Within your rights: dissociating wrongness and permissibility in moral judgment

Short title: *Wrongness and permissibility*

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

[1] Laboratorio de Emociones y Juicios Morales, Universidad de Los Andes, Bogotá, Colombia
[2] Philosophy Department, Providence College, Providence, RI
[3] Department of Psychology, Universidad de Los Andes, Bogotá, Colombia
[4] Department of Philosophy, Universidad de Los Andes, Bogotá, Colombia

*Please address correspondence to Samuel Murray, 105 Siena Hall, 1 Cunningham Sq., Providence, RI, 02918 (email: smurray7@providence.edu). All Authors contributed equally to this project.

Abstract:
Are we ever morally permitted to do what is morally wrong? It seems intuitive that we are, but evidence for dissociations among judgment of permissibility and wrongness are relatively scarce. Across 4 experiments (N = 1,438), we show that people judge that some behaviors can be morally wrong and permissible. The dissociations arise because these judgments track different morally relevant aspects of everyday moral encounters. Judgments of individual rights predicted permissibility but not wrongness, while character assessment predicted wrongness but not permissibility. These findings suggest a picture in which moral evaluation is granular enough to express reasoning about different types of normative considerations, notably the possibility that people can exercise their rights in morally problematic ways.

Keywords:
moral judgment; wrongness; permissibility; individual rights; suberogatory; moral encounters

Data availability statement:
Materials, data, analysis code, preregistrations, and additional analyses can be found at the OSF repository for the project: https://osf.io/jp7tg/

Acknowledgements:
The Authors thank Paul Henne, Matthew Stanley, and Walter Sinnott-Armstrong for helpful suggestions on experimental design and interpretation. This work was supported in part by the James S. McDonnell Foundation (Grant #2020-1200) and the John Templeton Foundation (Grant #60845).

1. Introduction

Are we ever permitted to do what is wrong? To some, this would be a contradiction. We might be *legally* allowed to do what is morally wrong, but if something is morally wrong, then it is also morally impermissible. Yet, certain examples suggest otherwise. It seems morally wrong to waste food, but to say that wasting food is impermissible feels pedantic. It may be selfish or thoughtless, but you have a right to do what you want with your food. It's yours, after all.

Some sophisticated normative theories make room for the possibility of permissible wrongdoing, or *suberogatory* behavior (Chisholm, 1963; Driver, 1992; Hurd, 1998). People sometimes underperform relative to moral standards without violating any imperatives. This might consist in behaving selfishly or callously toward others: this can be wrong, but it's not forbidden. While the normative basis of the suberogatory is contested (Heyd, 1982; Ullmann-Margalit, 2011), the basic idea seems to resonate with some commonsense moral intuitions (Barbosa & Jiménez-Leal, 2017; Dahl et al., 2020).

From this, we might expect to find people sometimes dissociating moral valence (i.e., rightness/wrongness) and permissibility. However, there is little evidence for this dissociation. Many studies of moral judgment often use single measures and, therefore, do not shed light on the dissociations among categories of judgments (Malle, 2021). But even studies that include multiple measures have failed to find them. Cushman (2008) found that judgments of valence and permissibility are both sensitive to the same kind of mental state information. O'Hara et al. (2010) found that moral judgments of wrongness, inappropriateness, and impermissibility varied only marginally. The variation was so minor that O'Hara et al. concluded: "the influence of wording variations on moral judgments [is] negligible" (p. 552). Kneer and Machery (2019) likewise found that judgments of permissibility and valence for negligent behavior did not differ significantly in

either a between-subjects or within-subjects design. This limited evidence would suggest valence and impermissibility do not dissociate.

In line with these results, some have argued on conceptual grounds that terms like 'forbidden' or 'wrong' are probably linguistic variations of some homogenous moral category (Björklund, 2003; Cushman et al., 2006; Greene et al., 2001; Koenigs et al., 2012). Common sense moral judgment is likely not granular enough to reflect differences between being forbidden, impermissible, wrong, and so on, despite what some everyday examples or sophisticated theories might suggest. "Impermissible" and "wrong," "obligatory" and "good" are, accordingly, linguistic variations conveying a singular mode of moral evaluation.

