little bit selfish, but then again she has the right to keep her organs”
- "While it was mean, the passenger has the right to refuse
- “Rude, but it is his lawn”
- “Giving an organ is a big thing to ask. It is something that is optional, and there is no mention of Alex asking him to do it so he is not required to offer”

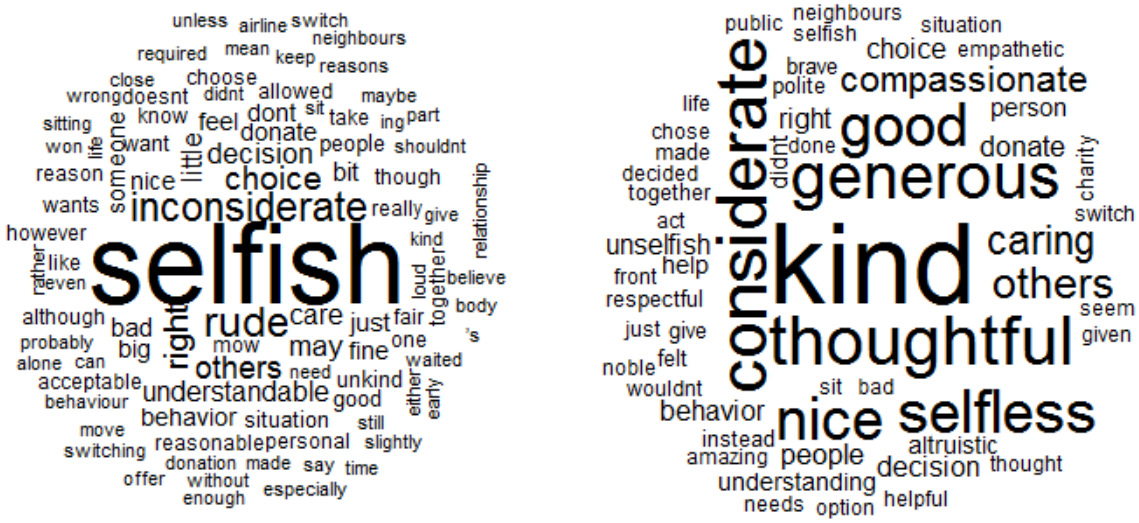


Figure 2. Wordcloud for suberogatory (left) and supererogatory (right) situations

While these results are exploratory, they provide evidence that moral thinking about the depicted situations is nuanced in ways that seem in line with the categories of suberogatory and supererogatory and more complex that simple good/bad, obligatory/impermissible evaluations. These findings made possible the use of the measures and procedures described in the subsequent experiments.

Experiment 2

In Experiment 1, we found some evidence that people describe certain moral situations in ways that are sensitive to the categories of the suberogatory and the supererogatory. This suggests that people distinguish between permissibility, obligation, and goodness. Here we test for quantifiable differences between these evaluative categories.

We asked people to rate the sub– and supererogatory situations along three dimensions: good/bad, permissible/impermissible, and obligatory/optional. We manipulated the Erogation category (within-subjects, sub– and supererogatory) and the Situation Type (between-subjects, action or omission). We hypothesized that people would judge suberogatory situations as worse than supererogatory situations, but that permissibility and obligatoriness ratings between the two situations would not be significantly different. We also expected permissibility ratings to be significantly higher than obligatoriness ratings in both supererogatory and suberogatory conditions, though the difference between permissibility and obligation would be greater in the suberogatory condition than in the supererogatory condition. We did not expect any differences between judging actions and omissions. We also collected data on attitudes towards duties, expecting that ratings along the Erogation category would be associated with these attitudes.

Participants

We ran a power analysis for a mixed ANOVA (between-within interaction), assuming an effect size of $f = 0.15$ using the software G*Power. This analysis suggested a sample size of 272 for a 0.95 power. To account for exclusions, we recruited 311 participants (186 women and 125 men, mean age = 32.77, SD =11.18) through Prolific Academic. We decided to switch to an alternative data strategy after collecting the data, using mixed linear models, given the problems of repeated measures analyses with independence and distributional assumptions (Singmann & Kellen, 2019). Our sample size, however, is consistent with a power of 0.9, assuming participants can be treated as a random factor to account for within-person response variability, with a mixed design (Singmann & Kellen, 2019; Westfall, 2015; See Supplementary materials for details).

Each person voluntarily participated in the study and received 38 pence as compensation. The average completion time was 3.11 minutes.

Materials and procedure

Materials were the same as Experiment 1. Each participant saw four of eight scenarios. We manipulated the moral category within participants and action/omission between participants, so each participant saw two supererogatory situations and two suberogatory situations, where all of them were either actions or omissions. For each vignette, participants evaluated the situation along different dimensions with a 100-point sliding scale. The dimensions included degree of permissibility (impermissible = 0, neither impermissible nor permissible = 50, permissible =100), degree of goodness (bad = 0, neither good nor bad = 50, good = 100), and degree of obligation (optional=0, neither optional nor obligatory = 50, obligatory = 100). The dimension order was randomized across trials. The scale appeared only with the anchors, the slider was always placed in the center of the scale, and participants were not given a numerical representation of where they placed the slider.

Participants viewed one vignette at a time and the sliders were placed on the same page as the vignette. After completing the study, participants indicated their age, gender, and political orientation (on a five-point Likert scale from *very liberal* to *very conservative*). We also collected information about personal sense of duty. Participants indicated agreement with a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree) with the idea that there are duties to respect your neighbors, to help anybody who needs help, and to help one's family members.

Results

Pre-registered analyses were integrated into a set of linear random effect models fitted with the lme4 R library (Bates, Mächler, Bolker, & Walker, 2015; R Core Team, 2019), with participant as a random intercept.[3] Pairwise comparisons were carried out using the emmeans package (Lenth, 2020) which allows degrees of freedom to be calculated with the Kenward Roger method and p values to be adjusted with the Tukey method. Confidence intervals for non-standardized simple differences are reported for ease of understanding.

Results are summarized in Table 1 and Figures 3 and 4. We fitted four nested models to contrast the interactive effect of the manipulated variables (see models 1 and 2) and to test the effect of differences between scenarios and endorsement of norms (models 3 and 4). Goodness of fit indicators (AIC, BIC and deviance) and the chi square test ($\chi^2(2)=393.2$, $p<.0.001$) favor selection of model 2. Significant interactions between judgment type and condition suggest differences both between and within the erogation category. Supererogatory behaviors were judged as better and more permissible ($M = 86.4$, $SD = 18.7$, $n = 622$ and $M = 82.0$, $SD = 20.6$, $n = 622$) than suberogatory behaviors ($M = 36.6$ $SD = 23.3$, $n = 622$ $M = 62.2$ $SD = 29.7$ $n = 622$) t-ratio Good: $t(3411) = -34.89$, $p < .001$, $M_{diff} = -49.85$, 95% CI [-52.2, -47.5] and t-ratio Permissible: $t(3411) = -19.7$, $p < .001$, $M_{diff} = -19.78$, 95% CI [-22.6, -17.0], regardless of whether

[3] It can be argued that the data of this experiment could be considered a cross-classified data set. However, the items in each situation type are not completely equivalent, which makes the corresponding items nested within situation type. Additional models were fitted with additional random term but since results are equivalent, we restrict the presentation here to the different conditions as fixed effects.

they were actions or omissions. Supererogatory behaviors were judged to be marginally more obligatory (i.e., less optional) (M = 30.9, SD = 32.7, n = 622) than the corresponding suberogatory responses (M = 23.5, SD = 26.3, n = 622) t-ratio: $t(3411) = -7.47$, $p < .001$, $M_{diff} = -7.47$, 95% CI [-10.8, -4.19] though it is clearly a smaller effect.

Interestingly, the size of the differences between permissibility and goodness ratings are vastly different when looking at sub and supererogatory responses. Within the suberogatory responses, this difference amounts to 25.6 points (Cohen's d = 0.96) while for the supererogatory category, this difference is only 4.46 points (Cohen's d = 0.22). That is, goodness and permissibility judgements are very similar within the supererogatory condition, but not for the suberogatory responses, where they are more clearly tracking different aspects of the situation.

386 Table 1. Summary of Models fitted for Experiment 2

	Model 1	**Model 2**	**Model 3**	**Model 4**
Omission	-2.74***	-6.411***	-7.33***	-7.26***
	(-4.65, -0.83)	(-10.500, -2.322)	(-10.41, -4.24)	(-10.36, -4.17)
OBLIGATORY	-34.30***	-17.449***	-38.12***	-38.12***
	(-36.41, -32.20)	(-21.389, -13.509)	(-41.07, -35.17)	(-41.07, -35.17)
PERMISSIBLE	10.57***	21.758***	7.58***	7.57***
	(8.47, 12.67)	(17.818, 25.698)	(4.63, 10.53)	(4.62, 10.52)
SUPER EROGATORY	25.70***	50.771***	25.70***	25.70***
	(23.98, 27.41)	(46.830, 54.711)	(23.99, 27.41)	(23.98, 27.41)
Mowing	-7.48***		-7.48***	-7.47***
	(-9.90, -5.05)		(-9.89, -5.06)	(-9.89, -5.05)
Newlyweds	-3.97***		-3.97***	-3.96***
	(-6.39, -1.54)		(-6.39, -1.55)	(-6.39, -1.54)
Raffle	-0.15		-0.16	-0.15
	(-2.58, 2.26)		(-2.58, 2.26)	(-2.57, 2.26)
DUTY TO FAMILY				-0.21
				(-0.99, 0.56)
DUTY TO HELP				-0.01
				(-0.79, 0.77)
DUTY TO NEIGHBORS				-0.31
				(-1.35, 0.71)
Omission: OBLIGATORY		8.75***	7.72***	7.718***

		(3.15, 14.35)	(3.53, 11.91)	(3.526, 11.91)
Omission: PERMISSIBLE		7.75***	6.04***	6.042***
		(2.15, 13.35)	(1.85, 10.23)	(1.85, 10.23)
Omission: SUPER EROGATORY		-1.84		
		(-7.44, 3.76)		
OBLIGATORY: SUPER EROGATORY		-41.34***		
		(-46.92, -35.77)		
PERMISSIBLE: SUPER EROGATORY		-28.36***		
		(-33.93, -22.78)		
Omission: OBLIGATORY: SUPER EROGATORY		-2.07		
		(-9.99, 5.84)		
Omission: PERMISSIBLE: SUPER EROGATORY		-3.419		
		(-11.34, 4.50)		
Constant	52.90***	39.74***	55.17***	58.14***
	(50.41, 55.40)	(36.86, 42.62)	(52.41, 57.94)	(51.41, 64.87)
N	3732	3732	3732	3732
Log Likelihood	-17580.00	-17367.00	-17570.00	-17569.00
AIC	35180.00	34763.00	35163.00	35168.00
BIC	35242.00	34850.00	35238.00	35261.00

***p < .01; **p < .05; *p < .1

387

Our results do, however, show that people strongly distinguish between goodness, permissibility, and obligation. Participants rated supererogatory behaviors as good, optional, and permissible; they rated suberogatory behaviors as bad, optional, and permissible. Despite both being rated permissible, supererogatory behaviors were rated as more permissible than suberogatory behaviors.

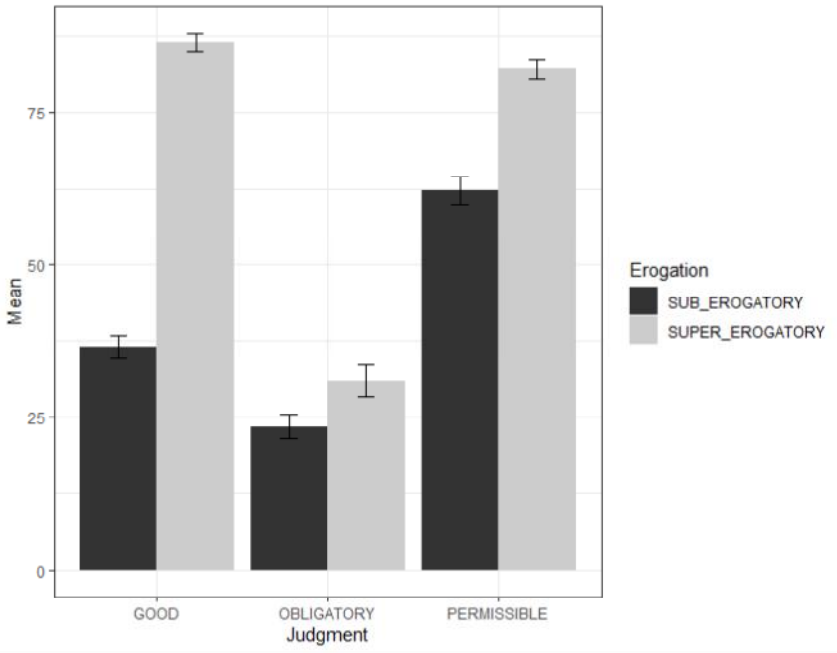


Figure 3. Scores by type of judgment. Error bars represent 95% confidence intervals.

We found a smaller interaction between judgment type and the action/omission dimension across erogation conditions. Actions were rated as better (M = 65.1, SD = 32.7, n = 628) than omissions (M =57.8, SD = 32.3, n = 616) t-ratio: $t(1539) = 4.82$, $p = .001$, $M_{diff} = 7.33$, 95% CI [3.71, 10.94], but there were no differences between permissibility and obligatory ratings for actions and omissions. The three-way interaction is not explored here but overall, there are no

important differences between erogation categories across the action/omission condition except for the goodness judgment, where actions are judged as better than omissions.

We explored the association between ratings for each type of judgment and participant's endorsement of statements about personal duties. We did not find any consistent association between judgment scores and responses pertaining to personal duties (see Table 2). There are significant correlations between endorsing different statements of personal duty (ranging from 0.13 to 0.34) but most correlations between sense of personal duty and different judgment categories were negligible (from -.01 to 0.07) and non-significant. The one exception is that beliefs about duties to help others significantly correlated with judgments of permissibility, but the correlation is very small.

Table 2. Bivariate correlations between personal norms and judgment scores

	Permissible	Good	Obligatory	Duty to neighbors	Duty to help
Permissible					
Good	0.55***				
Obligatory	-0.24**	-0.22***			
Duty to neighbors	-0.03	0.0	-0.02		
Duty to help	-0.07**	-0.02	0.07	0.33***	
Duty to family	-0.03	-0.02	0.01	0.30***	0.23***

***p < .01; **p < .05; *p < .1

Breaking down responses by scenario reveals some variability across vignettes (see Figure 4). For example, while donating a kidney to a cousin is judged to be better and more permissible than not donating a kidney, the same pattern does not hold in the raffle scenario. In this case, both situations are equally permissible, but donating the raffle prize is better than not. Notice however

that these small differences do not account for significant variance, according to the model fitting presented in Table 1.

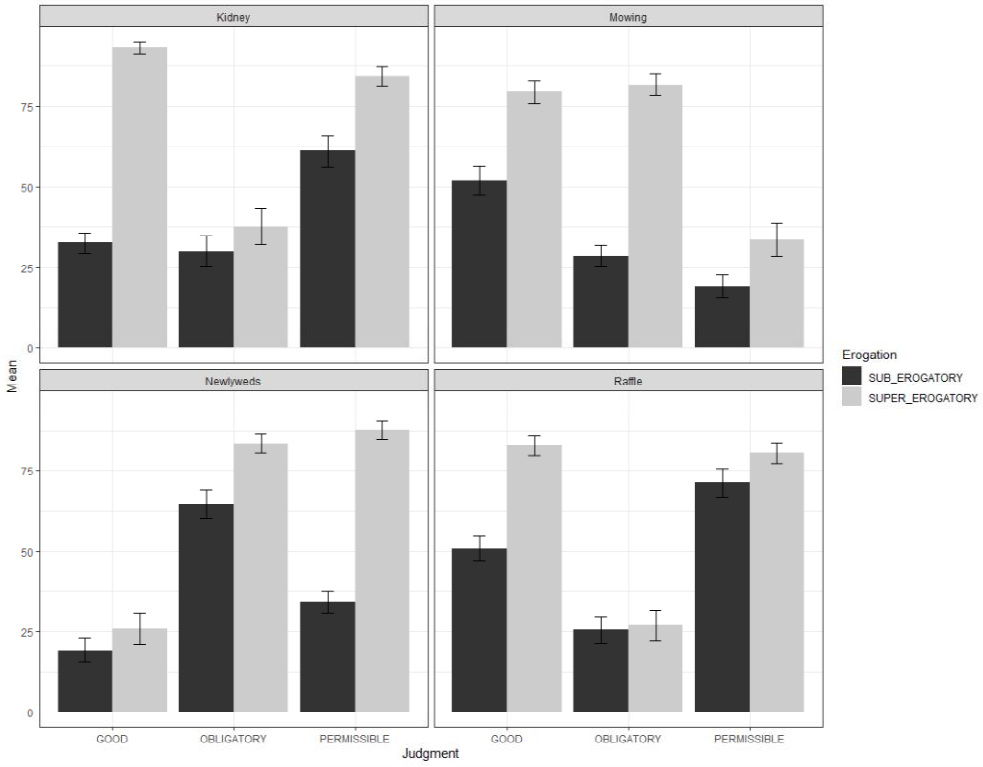


Figure 4. Scores by type of judgment and scenario. Error bars represent 95% confidence intervals

Discussion

These results reveal quantifiable differences between different evaluative categories that are employed in moral judgment. Similar to our results in Experiment 1, we found that judgments of permissibility, obligation, and goodness dissociate when people make judgments of suberogatory behavior. This shows that including additional measures allows variability in moral judgment to emerge. Moreover, beliefs about personal duties were negligibly correlated with different kinds of judgments. This suggests that people's responses to these situations are not indicative of an

underlying adherence to rules or generalized statements about what one ought to do, even though the duties we asked about specifically applied to the scenarios in the study.

We mentioned in the Introduction that the use of dilemmas to study moral judgment reinforces assumptions about the way to measure moral judgment. When assessing the relationships between different measures, researchers have found negligible differences between different judgment categories. We see now that including different measures of moral judgment allows for variability, challenging a central assumption about the relationship between different measures of moral judgment. This raises a question: is the lack of variability between judgment categories a function of the measures used to study moral judgment, or the dilemmas commonly used to elicit judgments? It might be the case that some measurement devices function as self-reinforcing demand effects. Alternatively, there might be something about dilemmas themselves that make a difference to moral judgment independently of the measures used. To begin answering these questions, we applied our measures to sacrificial dilemmas in Experiment 3 to see how they align with the results of this experiment.

Experiment 3

In this experiment, we consider different kinds of judgments in cases of sacrificial dilemmas. In our previous experiments, we found dissociations between different kinds of judgments. However, the situations used to test these differences differ significantly from sacrificial dilemmas, the stimuli most consistently used to elicit moral judgments. First, harm to another agent is unavoidable in a sacrificial dilemma, so there is no straightforward possibility of supererogatory behavior. Second, despite being unrealistic, they elicit distinctive emotional

448 reactions (Christensen et al., 2014) as measured by arousal and valence. Third, dilemmas are
449 situations where explicit rules come into conflict (e.g., prohibitions against causing harm and
450 prohibitions against allowing easily preventable harm).

451 We chose sacrificial dilemmas that have been widely used in empirical studies of moral
452 judgment. Based on an analysis from Christensen et al. (2014), we selected dilemmas that
453 produce the greatest variation in responses (varying in the use of personal force and the
454 inevitability of harm). We also excluded dilemmas where various interactive effects might
455 plausibly drive moral judgments (e.g., when causing harm is self-beneficial). Characteristics and
456 full text of the dilemmas selected are presented in the supplemental materials.

457 Sacrificial dilemmas are sometimes thought to bring out the contrasts between
458 deontological and utilitarian ethical intuitions, because each set of intuitions usually recommends
459 different behaviors in the face of a sacrificial dilemma. Hence, intuitions about what is right or
460 appropriate indicate alignment with one or the other theory. This suggests that people committed
461 (implicitly or explicitly) to either utilitarianism or deontology might produce different moral
462 judgments in reaction to sacrificial dilemmas. This presents a challenge. Because we are
463 measuring for dissociations among different judgment categories, and different normative
464 theories make different recommendations for navigating dilemmas, it is possible that each side
465 will cancel the other out, thereby giving the appearance of similarity between judgment
466 categories. To ensure that people with utilitarian tendencies do not cancel out people with
467 deontological tendencies, we used the Oxford Utilitarianism Scale (Kahane et al., 2018) to assess
468 trait utilitarianism across two dimensions of utilitarianism, impartial beneficence and
469 instrumental harm.

The character of this study is primarily exploratory but we expected to replicate some of the main findings in this literature and some of the patterns observed by Christensen et al. (2014). We expected utilitarian responses in personal dilemmas to be associated with lower ratings of goodness and permissibility relative to impersonal dilemmas (Greene et al. (2001b). We also expected behaviors that cause unavoidable deaths to be judged as better than situations where causing death was avoidable (Christensen et al., 2014).

Method

Participants

Given that our objective was to first to replicate the observed effects with sacrificial dilemmas and, second, to explore the impact of these variables on our new measures, we ran power analysis for a linear mixed model treating participants as random effects, using the effect size reported by Christensen et al (2014) for personal force of r = .75 (equivalent to a d = 2.2). This is the smallest of the effect sizes considered. This analysis suggested a sample of 60 participants for our mixed design for a power of .99 (Westfall, 2015). Since we wanted to make sure we would be able to observe differences in our new dependent variables, we aimed to collect data for 300 participants which would allow us to observe a significantly smaller effect (r = .17, d = .3). 304 participants (153 women and 151 men, mean age = 33.21, SD =11.62), recruited through Prolific Academic, took part in the study in exchange for 0.50 pence. The average completion time was of 7.21 minutes and none of the participants had taken part in our previous experiments.

490 **Materials and procedure**

491 Participants saw the four dilemmas and were randomly presented with the version where
492 the main character decides to cause harm (utilitarian response) or allow harm (deontological
493 response). That is, utilitarian/deontological response was a between-subject factors while
494 avoidability of the result (avoidable/unavoidable) and personal force (personal/impersonal) were
495 within subject variables. For each scenario, participants judged the main character's response
496 along dimensions of permissibility, obligation, goodness, and blameworthiness using the same
497 sliders as before. After judging the dilemmas, participants were presented with the items of the
498 *Oxford Utilitarianism Scale* (OUS) in a random order. Finally, they are asked some basic
499 demographic questions and were asked to rate their experience with this kind of dilemmas in a
500 scale from 1 (not at all familiar) to 5 (extremely familiar).

501 **Results**

502 Results are summarized in Table 3 and Figures 5 and 6. As before, we fitted a set of
503 random effect models with judgment type, personal/impersonal utilitarian/deontological,
504 avoidable/unavoidable, having experience with dilemmas and the score in the OUS as fixed
505 predictors allowing for a random effect of participant. Pairwise comparisons were calculated with
506 the same parameters specified for Experiment 2. The OUS had good reliability for its both
507 subscales (Instrumental Harm Cronbachs' $\alpha = .65$ and Impartial Beneficence Cronbachs' $\alpha =$
508 .72). The models differed in the specification of the interaction of the fixed effects. The best
509 fitting model has interactive effects for two-way interactions for all terms with the
510 personal/impersonal dimension. This model is presented in Table 3.