Still, some recent evidence pushes back against this singular view of moral judgment. Voiklis et al. (2016) found that justifications for judgments of valence (i.e., goodness vs badness) and permissibility differed when evaluating responses to sacrificial dilemmas. Permissibility judgments more often appealed to consequences, while valence judgments appealed to mental agency. Dahl et al. (2020) presented participants with vignettes that depicted an agent deliberating about whether to help another individual. In situations where helping behavior would incur high cost for low benefits *or* where individuals had no relationship to those needing help, 48% of participants claimed that people *should not help* but that it would be *okay to help*. These responses were categorized as suberogatory by the researchers. However, because of how the suberogatory was operationalized (as 'something that is OK to do but should not be done'), it is unclear whether participants were making these judgments in a moral register. For example, participants claimed that it was OK for a person on crutches to help someone who has fallen over, but that the individual *should not help*. Does this mean that it is wrong, but permissible, for a person on crutches to offer help? If so, why would it be *morally* wrong to help? The measures used by Dahl et al. make the results difficult to interpret whether participants are dissociating moral constructs.

1.1. A methodological issue?

Malle (2021) offers two potential explanations for the absence of evidence for distinctions among categories of moral evaluation. First, he suggests that judgments of valence are typically made retrospectively, while judgments of permissibility are typically made prospectively. Because experimental stimuli often depict actions that have been done, participants might interpret permissibility probes as asking about valence, thereby washing out potential differences between the two. Second, Malle claims that valence and permissibility are categorical concepts, though researchers often provide continuous scales for their measurement. Thus, when asked to assess valence and permissibility as continuous variables, participants interpret them in terms of scalar constructs (e.g., blame or badness).

We believe there is an alternative diagnosis. Our hypothesis is that many researchers have not used situations that might plausibly disentangle judgments of permissibility and valence. Thus, the lack of variability among categories of moral judgment might not be a measurement issue, as Malle suggests, but an artifact of the stimuli used to elicit moral judgments. To that end, we used different situations, where individuals face choices where every option is plausibly permissible, but some seem better or worse from a moral perspective. These situations, while being recognizable from everyday life, introduce a host of competing moral considerations related to people's rights and the moral characteristics they exhibit while exercising them. In so far as these moral considerations can be pitted against each other, these dissociations become observable.

Thus, in a way, we agree with Malle that current methodology is crucially limited. However, the issues of detecting dissociations among moral judgments goes beyond methodology (whether this refers to either *materials* or *measurement*). To the extent that commonsense morality not only makes demands of many different types, but also institutes a variety of entitlements (what people

111 have a right to do), moral evaluation is sensitive enough to carve distinctions between, for example,
112 what counts as impermissible and what counts as wrong.

113 Let us be clear about this. Many moral judgments are remarkably simple: "That's bad",
114 "You're a true friend", and so on. This simplicity might indicate that moral judgments are an
115 expression of an underlying *monolithic* construct of moral propriety (either rightness or
116 wrongness). Such an assumption is implicit even among frameworks that recognize distinct
117 domains of morality, such as Shweder's Big Three (Shweder et al., 1997) or Moral Foundations
118 Theory (Haidt, 2001). For example, although Moral Foundations Theory recognizes that moral
119 evaluation reflects different concerns (encapsulated in the foundations of care, loyalty, etc.), the
120 theory characterizes moral evaluation in terms of the application of a unified concept of wrongness
121 (Graham et al., 2013). In other words, care, sanctity, and loyalty violations are wrong for different
122 reasons, but they are all still *wrong*.

123 Consider now sophisticated turns of phrase, such as "You shouldn't have done that", "You
124 weren't supposed to do that", and "You had no right to do that". Some of the surface-level
125 variability in the expressions of moral judgment corresponds to genuine variation in the *content* of
126 those judgments. That is, independently of which specific actions are referred to here, judging that
127 something shouldn't have been done is different from saying that person was not morally permitted
128 to do it. Each judgment, as the evidence we present below indicates, potentially responds to moral
129 considerations that are not just thematically different (harm vs. loyalty) but are of a different
130 normative kind.