511 Table 3. Best fitting model for Experiment 3.

	score
BLAME	-5.15**
	(-9.53, -0.77)
OBLIGATORY	-2.98
	(-7.36, 1.40)
PERMISSIBLE	12.78***
	(8.40, 17.16)
PERSONAL	9.22***
	(5.40, 13.04)
UTILITARIAN	2.46
	(-1.03, 5.95)
UNAVOIDABLE	2.90*
	(-0.55, 6.35)
BLAME: PERSONAL	4.49**
	(0.13, 8.85)
OBLIGATORY: PERSONAL	-4.91**
	(-9.27, -0.55)
PERMISSIBLE: PERSONAL	-4.95**
	(-9.31, -0.59)
PERSONAL: UTILITARIAN	-14.00***
	(-17.19, -10.82)
BLAME: UNAVOIDABLE	0.23
	(-4.13, 4.59)
OBLIGATORY: UNAVOIDABLE	-1.82
	(-6.18, 2.54)
PERMISSIBLE: UNAVOIDABLE	-4.94**
	(-9.30, -0.58)
PERSONAL: UNAVOIDABLE	-4.89***
	(-7.98, -1.81)
BLAME: UTILITARIAN	6.82***
	(2.46, 11.19)
OBLIGATORY: UTILITARIAN	1.50
	(-2.86, 5.86)
PERMISSIBLE: UTILITARIAN	-2.37
	(-6.74, 1.99)
Constant	47.87***

	(44.53, 51.21)
N	4864
Log Likelihood	-23017.00
AIC	46074.00
BIC	46204.00

$^{***}p < .01$; $^{**}p < .05$; $^{*}p < .1$

There are significant main effects of judgment type and the utilitarian status of the response judged (Deontological, $M = 52.39$, $SD = 28.3$ and Utilitarian $M = 49.18$, $SD = 28.4$, , $t(3924) = 3.77$, $p < .001$, $M_{diff} = 3.21$, 95% CI [1.61, 4.80]), but not of its being personal (Personal $M = 51.64$, $SD = 25.8$, and Impersonal $M = 49.87$, $SD = 30.8$, $n = 2432$, $t(3411) = 1.99$, $p = .046$, $M_{diff} = 1.77$, 95% CI [0.17, 3.37]) or avoidable (Avoidable $M = 51.46$, $SD = 25.01$, and unavoidable $M = 50.04$, $SD = 31.4$, $n = 2432$, $t(3411) = 1.49$, $p = .13$, $M_{diff} = 1.42$, 95% CI [-0.17, 3.02]). These variables only have interactive effects, which shows that there is not an overall effect of the manipulation across all judgment dimensions.

Figure 5 shows how personal deontological responses are judged as better ($M = 61.75$, $SD = 28.89$, $n = 303$), considerably more permissible ($M = 65.76$, $SD = 28.56$, $n = 303$), more obligatory ($M = 51.38$, $SD = 33.94$, $n = 303$) and less blameworthy ($M = 42.31$, $SD = 27.59$, $n = 303$) than the corresponding utilitarian responses (Good $M = 39.19$, $SD = 28.12$, $n = 305$, Permissible $M = 43.54$, $SD = 31.06$, $n = 305$, Obligatory $M = 36.53$, $SD = 31.23$, $n = 305$, and blame $M = 60.01$, $SD = 27.22$, $n = 305$) but only for the unavoidable outcomes (all pairwise comparison significant at 0.01). When the outcome is avoidable, the pattern is the opposite for both impersonal and personal responses. The main effect of personal contact consists in making the deontological response more acceptable (less blameworthy, more obligatory, better and more permissible) than the corresponding utilitarian ratings (See lower panel, Figure 5). Overall this

picture is consistent with prior findings where the effect of these variables on dilemma responses is conditional on several factors (See Christensen et al. (2014)).

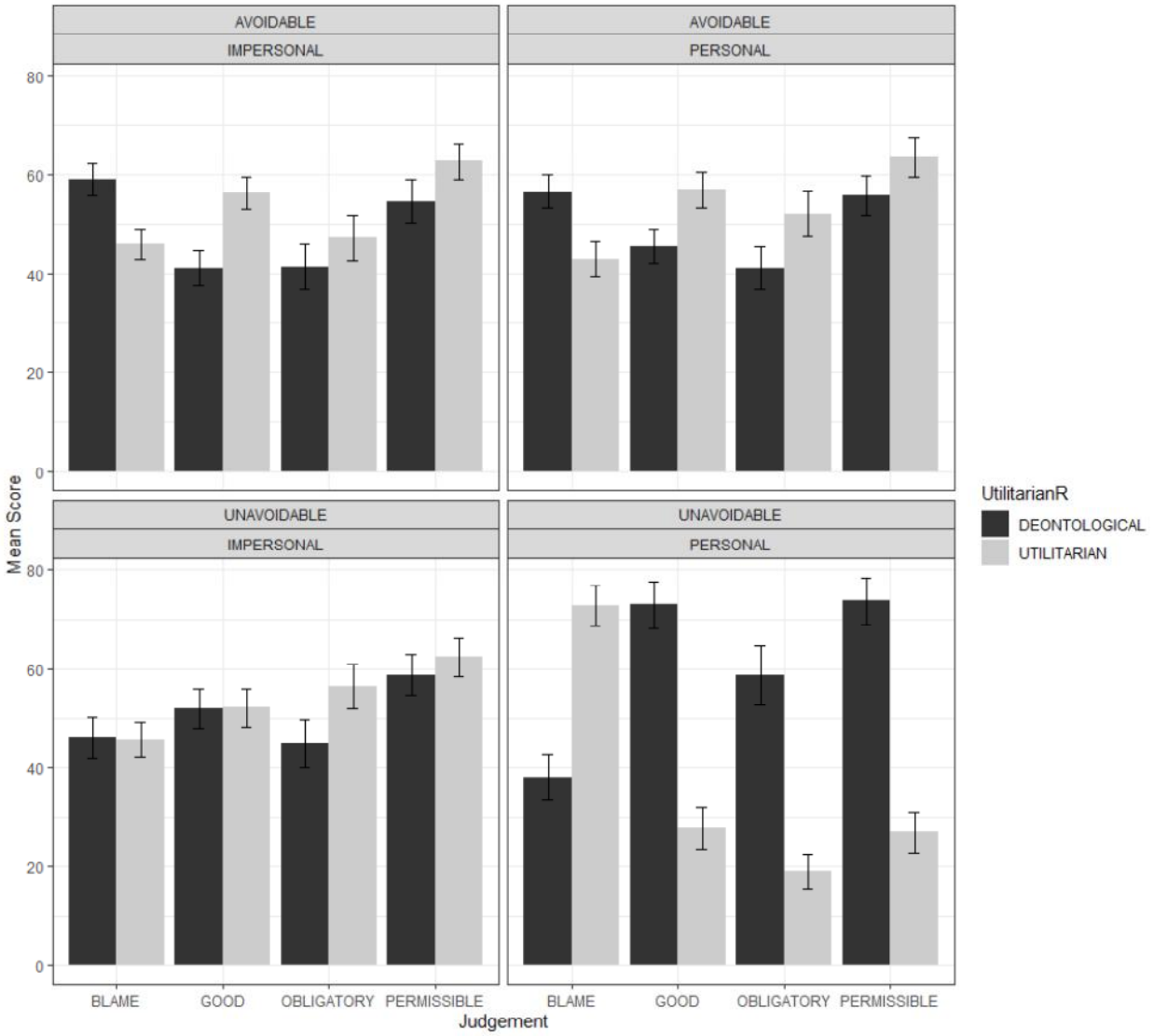


Figure 5. Mean ratings for each judgment type by utilitarian status, personal force and avoidable. Error bars represent 95% confidence intervals.

In the model fitting process, neither experience with dilemmas nor the overall score of the OUS (or its subscales) resulted in significant predictors of participant judgments. Moreover, for all but one dilemma type (Unavoidable, Personal) we did not find significant differences between judgments across trait utilitarian tendencies.

We wanted to see, however, whether more fine-grained distinctions in trait utilitarianism might allow unnoticed dissociations to emerge. To do this, we assigned participants to a low, medium and high utilitarianism group, by splitting participants aggregate scores on the OUS by the 33rd and 66th percentile. Zooming in on participants with the lowest and highest utilitarian scores (See Figure 6), we see that participants in the Low Utilitarianism Group (left panel) clearly judge causing harm as less permissible, less optional, and more blameworthy than allowing harm. People in the High Utilitarianism Group (right panel) show an opposite pattern, albeit not as distinctive (See Supplemental materials for more information on these comparisons). Note, however, that these patterns only emerge after making exploratory, post-hoc data analyses. Even doing this, which might otherwise be considered methodologically problematic, we weren't able to see the clearly marked dissociations observed in Experiment 2.

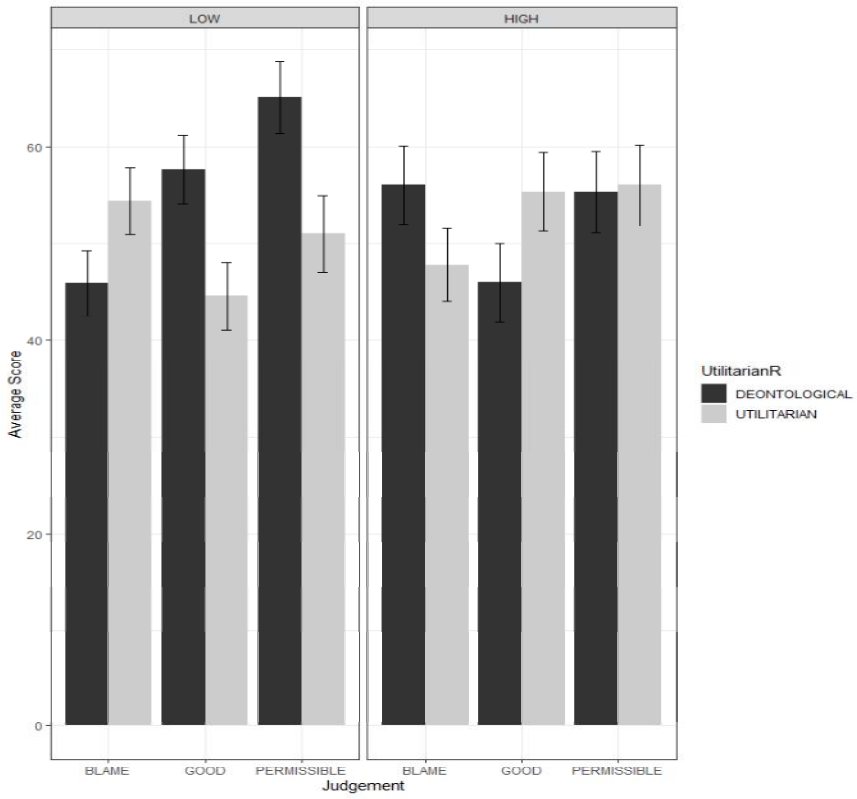


Figure 6. Mean ratings by of judgment, grouped by Utilitarianism level (panels) and Utilitarian response (bars)

Permissibility judgments tend to be higher than goodness judgments, suggesting that even though they are related, their difference in magnitude suggests that these judgments are tracking different features of the actions depicted in the scenario. This difference is larger and statistically significant for participants with low utilitarian tendencies, who consider harm relatively more permissible (M = 50.95, SD = 30.51) than it is good (M = 44.55, SD = 27.23, , t(448) =-2.36, p -=.02, M_{diff} =-6.40, 95% CI [-11.72, -1.08]). With participants who score in the higher end of utilitarianism, there is virtually no difference between these judgments (permissibility M = 56.02, SD = 27.52 vs obligatoriness M = 55.35, SD = 26.71, , t(338) =-0.23, p =.81, M_{diff} =-0.66, 95% CI [-5.45, 5.12]). A similar pattern is found with the obligatoriness levels: participants who score

low in utilitarianism consider causing harm optional (as revealed by a one-sample t test against the indifference point $M = 37.09$, $SD = 31.02$, $t(227) = -6.79$, $p < .001$, 95% CI [33.98, 41.10]) and participants who score high in utilitarianism consider *not* causing harm as optional ($M = 44.20$, $SD = 29.81$, $t(169) = -0.69$, $p = .48$, 95% CI [43.70, 73.0]). These tendencies are statistically significant when explored with a fixed effects model with these terms (see Supplementary materials). Bear in mind that this model is exploring a tendency rather than testing a hypothesis. As such, it lacks confirmatory value.