131

132 ### *1.2.* The suberogatory and supererogatory

133 There has been ample discussion among philosophers regarding the possibility of
134 supererogatory action (Archer, 2018). People can seemingly do things that, though admirable, are

not required (e.g., volunteering at a local animal shelter). Notably, if supererogatory action is possible, then the contrary also seems possible: people can do things that are loathsome without violating an obligation (Driver, 1992; Hurd, 1998). For example, someone might not offer to proctor the exam of a sick colleague despite being available. Suberogatory behavior is wrong, but not because one fails to discharge a duty; rather, suberogatory behavior seems wrong because it manifests something negative about one's moral character.

In failing to do a supererogatory action, one need not do something wrong. 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 wrong, whereas helping is good despite not being required. People, then, sometimes encounter certain conflicts in their day-to-day experiences of morality: conflicts between equally permissible right and wrong options. These moral encounters (Monin et al, 2007) differ in their normative structure from the dilemmas typically used to study moral judgment, because every option is in principle permissible and people have the right to pursue each option (Sinnott-Armstrong, 1984; Christensen et al., 2014). However, it would be wrong to pursue some options. To this extent, using these encounters as stimuli offers a distinctive opportunity to study the granularity of moral judgment in everyday life.

The evaluation of super- and suberogatory behavior provides additional nuance in the debate over whether moral judgments are act-based or person-based. Act-based models of moral judgment claim that such judgments are primarily evaluations of actions (Cushman, 2015; Malle et al., 2014; Malle, 2021). Person-based models of moral judgment claim that such judgments are primarily evaluations of enduring states of persons (Pizarro & Tanenbaum, 2012; Uhlmann et al., 2015). If the suberogatory is represented in psychological categories of moral evaluation, this

158 would suggest that different judgments are keyed to different aspects of a situation. In this way,
159 some judgments might tend to be more act-based (e.g., permissibility) while other judgments might
160 tend to be more person-based (e.g., wrongness). In arguing for a more complex picture of moral
161 judgment, we open the possibility that different kinds of information-processing characteristics
162 underlie different forms of judgment.

163 ### *1.3.* The present study

164 The present study provides evidence that people sometimes judge wrong actions to be permissible.
165 This, in turn, suggests that folk psychological categories of moral evaluation exhibit interesting
166 dissociations that reflect relatively fine-grained distinctions among normative concepts (Bennis et
167 al., 2010; Sinnott-Armstrong, 2016).

168 This study provides insight into both the logic of moral judgment and the psychological
169 structure of moral categories. It does so by addressing a methodological limitation in current
170 research on moral judgment. Researchers typically ask participants to assess the perceived
171 normative properties of a situation in terms of a single dimension, including: disapproval (Van
172 Dillen et al., 2012), wrongness (Cheng et al., 2013; Schnall et al., 2008; Wheatley & Haidt, 2005),
173 acceptability (Young et al., 2012; Greene et al., 2001a), and blameworthiness (Siegel et al., 2017;
174 Young et al., 2010; Cushman, 2008). Even when researchers provide multiple measures, they
175 instruct participants to interpret these various measures in terms of a single construct (Kahane et
176 al. 2018, p.139). Here, we provide participants with multiple measures of moral judgment
177 (wrongness or rightness, permissibility, and obligatoriness) without presuming that these measures
178 map to the same underlying construct.

179 Experiments 1a and 1b found quantitative evidence that people distinguish between the
180 badness or wrongness of an action and its permissibility across several scenarios. In Experiment 2,

we used vignettes that described scenarios involving *harm* adapted from classic philosophical thought experiments about abortion and property rights. We found the same pattern of dissociations in judgments of badness and permissibility. In Experiments 3 and 4 we tested directional hypotheses about potential drivers of this dissociation. In Experiment 3, we found that judgments about individual rights predicted judgments of permissibility for suberogatory behavior, but do not predict judgments about valence (rightness or wrongness) or responsibility (praise or blame). In Experiment 4, we found that judgments about character predicted judgments of valence but not judgments of permissibility. This is preliminary evidence that judgments of permissibility track perceived individual rights, while judgments of wrongness track character evaluations.