Discussion

We wanted to see whether adding additional measures allowed variability in moral judgment to emerge, even when using dilemmas as stimuli. However, when we had people make judgments about dilemmas, the variability we observed in Experiment 2 nearly disappears. This is not an effect driven by Utilitarian members of the participant pool. Even when we separate people according to their utilitarian tendencies (which tendencies pull *against* making the distinctions outlined in Experiment 2), we do not find strong dissociations between the categories of goodness, permissibility, and obligation. The dissociations are weak enough that they might seemingly justify the claim that **"**the influence of wording variations on moral judgments was negligible" (O'Hara et al., 2010).

Our results suggest that changes in measurement alone are not sufficient to indicate the underlying complexity of moral judgment. The kind of situation being evaluated makes a difference to moral judgment. When we limit ourselves to using only dilemmas, we generate an overly narrow view of moral thinking. Worse, it hinders the ability to make useful generalizations about the operations of moral cognition in general.

The lesson is that we need more than different measures. The situation being evaluated appears to have an influence on the shape of moral judgment. Our materials in Experiment 2 provide a first step toward expanding the range of situations used as stimuli to study moral judgment. In our next two experiments, we wanted to introduce different measures and scenarios, as well as explore potential connections between moral judgments and generalized rules.

Experiment 4

In Experiment 3 we found that judgment dimensions do not behave uniformly when used to judge responses in sacrificial dilemmas. That is, utilitarian and deontological responses are judged to be permissible, optional, blameworthy and good at different levels, depending on variables known to impact participant responses (Christensen et al. (2014)). However, unlike the pattern observed in Experiment 2, differences between categories were generally non-significant. Moreover, significant differences tended to be small, to arise from interactions, and to be driven mainly by individual differences in trait utilitarianism. This shows that sacrificial dilemmas do not elicit differences in judgment that are found when using different erogation conditions. Conversely, this suggests that not all situations of moral conflict could be assessed in terms of the dimensions proposed.

Additionally, in Experiment 3, participants made judgments of blameworthiness for the response of the person allowing, or causing harm. This marks an important difference with the procedure used in Experiment 2. In Experiment 2, we found that individuals rate scenarios in a way that reflects the categories of suberogatory and supererogatory. However, we did not find that individuals rate suberogatory and supererogatory behaviors as equally permissible and optional, as we had initially hypothesized, despite finding a difference in judgments of goodness.

One explanation of this failure might be that people disapprove of suberogatory behaviors, but the various judgment types (permissibility, obligation, and goodness) do not have any obvious place for people to register this disapproval. Hence, disapproval might be skewing judgments in a way that drives the differences between permissibility and obligatoriness across scenarios. To correct for this, and to replicate the findings of Experiment 2, we ran the experiment again with an additional judgment type of blameworthiness.

Participants

We recruited 311 participants (166 women and 145 men, mean age = 31.95, *SD* =11.00) through Prolific Academic. Sample size was determined using the rationale of Experiment 2. All participants voluntarily agreed to participate in the study and were monetarily compensated (38 pence). The average completion time was 4 minutes.

Materials and procedure

We used the same materials and procedures from Experiment 2. We included an additional measure for participants to rate the blameworthiness of the character in the vignette (praiseworthy = 0, neither praiseworthy nor blameworthy = 50, blameworthy =100).

Results

Overall, the results replicated the pattern observed in Experiment 2 (see Figure 7). We fitted the same random effect models as in Experiment 2 and performed the same pairwise comparisons. Model 2, which includes the interactive effect of judgment type, omission and erogation condition, was the best model. Table 4 below presents only the two best fitting models.

627

628 Table 4. Summary of Models fitted for Experiment 4

	Model 1	**Model 2**
Omission	3.90**	5.70***
	(0.19, 7.60)	(2.40, 8.90)
GOOD	-19.00***	29.00***
	(-23.0, -16.0)	(26.0, 32.0)
OBLIGATORY	-41.00***	-16.00***
	(-45.00, -38.00)	(-19.0, -13.0)
PERMISSIBLE	4.20**	40.0***
	(0.47, 7.90)	(37.00, 43.00)
SUPER EROGATORY	-50.00***	5.20***
	(-53.00, -46.00)	(3.60, 6.80)
Mowing		-1.0
		(-4.20, 0.460)
Newlyweds		-1.400
		(-3.70, 0.92)
Raffle		0.260
		(-2.00, 2.60)
Omission:GOOD	-12.00***	-14.00***
	(-18.00, -7.20)	(-19.00, -9.30)
Omission:OBLIGATORY	3.70	0.180
	(-1.50, 9.00)	(-4.40, 4.80)

Omission:PERMISSIBLE	-7.90***	-10.00***
	(-13.00, -2.70)	(-15.0, -5.70)
Omission:SUPER EROGATORY	3.50	
	(-1.70, 8.80)	
GOOD: SUPER EROGATORY	97.00***	
	(92.0, 102.0)	
OBLIGATORY: SUPER EROGATORY	51.00***	
	(46.0, 56.0)	
PERMISSIBLE: SUPER EROGATORY	72.00***	
	(67.00, 77.00)	
Omission: GOOD: SUPER EROGATORY	-3.00	
	(-10.00, 4.40)	
Omission:OBLIGATORY: SUPER EROGATORY	-7.10*	
	(-14.0, 0.36)	
Omission:PERMISSIBLE: SUPER EROGATORY	-4.80	
	(-12.00, 2.60)	
Constant	62.00***	35.00***
	(59.00, 64.00)	(32.00, 38.00)
N	4976	4976
Log Likelihood	-22765.00	-23867.00
AIC	45566.00	47761.00
BIC	45683.00	47852.00

***p < .01; **p < .05; *p < .1

629

As before, both permissibility and goodness ratings were significantly higher for the supererogatory (Permissibility M = 85.4, SD = 19.5, n = 622 and goodness M =85.6, SD = 20.9, n = 622) than for the suberogatory condition (M = 63.9, SD = 29.9, n = 616 and goodness M =38.0, SD = 20.5, n = 622) permissibility: $t(4580) = -16.08$, $p < .001$, $M_{diff} = -21.5$, 95% CI [-24.3, -18.7], goodness: $t(4580) = -35.54$, $p < .001$, $M_{diff} = -47.2$, 95% CI [-49.9, -45.3]. However, the difference between obligatory ratings was not replicated ($t(4580) = 0.22$, $p = .82$, $M_{diff} = 0.32$, 95% CI [-2.77, 3.41]). Blame ratings differed significantly across the suberogatory and supererogatory conditions. Participants rated suberogatory behaviors as more blameworthy (M = 63.70, SD = 20.77, n = 622) than supererogatory behaviors (M = 15.76, SD = 20.62 , n = 622) $t(4580) = 35.82$, $p < .001$, $M_{diff} = 47.9$, 95% CI [45.6, 50.2]. As in Experiment 2, there are significant differences between scenarios, with supererogatory responses in the *Kidney* scenario having the highest positive scores (as better, more permissible, more praiseworthy) and suberogatory responses in the *Mowing* scenario the most negative (as worse, less permissible, less optional, and more blameworthy).

Blame ratings were not significantly correlated with statements about personal sense of duty across either scenarios or conditions. The same overall pattern of correlations between the other ratings and statements about duties was observed (See Table 2, Experiment 2).

Supererogatory behaviors are consistently rated as praiseworthy, while suberogatory behaviors are consistently rated as blameworthy. However, the magnitude of praiseworthiness judgments is much greater in the case of supererogatory behaviors. This asymmetry suggests that supererogatory behaviors deserve more praise than the corresponding suberogatory behaviors

deserve blame. (see Figure 7). This asymmetry has been repeatedly observed in other studies (Monroe et al., 2018, Pizarro et al., 2007).

It is worth noting that, again as in Experiment 2, there are no important differences between the permissibility and goodness means in the supererogatory condition (M = 85.61, SD = 20.86 vs M = 85.43, SD = 19.46 respectively) while for the suberogatory responses we find a difference of 25.88 points between these (Cohen's d = 1.01), suggesting that participants are considering different information when using assessing permissibility and goodness in these situations.

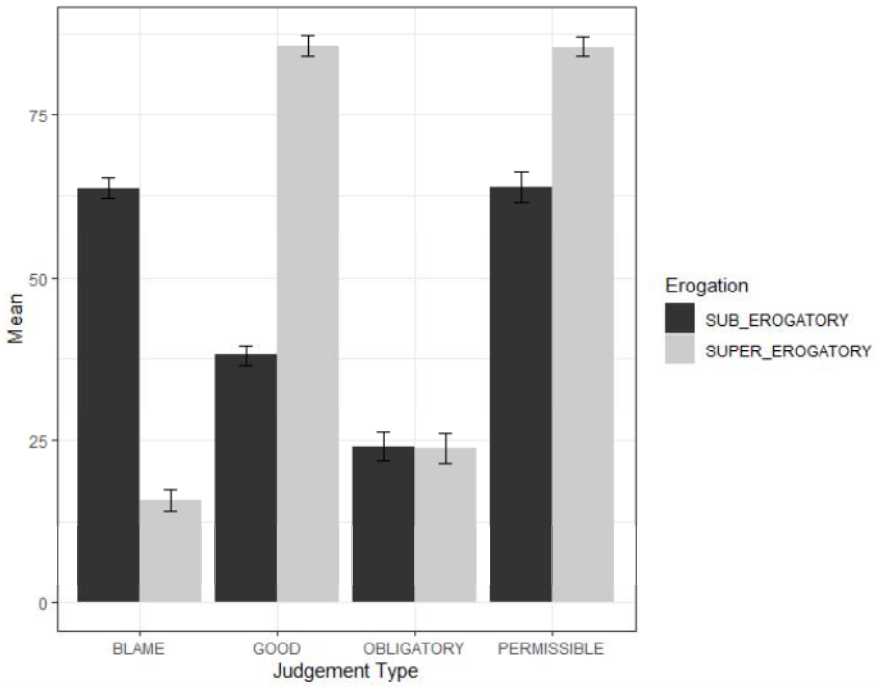


Figure 7. Scores by type of judgment Experiment 4. Error bars represent 95% confidence intervals

Discussion

In this experiment, we replicated the key findings of Experiment 2. When comparing judgments about suberogatory and supererogatory behaviors, people distinguish between the goodness, permissibility, and blameworthiness of the action. However, people do recognize suberogatory and supererogatory behaviors as equally optional. Despite statistically significant differences between the permissibility of suberogatory and supererogatory actions, participants rated suberogatory behaviors as permissible.

Now that we have provided some evidence for distinctions between these evaluative categories, we turn to another question: what factors differentially affect these judgments? As we saw in Experiment 2, beliefs about personal duties do not relate significantly to people's judgments. The significant difference in blame judgments suggests that perhaps norms of social sanctioning help to inform at least one kind of judgment. To understand the reasoning underlying these judgments, we conducted two additional experiments.

Experiment 5

In this experiment we explored how judgments of deserved social sanction are associated with the different kinds of moral judgment identified in previous experiments. Thus far, we have established that some situations can be understood in terms of suberogatory or supererogatory characterizations. In order to understand people's responses to these situations, we need to deploy a richer array of judgment dimensions. In this experiment, we explored whether suberogatory responses are associated with potential behavioral consequences, such as social punishment. Plausibly, social punishment is, as much as expressions of blame, a behavioral marker of underlying moral judgment.

We also explore whether attitudes about rights predict different kinds of judgments. The situations that we describe often bring into conflict an individual right and a kind or generous thing that could be done for others. For example, someone has the *right* to keep the seat they paid for on the airplane, but it is a kind thing to offer one's seat when asked to switch. Because of this, we thought that attitudes towards *rights*, instead of duties, might usefully predict moral judgments. This would partially vindicate the rule-based conception of morality, as one might hold that there are generalized functions from rights to prohibited behaviors (even if there are fewer—if any—such functions from rights to *prescribed* behaviors). Moreover, we found this pattern of reasoning in the open responses for Experiment 1. All of this suggests that personal sense of rights might provide some details about the range of inputs to which moral judgment is sensitive.

Participants

309 participants (165 women and 144 men, mean age = 39.89, SD =13.76), recruited through Prolific Academic, took part in the study in exchange for 60 pence. The average completion time was of 4.87 minutes and none of the participants had taken part in the previous experiments.

Materials and procedure

We used the same materials and procedures from Experiment 4 with two changes. First, people had to judge whether the response merited social sanction or recognition by using a slider with an underlying scale going from 0 to 100 (social sanction = 0, neither = 50, social recognition=100). Second, people provided information about their attitudes towards rights by answering these three questions with a 7-point Likert scale (1= Strongly disagree, 7 = Strongly

agree): 1) Everyone has a right to do anything they want on their property (as long as they are not hurting anyone else); 2) Everyone has the right to bodily autonomy; 3) Everyone has a right to make use of their money (or goods) as they see fit.

Results

As in Experiments 2 and 4, we fitted a series of random effect models which resulted in a best fitting model that includes the interactive effect of judgment type, omission, erogation condition and scenario. Crucially, including terms for attitudes about rights did not significantly improve the fit of this model.

Table 5. Summary of Models fitted for Experiment 5

	score
Omission	-8.08***
	(-10.95, -5.22)
OBLIGATORY	-22.94***
	(-26.74, -19.15)
PERMISSIBLE	19.74***
	(15.94, 23.54)
SANCTION	-2.66
	(-6.45, 1.14)
SUPER EROGATORY	44.90***
	(42.21, 47.59)
Mowing	-5.24***
	(-8.25, -2.24)
Newlyweds	-5.14***
	(-7.04, -3.23)
Raffle	3.68**
	(0.68, 6.69)
Omission: OBLIGATORY	15.28***
	(9.90, 20.65)

Omission: PERMISSIBLE	9.13***
	(3.75, 14.51)
Omission: TypeSANCTION	7.98***
	(2.60, 13.35)
OBLIGATORY: SUPER EROGATORY	-39.76***
	(-45.13, -34.39)
PERMISSIBLE: SUPER EROGATORY	-22.18***
	(-27.55, -16.81)
SANCTION: SUPER EROGATORY	-14.17***
	(-19.54, -8.80)
Omission: OBLIGATORY: SUPER EROGATORY	-3.29
	(-10.90, 4.31)
Omission: PERMISSIBLE: SUPER EROGATORY	-5.62
	(-13.22, 1.99)
Omission: SANCTION: SUPER EROGATORY	-3.65
	(-11.26, 3.96)
Constant	45.96***
	(43.03, 48.90)
N	4944
Log Likelihood	-22788.0
AIC	45617.0
BIC	45747.0

***p < .01; **p < .05; *p < .1

The pattern found in Experiments 2 and 4 was replicated here (see Figure 8). Supererogatory behaviors were judged as better and more permissible (M = 85.16, SD = 19.31, n = 618 and M = 84.47, SD = 20.18, n = 618) than suberogatory behaviors (M = 40.26 SD = 23.99, n = 618 and M = 64.55, SD = 30.09, n = 618) Good: $t(4619) = -32.72$, $p < .001$, $M_{diff} = -44.4$, 95% CI [-47.3, -42.5] and Permissible: $t(4619) = -14.51$, $p < .001$, $M_{diff} = -19.9$, 95% CI [-22.8, -17.1], regardless of being actions or omissions. Supererogatory behaviors were again judged to be less optional (M = 28.43, SD = 30.61, n = 618) than the corresponding suberogatory responses

721 (M = 24.93, SD = 26.34, n = 618), t(4619) = -2.55, p = .01, M_{diff} = -3.50, 95% CI [-6.68, -0.31].
722 As in Experiment 4 this difference only emerged when an additional measure was included
723 (blameworthiness in Exp 4 and social sanction in this Exp) suggesting the presence of a joint
724 evaluation effect (Hsee, Blount, Loewenstein, & Bazerman, 1999).

725 Supererogatory behaviors were judged to merit more social recognition than suberogatory
726 responses (M = 70.49, SD = 25.41, n = 618 vs M = 41.59, SD = 20.57, n = 618, t(4619) = -21.07,
727 p < .001, M_{diff} = -28.9, 95% CI [-31.5, -26.3]). Crucially, both means for social recognition and
728 social sanction are significantly different from the indifference point (50 in our scale)
729 (Supererogatory t(617) = 20.1, p < .001, M = 70.5, 95% CI [68.5, 72.5] and suberogatory t(617)
730 = -10.2, p < .001, M = 41.5, 95% CI [40.0, 43.2]. Mirroring the asymmetric pattern observed for
731 blameworthiness ratings, supererogatory responses deserve more recognition than sanction is
732 deserved by suberogatory responses. Also, as in the previous studies there are important
733 differences between scenarios, even when the overall pattern is consistent. For example, not
734 giving up the raffle prize does not merit neither recognition nor sanction (M = 50.82, SD = 22.06,
735 n =155) while mowing your lawn early morning is clearly disapproved (M = 34.23, SD = 19.76,
736 n = 155).

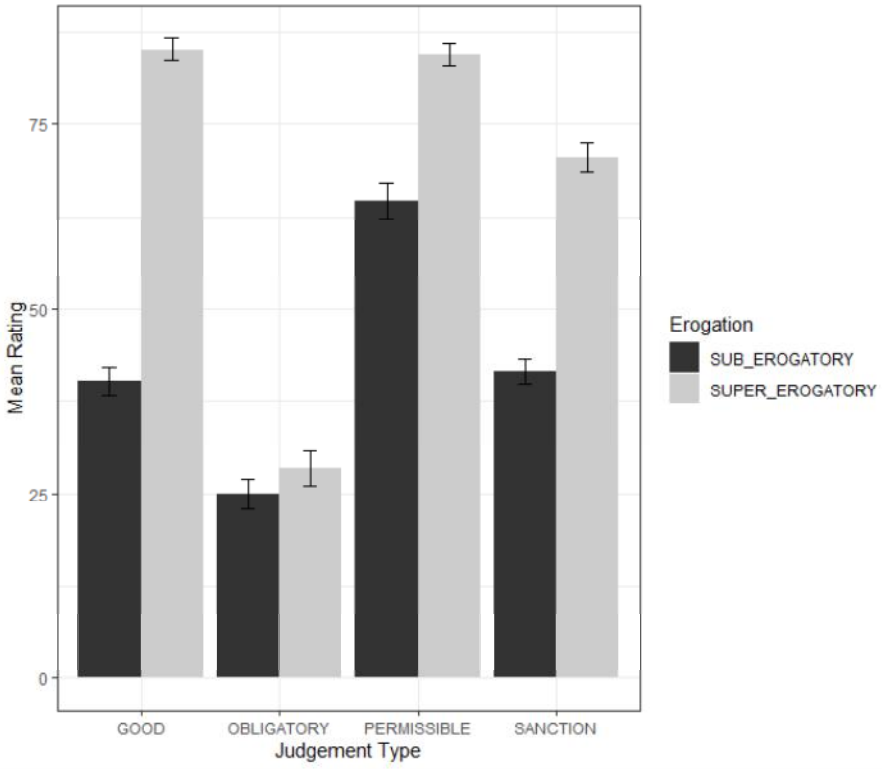


Figure 8. Scores by type of judgment and condition. Error bars represent 95% confidence interval.

To explore the association between judgment scores and attitudes towards rights we first observed the strength of the association between attitudes towards rights. Correlations are medium to small in size (see Table 6) and there is low internal consistency (Cronbach's alpha = 0.50). We then calculated the correlations between judgment scores and these attitudes towards rights (Table 6). Similar to what happened with questions about duties, the correlations are low and only significant for two associations: (1) bodily autonomy and permissibility, and; (2) obligatoriness and right to property. Only when considering particular scenarios, moderately stronger correlations start to emerge (for example, permissibility ratings in the *Kidney* scenario are significantly correlated with attitudes on bodily autonomy r (309) = .12, $p < 0.01$)

Table 6. Bivariate correlations between attitudes towards rights and judgment scores

	Permissible	Good	Obligatory	Sanction	Right to Money	Bodily autonomy
Permissible						
Good	0.54***					
Obligatory	-0.24**	-0.04***				
Sanction	0.24***	0.40***	0.14***			
Right to money	0.03	0.04	-0.05	0.01		
Bodily autonomy	-0.16**	0.05	-0.14	0.00	0.16***	
Right to property	0.05	0.02	0.06*	0.01	0.44***	0.17***

$^{**}p < .001$; $^{**}p < .01$; $^{*}p < .05$

Discussion

We find the same dissociations observed in Experiments 2 and 4. Including the category of sanctions, however, brings out something interesting. There is a much closer relationship between goodness and sanction than there is between goodness and blame in suberogatory cases. The relationship does not hold in cases of supererogatory behavior, where judgments of goodness were statistically distinct from judgments of sanction (positive recognition, in this case). This suggests that people think suberogatory actors should be sanctioned, though supererogatory actors need not necessarily be positively recognized. This, however, might be a function of the cases we used rather than marking out an intrinsic difference between commonsense thinking about the sub- and supererogatory.

We also identified an interesting relationship between sanction and goodness. Actions rated as bad received similar ratings of sanction. In the previous experiment, we saw that ratings of badness do not align with ratings of permissibility. Permissibility ratings are more closely aligned with ratings of blame. This suggests that the permissibility of subererogatory behaviors is related to blameworthiness, whereas the badness of suberogatory behaviors is related to public

765 sanction. However, there is an asymmetry between subererogatory and supererogatory behaviors.
766 We discuss this pattern more in the General Discussion.

767 Lastly, we did not find any interesting correlations between beliefs about rights and
768 different kinds of moral judgment. This was similar to our findings about individual sense of duty
769 in Experiment 2. This suggested the absence of general function from facts about individual
770 rights to behavior evaluation in moral encounters, providing further evidence that commonsense
771 thinking about morality does not reflect a structured system of rules.

772 **Experiment 6**

773 So far we have shown that judgments of permissibility, goodness, and obligation track
774 distinct features of a situation. Evaluations of supererogatory and suberogatory behaviors brings
775 these distinctions to light. One question that remains is what factors drive these different
776 judgments. The previous experiments offered some suggestions. In Experiment 1, we found that
777 participants used the language of rights to offer justifications for suberogatory behavior.
778 Additionally, they used virtue terms and aretaic categories in free descriptions of the scenarios.
779 Experiments 2, 4 and 5 suggest, however, that people are not explicitly taking into account their
780 attitudes towards duties (Exps. 2 and 4) or rights (Exp 5). This suggests two things. Judgments of
781 goodness might be sensitive to what people think a decent or virtuous person would do in a
782 situation, while judgments of permissibility track what some individual has the right to do in a
783 given situation.

784 If these suggestions are correct, it would explain why judgments of permissibility and
785 goodness often coincide: what is decent or virtuous often overlaps with what one has a right to

do. However, it also explains why the two concepts dissociate in cases of suberogatory behavior. Suberogatory situations arise when someone has a right to do something that a virtuous or decent person would not do. If this is the case, it is the absence of duties and the corresponding presence of rights that are characteristic of suberogatory situations. These characteristics are, however, not framed as rules, but are instead low-level features of the situation.

In this experiment, we explore the justifications people offer for their judgments of permissibility and goodness to see whether these different factors explain the distinction between these two kinds of judgment. We also include scenarios that are more perspicuously *moral* than the ones previously used (e.g. lawn mowing or seat ownership). By providing situations with a higher moral charge, we can be more confident that the responses observed so far are not merely tracking the perceived conventional permissibility or social aptness of the behaviors evaluated.

Method

Participants

We recruited 316 participants (160 women and 153 men, mean age = 33.30, SD =10.81) through Prolific Academic to participate in the study in exchange for 60 pence. Sample size was set to reproduce findings of Experiments 2, 4 and 5 but with a within participants design in this case, which would allow us to increase statistical power. The average completion time was 10 minutes and none of the participants had taken part in the previous experiments.

Materials and procedure

We constructed two new scenarios and created a suberogatory and supererogatory version of each. The two scenarios (adapted from Thomson, 1971 and Nozick, 1974, respectively) are described below [suberogatory version in brackets]:

Violinist Alex is driving home from work on the highway when she gets into an accident that knocks her unconscious. When she wakes up, she finds herself in a hospital bed. She's also connected to another individual through a series of wires and tubes. A doctor enters the room and explains to Alex that she is fine, but the individual she's connected to suffered some severe damage to internal organs. Alex has the right blood type to help, and—since she was unconscious—the doctor decided to connect Alex to keep the other individual alive for the time being. The doctor explains that Alex can unplug herself if she chooses, but the individual will most likely die. The individual will recover from these injuries in about a month (give or take a few days), after which time Alex can unplug herself and leave. After a few hours of pondering what to do, Alex decides stay plugged in for the month [*to unplug herself*].

Well Jones finds a large freshwater source on his property, so he digs a well as a way of claiming the water. A few weeks later, the town where he lives begins experiencing a drought, which was completely unpredictable. Town representatives visit Jones to ask whether they can use his water to alleviate some of the drought. Without Jones' help, the town will likely run out of water in a few days. If Jones donates some of his water, however, he might experience the effects of the drought in the unlikely event that the drought prolongs for too long. After considering what to do, Jones decides to offer his water [*declines to offer his water*]

To test variation against a known benchmark, we also included the *Newlyweds* scenario (See Exp. 5). Participants read the informed consent and then rated each one of the three vignettes. The order of the vignettes was randomized across participants. Each participant was randomly assigned to see either the supererogatory or suberogatory condition of each scenario and the rest of the procedure was the same as in Experiment 4. The only difference is that, after

828 providing judgment for the dimensions requested, participants were asked to explain their ratings
829 in their own words (cf. Christensen et al., 2014).

830 **Results**

831 Overall, numerical ratings closely followed the pattern observed in prior Experiments (See
832 Figure 9). Suberogatory behaviors are rated as worse (M = 34.19, SD = 23.71, n = 468 vs M =
833 86.37, SD = 19.80, n = 480) , more blameworthy (M = 64.13, SD = 22.46, n = 468 vs M = 14.92,
834 SD = 19.12, n = 480), and less permissible (M = 28.17, SD = 32.31, n = 468 vs M = 83.15, SD =
835 22.27, n = 480) than supererogatory responses (all significant pairwise comparisons). Unlike
836 Experiment 5, there is no significant difference in obligatoriness judged (Suberogatory M =
837 28.17, SD = 32.31, n = 468 vs M = 26.64, SD = 31.75, n = 480). As before, we fitted random
838 effect models, with a random intercept for participants. However, models with this term resulted
839 in singular fits, due to lack of variation for the random intercept for participant and suggesting an
840 overcomplex random structure (Matuschek, Kliegl,Vasishth, Baayen & Bates (2017). Therefore,
841 we fitted only fixed effect models, where the best model includes the interaction of judgment
842 type, scenario and erogation condition (results can be seen in Table 7).

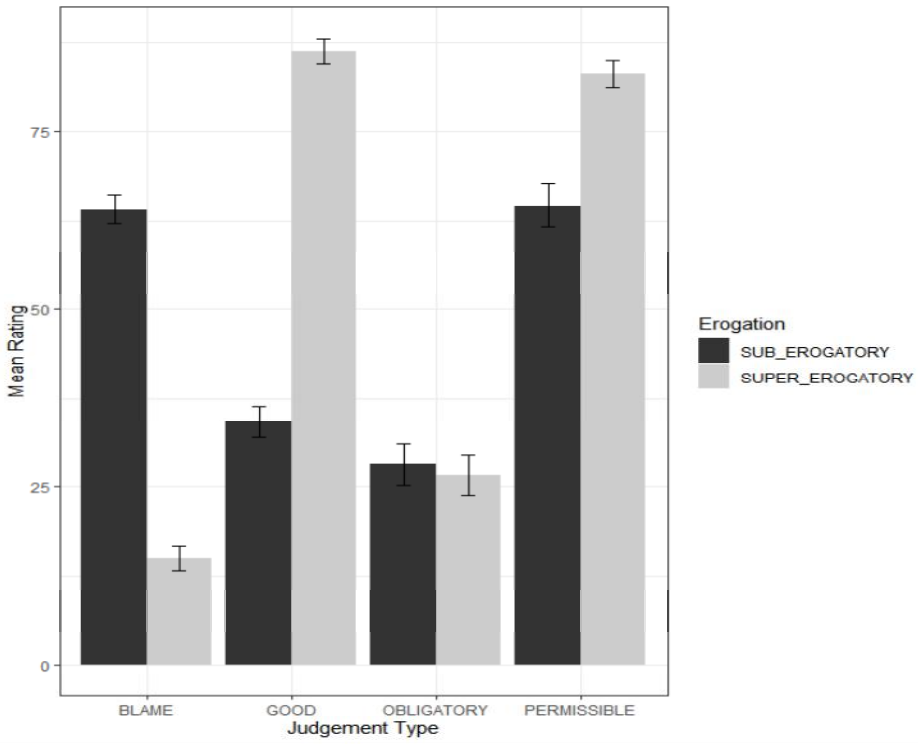


Figure 9. Mean scores by type of judgment Experiment 4. Error bars represent 95% confidence intervals

There is also a significant between-scenario variation. For example, only in the *Well* scenario people judged the suberogatory behavior, not giving water, as significantly less optional than the supererogatory behavior ($M = 31.26$, $SD = 31.75$, $n = 160$ vs $M = 38.81$, $SD = 34.63$, $n = 156$, Good: $t(310)$[4] $= 2.64$, $p < .001$, $M_{diff} = 7.55$, 95% CI [0.19, 14.9]). In this scenario also occurred the most extreme values for suberogatory responses.

[4] Degrees of freedom correspond to the Welch t-test, since the library *emmeans* (Lenth, 2020) calculated the asymptotic result for this comparison (value from the *z* distribution).

851

Table 7. Summary of Best Model fitted for Experiment 6

	score
GOOD	-18.65***
	(-24.33, -12.97)
OBLIGATORY	-43.06***
	(-48.74, -37.38)
PERMISSIBLE	14.01***
	(8.33, 19.70)
SUPER_EROGATORY	-35.49***
	(-41.08, -29.89)
Violinist	3.61
	(-2.02, 9.24)
Well	13.13***
	(7.48, 18.79)
GOOD: SUPER_EROGATORY	77.22***
	(69.31, 85.13)
OBLIGATORY: SUPER_EROGATORY	36.02***
	(28.11, 43.93)
PERMISSIBLE: SUPER_EROGATORY	50.07***
	(42.16, 57.98)
GOOD: Violinist	-7.79*
	(-15.75, 0.17)
OBLIGATORY: Violinist	10.88***
	(2.92, 18.84)
PERMISSIBLE: Violinist	-10.41**
	(-18.37, -2.45)
GOOD: Well	-25.91***
	(-33.91, -17.91)
OBLIGATORY: Well	10.22**
	(2.22, 18.22)
PERMISSIBLE: Well	-29.89***
	(-37.89, -21.89)
SUPER_EROGATORY: Violinist	-17.16***
	(-25.07, -9.25)

SUPER_EROGATORY: Well	-24.19***
	(-32.10, -16.28)
GOOD: SUPER_EROGATORY: Violinist	27.74***
	(16.56, 38.93)
OBLIGATORY: SUPER_EROGATORY: Violinist	19.66***
	(8.47, 30.84)
PERMISSIBLE: SUPER_EROGATORY: Violinist	12.94**
	(1.76, 24.13)
GOOD: SUPER_EROGATORY: Well	44.97***
	(33.79, 56.16)
OBLIGATORY: SUPER_EROGATORY: Well	16.10***
	(4.91, 27.29)
PERMISSIBLE: SUPER_EROGATORY: Well	39.84***
	(28.66, 51.03)
Constant	58.52***
	(54.51, 62.54)
N	3792
Log Likelihood	-17629.00
AIC	35306.00

***p < .01; **p < .05; *p < .1

Qualitative data

For the qualitative analysis of the open responses, we used two coders. Both were blind to the individual ratings associated with each open response and one coder was completely blind to the objective of the study. Coders used ten predefined categories to sort responses. We report agreement between coders (Cohens Kappa) and results from fitting loglinear models on the classification frequencies.

There were 4003 unique written explanations of ratings (1004 for goodness ratings, 999 for obligatory ratings, 997 for permissibility ratings and 1003 for blame/praise ratings). Coders

were given ten categories. *Character* when the response included some mention of personality or character traits; *Rights (presence or absence)* when responses appealed to what the protagonist can morally do, is allowed to do, or its explicit negation; *Duties (presence or absence)* when responses appeal to what the protagonist is morally obliged to do, or its explicit negation; *Neither* when participants appealed to norms that could not be considered rights nor duties, etc; *Outcomes* when the justification was based on the consequences for the "victims", negative or positive; *Values* actions performed had a clear valence; *Justification* when appealing to reasons not based on consequences; *Other* if the response did not fall in the previous categories but was common enough to merit its own. Coders were given an example of each one of the categories. Raters completed coding independently. They were also told that they could classify any given response in more than one category.

For example, one response to the *Well* vignette reads: "He didn't have to donate the water, and even if he didn't that wouldn't be necessarily a bad thing." This was rated as indicating 'Absence of Duties' and 'Values'. Another response (to the *Newlyweds* vignette) reads: "No one should have to switch a seat unless they want to. They bought the seat they were in and would have been justified in staying in that seat." This was rated as indicating 'Justification' and 'Presence of Rights'.

Raw agreement between coders and Cohen's Kappa for each category are presented in Table 8. The category *Neither* is not presented because there was virtually no agreement in its use. Most of the unique responses were classified into at least 2 categories (58%). Half of the categories show a weak reliability between coders (outcomes, values, justification and absence of

rights) while the other half shows an overall good agreement (presence of duties, character, presence of rights and absence of duties).

Table 8. Cohen's kappa agreement between coders for the eight categories. Columns L and U are the Lower and Upper limits of Cohen's confidence interval.

	Kappa	L	U	% Classified as	% Agreement
Abs duties	0.81	0.79	0.84	11.02	95.68
Pres rights	0.75	0.73	0.77	21.06	90.50
Character	0.69	0.65	0.72	8.61	94.18
Pres duties	0.63	0.58	0.68	5.65	95.25
Outcomes	0.39	0.35	0.43	11.86	85.08
Values	0.38	0.35	0.42	20.28	77.18
Justification	0.37	0.34	0.41	13.7	82.35
Abs rights	0.34	0.27	0.41	3.13	95.15

To determine the association between justification category and the variables manipulated in this experiment, we submitted the classification frequencies to a loglinear analysis with a saturated model including erogation condition (sub and supererogatory), judgment dimension (goodness, permissibility, obligatoriness, blameworthiness) and justification classification category (presence of duties, character, presence of rights and absence of duties). The three-way loglinear model produced a model that retained all interactive effects (likelihood ratio χ^2 (0) = 0, p = 1), indicating a significant interaction between the variables fitted to the model (χ^2 (12) = 102.53, $p < 0.001$). Judgments of permissibility and obligatoriness were mainly sensitive to presence of rights and these explanations appealing to the presence of rights were more prevalent in suberogatory cases. Judgments of permissibility were also justified by appealing mostly to the absence of duties and the presence of rights (>40%) but in this case there is no asymmetry

between sub and supererogatory response. Judgments of goodness were mostly justified by appealing to character and again the suberogatory responses were justified by relying on presence of rights. Additionally, there are more character justifications for supererogatory responses based on character than for the corresponding suberogatory responses.

The qualitative analysis of the open explanations suggests that people do systematically call to mind considerations on duties and rights when making judgments about permissibility and obligatoriness. While this is not surprising, the interesting thing here is that these considerations are also significant for judgments of blameworthiness and goodness. Duties and rights showed up, not as blank generalizations – as explored in Experiments 2 and 5 – but as situated appeals to factors of the situation.

Discussion

As before, we replicated the basic pattern found in our other experiments using two situations that are more clearly moral. This provides strong evidence for our claim that differences in evaluative categories track differences in kinds of *moral* judgments.

We also used open responses to allow participants to supply their own reasoning for making different moral judgments. Our results are striking. People appear to be sensitive to many kinds of information differentially when making moral judgments. Judgments of praise and blame are sensitive to character considerations. Judgments of goodness seem primarily sensitive to character considerations, whereas judgments of badness seem sensitive to both character considerations and the presence of rights. This, however, might be a function of the situations we provided. When people make judgments of badness, they were often pointing to the fact that the

bad behavior was fully within the rights of the individual, even though the behavior was indicative of vicious character. The fact that people mention both points to the relevance of both kinds of information to making a judgment of badness. Judgments of permissibility appear to track the presence of rights and the absence of duties, whereas judgments of obligatoriness track the presence of rights and the presence or absence of duties. We offer an interpretation of this in the general discussion.

Finally, we should note that the categories used to sort open responses were generated *a priori*. Bottom-up, data-driven methods might reveal a more perspicuous classificatory scheme that might generate surprising results about the kinds of information relevant to making different kinds of moral judgment.

General Discussion

In this paper, our goal was to develop the idea that morality is not exhausted by a system of rules and explore some of its implication for our understanding of the psychology of moral judgment. To do this, we've made a case for moving beyond the use of dilemmas in the study of moral judgment and propose a wider array of measures and situations. In this General Discussion, we summarize how our findings apply to the measurement of moral judgment, the cases used to elicit judgments, and the rule-based view of morality.

Measures and Categories of Moral Judgment

We found that judgments across different moral categories are dissociable. In particular, judgments of goodness and permissibility come apart depending on the situation presented.

People form judgments about situations that conform to the category of the suberogatory (bad, but permissible). This challenges the idea that different measures of moral judgment tap into the same underlying reasoning process. The resulting picture of moral judgment is that people are variably sensitive to different kinds of considerations when making different kinds of judgments (see Barbosa and Jiménez-Leal, 2017). While this presents a more granular picture of moral judgment, it also opens up new questions. When do people prefer different kinds of information in making moral judgments? What sorts of considerations predominantly drive different kinds of judgments?

Our results provide some initial suggestion for how to answer these questions. In many of our experiments, ratings of obligation remained consistently low. This indicates that people considered both suberogatory and supererogatory behaviors as optional rather than obligatory. The qualitative data from Experiment 6 suggests that ratings of obligation are predominantly tracking the presence of duties or explicit rules that prohibit or prescribe conduct. Because the situations considered here are not composed of conflicts between rules, there is an absence of duties that defines how one ought to behave. Hence, ratings of obligation are low.

Results from Experiment 5 indicate that ratings of badness align with ratings of sanction. That is, the degree to which one rates a suberogatory action as bad is related to amount of sanction one deserves in light of doing something suberogatory. This is importantly different from the results of Experiment 4, which indicated that ratings of permissibility are aligned with ratings of blame.

These results suggest that folk psychological categories of blame and sanction might dissociate. Because these categories differentially associate with others, it might be the case that

these judgments track different features of a situation. Understanding this difference might provide further insight into what judgments of permissibility and goodness indicate about folk psychological evaluation. This is an important result to investigate in future work. Moral psychologists and philosophers have assumed that these constructs significantly overlap and, accordingly, have used them interchangeably (Bendor & Swistak, 2001; Deutsch & Gerard, 1955; Cialdini, Reno, & Kallgren, 1990; Scanlon, 2008; McKenna, 2013; Bennett, 2012). Future work should investigate further whether and under what circumstances there are reliable dissociations between these categories.

Notably, some have mentioned the need for using new measures in studying moral judgment (see Uhlmann, Pizarro, & Diermeier, 2015), arguing that folk psychological categories of judgment are fundamentally directed at personal evaluation rather than behavior evaluation and routinely employs aretaic rather than deontic concepts. Our results suggest that this is partially true. People do show an interest in personal evaluation using aretaic concepts. However, we also find that behavior evaluation and deontic concepts also play a significant role. This suggests that social cognition is sensitive to both act-based and person-based considerations, as well as a wide conceptual repertoire in normative evaluation.

Moral Situations

Using a variety of measures is not sufficient to bring out the underlying variability of moral judgment. In Experiment 3 (on sacrificial dilemmas), we did not find similar dissociations between judgment categories as we did in our other experiments. This was the case even when controlling for underlying ethical tendencies (i.e., low or high trait utilitarianism). Even among low trait utilitarians, different kinds of judgments never significantly differed from each other.

This suggests that dilemmas themselves are qualitatively distinct from the kind of moral encounters we used in our experiments.

As discussed earlier, assumptions about methodology and dilemmas mutually reinforce each other. We can now clarify this point further. The use of dilemmas as stimuli functions as a demand effect that gives the appearance of commonsense moral thinking exhibiting a rule-based, hierarchical structure. In providing scenarios that explicitly bring different sets of rules into conflict, researchers have set people up to exhibit judgments that appear to preferentially select some rules over others (deontological prescription vs. maximization principles). This, in turn, blurs the distinctions between different judgment categories. Once we free moral judgment of the constraints of dilemmas, variegated measures capture better the variability and nuance of judgment. That variability is only possible when using situations that do not bring different sets of rules into conflict.

Different situations also allow for a different kind of variability. Moral dilemmas have mostly been constructed out of examples designed to test abstract philosophical principles. which has led to worries about their ecological validity (Kahane, 2015; Dahl & Oftedal, 2019). A related but, perhaps more significant, concern is that dilemmas, because of their abstract and artificial nature, likely occlude differences in moral attitudes that arise from socio-cultural variation. Thus, in the limited number of studies that have been conducted outside the urban areas of North America and Western Europe, researchers have found responses similar to those found among WEIRD populations (Abarbanell & Hauser, 2010; Barrett et al., 2016; Koenigs et al., 2007; Perkins et al., 2013; Szekely & Miu, 2015; Johnson, Danko, Huang, Park, Johnson, & Nagoshi, 1987).

By contrast, it is likely that responses that are deemed suberogatory in one community, such as tipping 10%, might be considered supererogatory by another one. As cultural variability mediates moral learning, it is plausible to conjecture that different cultures assign varying weights to values, virtues, and rules in the justification of their moral judgments (Graham, Meindl, Beall, Johnson & Zhang, 2016). Consequently, cultural differences can be observed not as variation in a dimension (e.g. being more or less utilitarian) but as a particular pattern of judgements and their justifications. At present, we do not have results that speaks directly to these issues—although we currently have relevant work in preparation. We do think, however, that moving beyond dilemmas opens up the possibility for this variability to emerge.

Rule-based Morality

We mentioned at the outset that not all moral situations reflect conflicts between rules. Our results show that when people approach some of these situations and are asked to evaluate them, they bring to bear a variety of considerations, not all of which are codifiable in rules. This strongly suggests that moral judgment does not necessarily rely on the application of abstract *moral* principles.

In Experiments 2, 4 and 5, we asked participants to fill questionnaires aimed at measuring general attitudes towards duties and rights. We failed to find correlations between their attitudes toward abstract claims about the nature of rights and duties and their judgments about particular situations. This reflects a tendency of people to identify features of situations that guide judgment without having explicit representations of abstract principles play a causal role in forming moral judgments (Graham et al., 2013). Likewise, the open responses we collected in Experiment 6 suggest that rules play a role in structuring commonsense moral thinking alongside axiological

and characterological considerations, at least when it comes to explaining proffered moral judgements. In particular, as indicated by the responses under the "absence of duties" category, participants judged actions to be good, bad, etc., while explaining that these didn't violate a prohibition (in suberogatory cases) or were not generally mandated (in supererogatory cases). Instead, their evaluations seem overwhelmingly responsive to considerations about rights, values, and character traits.

For all we have said, it may turn out that precise algorithms adequately characterize the computations performed in forming moral judgments. Alternatively, moral judgment might be supported by an architecture that functions as a model-based system without having to explicitly represent rules (see Crockett, 2013; Cushman, 2015; Brownstein, 2018). Our interest here, however, does not lie with the computations or cognitive architecture behind moral judgment, but with the content of morality. Our claim is that most research on moral judgment has, thus far, assumed that the content of moral judgment can be measured in terms of its alignment with some structured system of moral rules. But fixating on rules provides only a partial window into the moral life, one that fails to reveal the complexities and subtleties of commonsense moral judgment. When moral judgment is removed from the narrow frame of sacrificial dilemmas, the appearance of rules in moral judgment evaporates.

Finally, our argument here does not assume that if morality has a rule-based structure, then commonsense moral judgment ought to exhibit perfectly coherent and systematic principles. Even Sidgwick (1981) admits that the maxims of commonsense morality are "somewhat vague generalities" (p. 342). Still, on his view and on the view of those who follow him, commonsense morality can be refined through rigorous theorizing to *approximate* the structure of an ideal

normative theory. It's made of the right sort of stuff, as it were, to function as a rule-based system.

The judgments that we have identified in these experiments, however, imply that this is an incomplete view of the moral framework reflected in moral cognition. Consider that people regularly judge that some behavior is bad but permissible across a range of different scenarios. If there are rules constitutive of commonsense morality, what must they be like such that they allow for such judgments? How can we make sense of such judgments within a rule-governed system? Recall that one key feature of the rule-based system of morality is the inter-definability of moral concepts. Whatever is permissible is in line with the rules of morality; whatever is bad is bad because it goes against those rules.

One option is to say that separate domains of rules govern separate judgments. But that requires giving up on the notion of inter-definability that is central to the rule-based system of morality. Another option is to say that one kind of judgment is properly moral and the other is not. However, there appears to be no principled way of stating that either judgments of permissibility or goodness are properly moral while excluding the other. A third option is to say that there *are* systematic principles underlying these judgments, but people do not understand what these are. Hence, they are making a mistake in dissociating permissibility and goodness. Lastly, one could argue that morality is self-contradictory.

While all of this is possible in principle, the attempt to defend these answers in the face of the evidence presented here begins to look like the imposition of researcher assumptions rather than an investigation into commonsense moral thinking. At this point, one begins to wonder if

the substantive assumption about morality being a system of rules is worth the cost. We suggest that it is not, and that it is time to consider what morality beyond the rules would look like.

References

Abarbanell, L. & Hauser, M.D. 2010. Mayan morality: An exploration of permissible harms. *Cognition* 115:2, 207-24.

Anderson, A.R. (1958). A reduction of deontic logic to alethic modal logic. *Mind,* 67, 100-103.

Andrade, G. (2019). "Medical ethics and the trolley Problem," *Journal of Medical Ethics and History of Medicine*, 12:3.

Archer, A. (2018). Supererogation. *Philosophy Compass*, *13*(3), 1–9. doi:10.1111/phc3.12476

Barbosa, S., & Jiménez-Leal, W. (2017). It's not right but it's permitted: Wording effects in moral judgement. *Judgment and Decision Making*, *12*(3), 308.

Barrett, H.C., Bolyanatz, A., Crittenden, A.N., Fessler, D.M.T., Fitzpatrick, S., Gurven, M., Henrich, J., Kanovsky, M., Kushnick, G., Pisor, A., Scelza, B.A., Stich, S., von Reuden, C., Zhao, W., & Laurence, S. 2016. Small-scale societies exhibit funamdental variation in the role of intentions in moral judgment. *Proceedings of the National Academy of the Sciences* 113:17, 4688-93.

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *ArXiv Preprint ArXiv:1406.5823*.

Bauman, C. W., McGraw, A. P., Bartels, D. M., & Warren, C. (2014). Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology. *Social and Personality Psychology Compass*, *8*(9), 536–554.

Białek, M. & De Neys, W. (2017). Dual processes and moral conflict: Evidence for deontological reasoners' intuitive utilitarian sensitivity. *Judgment and Decision Making*, *12*(2), 148–167.

Bennett, C. 2012. The Expressive Function of Blame. In J. Coates and N. Tognazzini (Eds.) *Blame: Its Nature and Norms* (Oxford: Oxford University Press).

Björklund, F. (2003). Differences in the justification of choices in moral dilemmas: Effects of gender, time pressure and dilemma seriousness. *Scandinavian Journal of Psychology*, *44*(5), 459–466.

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*(JUL), 1–18. https://doi.org/10.3389/fpsyg.2014.00607

Christensen, J. F., & Gomila, A. (2012). Moral dilemmas in cognitive neuroscience of moral decision-making: A principled review. *Neuroscience & Biobehavioral Reviews*, *36*(4), 1249–1264.

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

Crockett, M. (2013). Models of Morality. *Trends in Cognitive Sciences* 17:8, 363-66.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353–380.

Cushman, F., Young, L., & Greene, J. D. (2010). Our multi-system moral psychology: Towards a consensus view. *The Oxford Handbook of Moral Psychology*, 47–71.

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science*, *17*(12), 1082–1089.

Dahl, F. A., & Oftedal, G. (2019). Trolley Dilemmas Fail to Predict Ethical Judgment in a Hypothetical Vaccination Context. *Journal of Empirical Research on Human Research Ethics*, *14*(1), 23–32.

Darley, J. M., & Shultz, T. R. (1990). Moral rules: Their content and acquisition. *Annual Review of Psychology*, *41*(1), 525–556.

Djeriouat, H., & Trémolière, B. (2014). The Dark Triad of personality and utilitarian moral judgment: The mediating role of Honesty/Humility and Harm/Care. *Personality and Individual Differences*, *67*, 11–16.

Driver, J. (1992). The suberogatory. *Australasian Journal of Philosophy*, *70*(3), 286–295.

Everett, J. A. C., & Kahane, G. (2020). Switching Tracks? Towards a Multidimensional Model of Utilitarian Psychology. *Trends in Cognitive Sciences*.

Fauconnet, P. (2013). *La responsabilité. Étude de sociologie*. Presses Électroniques de France.

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

Gilligan, C. (1982). *In a different voice: Psychological theory and women's development*. Harvard University Press.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral

foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology* (Vol. 47, pp. 55–130). Elsevier.

Graham, J., Meindl, P., Beall, E., Johnson, K. M., & Zhang, L. (2016). Cultural differences in moral judgment and behavior, across and within societies. *Current Opinion in Psychology*, 8, 125-130.

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

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

Haidt, J., & Baron, J. (1996). Social roles and the moral judgement of acts and omissions. *European Journal of Social Psychology*, 26(2), 201-218.

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

Holyoak, K. J., & Powell, D. (2016). Deontological coherence: A framework for commonsense moral reasoning. *Psychological Bulletin*, *142*(11), 1179.

Hornik, K., Zeileis, A., & Meyer, D. (2006). The strucplot framework: visualizing multi-way contingency tables with vcd. *Journal of Statistical Software*, *17*(3), 1–48.

Hsee, C. K., Loewenstein, G. F., Blount, S., & Bazerman, M. H. (1999). Preference reversals between joint and separate evaluations of options: a review and theoretical analysis. *Psychological Bulletin*, *125*(5), 576.

Hurd, H. M. (1998). Duties beyond the call of duty. *JRE*, *6*, 3.

Johnson, R.C., Danko, G.P., Huang, Y.-H., Park, J.Y., Johnson, S.B., & Nagoshi, C.T. 1987. Guilt, shame, and adjustment. *Personatliy and Individual Differences* 8:3, 357-64.

Kahane, G. (2015). Sidetracked by trolleys: Why sacrificial moral dilemmas tell us little (or nothing) about utilitarian judgment. *Social Neuroscience*, *10*(5), 551–560. https://doi.org/10.1080/17470919.2015.1023400

Kahane, G., Everett, J. A. C., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review*, *125*(2), 131.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. 2007. Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature* 446, 908-11.

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

1170 Kohlberg, L., & Hersh, R. H. (1977). Moral development: A review of the theory. *Theory into*
1171 *Practice*, *16*(2), 53–59.

1172 Kohlberg, L., Levine, C., & Hewer, A. (1983). *Moral stages: A current formulation and a*
1173 *response to critics.*

1174 Lee, J. J., & Gino, F. (2015). Poker-faced morality: Concealing emotions leads to utilitarian
1175 decision making. *Organizational Behavior and Human Decision Processes*, *126*, 49–64.

1176 Lenth, R. (2020). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package
1177 version 1.4.5. https://CRAN.R-project.org/package=emmeans

1178 Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error
1179 and power in linear mixed models. *Journal of Memory and Language*, 94, 305-315.

1180 McKenna, M. 2013. *Conversation and Responsibility* (Oxford: Oxford University Press).

1181 Mikhail, J. (2011). *Elements of moral cognition: Rawls' linguistic analogy and the cognitive*
1182 *science of moral and legal judgment*. Cambridge University Press.

1183 Moll, J., & de Oliveira-Souza, R. (2007). Moral judgments, emotions and the utilitarian brain.
1184 *Trends in Cognitive Sciences*, *11*(8), 319–321.

1185 Monin, B., Pizarro, D. A., & Beer, J. S. (2007). Deciding versus reacting: Conceptions of moral
1186 judgment and the reason-affect debate. *Review of General Psychology*, *11*(2), 99–111.

1187 Monroe, A. E., Dillon, K. D., Guglielmo, S., & Baumeister, R. F. (2018). It's not what you do,
1188 but what everyone else does: On the role of descriptive norms and subjectivism in moral
1189 judgment. *Journal of Experimental Social Psychology*, *77*(March), 1–10.
1190 https://doi.org/10.1016/j.jesp.2018.03.010

1191 Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., van't
1192 Veer, A. E., & Vazire, S. (2019). Preregistration is hard, and worthwhile. *Trends in*
1193 *Cognitive Sciences*, *23*(10), 815–818.

1194 Nozick, R. (1974). *Anarchy, state, and utopia* (Vol. 5038). New York: Basic Books.

1195 O'Hara, R. E., Sinnott-Armstrong, W., & Sinnott-Armstrong, N. A. (2010). Wording effects in
1196 moral judgments. *Judgment and Decision Making*.

1197 Patil, I., Zucchelli, M., Kool, W., Campbell, S., Fornasier, F., Calo, M., Silani, G., Cikara, M., &
1198 Cushman, F. 2020. Reasoning supports utilitarian resolutions to moral dilemmas across
1199 diverse measures. *Journal of Personality and Social Psychology*.

1200 Perkins, A.M., Leonard, A.M., Weaver, K., Dalton, J.A., Mehta, M.A., Kumari, V., Williams,
1201 S.C.R., & Ettinger, U. 2013. A dose of ruthlessness: Interpersonal moral judgment is
1202 hardened by the anti-anxiety drug lorazepam. *Journal of Experimental Psychology: General*
1203 142:3, 612-20.

Piaget, J. (2013). *The moral judgment of the child*. Routledge.

Pizarro, D., Uhlmann, E., & Salovey, P. (2003). Asymmetry in judgments of moral blame and praise: The role of perceived metadesires. *Psychological Science*, *14*(3), 267–272.

Rest, J. R. (1992). *Development in judging moral issues*. U of Minnesota Press.

Rosen, B. (1980). Moral dilemmas and their treatment. *Moral Development, Moral Education, and Kohlberg. B. Munsey (Ed).(1980)*, 232–263.

Scanlon, T.M. 2008. *Moral Dimensions* (Cambridge, MA: Harvard University Press).

Sidgwick, H. (2019). *The methods of ethics*. Good Press.

Simpson, E.L. 1974. Moral development research: A case study of scientific cultural bias. *Human Development* 17, 81-106.

Singmann, H., & Kellen, D. (2019). An Introduction to Mixed Models forExperimental Psychology. In D. H. Spieler & E. Schumacher (Eds.),*New Methods in Cognitive Psychology*. Psychology Press.

Snarey, J.R. 1985. Cross-cultural universailty of social-moral development: A critical review of Kohlbergian research. *Psychological Bulletin* 97:2, 202-232.

R Core Team. (2019). *R: A language and environment for statistical computing*.

Szekely, R.D. & Miu, A.C. 2013. Incidental emotions in moral dilemmas: The influence of emotion regulation. *Cognition & Emotion* 29:1, 64-75.

Thomson, J. J. (1976). A defense of abortion. In *Biomedical ethics and the law* (pp. 39–54). Springer.

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science*, *10*(1), 72–81.

Westfall, J. (2015). PANGEA: Power analysis for general ANOVA designs. *Unpublished manuscript.* Available at *http://jakewestfall. org/publications/pangea. pdf*.

Walker, L.J. 1984. Sex differences in the development of moral reasoning: A critical review. *Child Development* 55:3, 677-691.

Young, I. M., (1985). Impartiality and the civic public: Some implications of feminist critiques of moral and political theory. *Praxis International*, *5*(4), 381–401.