We preregistered Experiments 1a, 2, 3, and 4 to clearly establish design and analysis plans and distinguish the confirmatory and exploratory aspects of our research. Materials, data, and code for all experiments are available on the OSF page of the project (https://osf.io/jp7tg/). The IRB of the Universidad de los Andes approved this study.

2. Experiment 1a

2.1. Methods

2.1.1. Participants

We recruited 311 participants through Prolific Academic (M_{age} = 32.77, SD_{age} = 11.2, 60% female). Sample size was determined through an *a priori* power analysis using G*Power software for a mixed ANOVA. We switched to using linear mixed models after collecting data given the problems of repeated measures analyses with independence and distributional assumptions (Singmann & Kellen, 2019). Our sample size, however, is consistent with 95% power to detect small effects (*d*

= 0.21) based on a two-tailed one-sample *t*-test at standard error thresholds, which is the primary analysis used in this experiment.

2.1.2. Materials and procedure

Each vignette, adapted from Driver (1992), described an individual faced with a choice between a suberogatory and a supererogatory option. Additionally, to account for possible asymmetries between actions and omissions (Haidt & Baron, 1996), we created action and omission versions of each scenario. This generated eight vignettes, described below (suberogatory versions 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.

231 She can come up front to claim her prize or she can let a cash equivalent go to a
232 hurricane relief fund." After hearing the news, Alex looked excited: "Even though I
233 have the winning ticket and Disney World sounds fun, I am going to donate the prize
234 to one of the charities" [*After hearing the news, Alex looked excited: "I have the*
235 *winning ticket! Even though I don't really care much about Disney World, I am going*
236 *to claim the prize anyway"*].
237
238 Participants were presented with an action and omission version of both suberogatory and
239 supererogatory behavior. For each vignette, participants were asked to make three judgments using
240 100-pt. sliders anchored at the midpoint:
241
242 *Permissibility*: To what extent do you consider [condition-specific behavior] to be morally
243 permissible or impermissible? (0 = Impermissible, 50 = Neither permissible nor
244 impermissible, 100 = Permissible).[1]
245 *Valence*: To what extent do you consider [condition-specific behavior] to be morally good
246 or bad? (0 = Bad, 50 = Neither good nor bad, 100 = Good).
247 *Obligatory*: To what extent do you consider [condition-specific behavior] to be optional or
248 obligatory? (0 = Optional, 50 = Neither optional nor obligatory, 100 = Obligatory).
249
250 All items were randomized across trials.
251
252 ***2.1.3. Data analysis approach***

[1] We interpreted 50 as an indifference point. Experiments 3 and 4 replicate similar patterns among different judgments using different midpoints (Unsure/Not a clear case). This suggests that participants treat the midpoints as indifference points in each experiment.

Linear mixed-effects models were fitted with the *lme4* package (Bates et al., 2015; R Core Team, 2022). Per our pre-registered analysis plan, participants and vignettes were modelled as random factors to allow generalizing beyond our specific sample and materials (Baayen et al., 2008). We calculated a model for each judgment category (valence, permissibility and obligatoriness) and entered Erogation Category, Situation Type and their interaction as fixed effects, where categorical predictors were effect-coded to be able to estimate their main effect (Singmann & Kellen, 2019). We followed a maximal-to-minimal modelling process (Barr et al., 2013) so that if a model failed to converge, we eliminated the random intercepts closer to zero (Barr et al., 2013; Brauer & Curtin, 2018; Meteyard & Davies, 2020).

We reported the fixed model estimates and pairwise comparisons using the *emmeans* package in R (Lenth, 2020) with degrees of freedom calculated with the Kenward Roger method and *p*-values corrected with the Tukey method. For Experiments 1a/b and 2, our primary analyses consist of comparing mean-centered responses to an 'indifference point'. This reflects an attempt to infer categorical claims (e.g., about what participants judge to be wrong or permissible or impermissible) from continuous data.

We do not report effect sizes for individual model terms since there is no widely accepted method of calculating them for linear mixed models. Confidence intervals for non-standardized simple differences are reported for ease of understanding and the precise structure of each model are stored in the OSF repository. All reported analyses were preregistered unless otherwise specified.

2.2. Results

Results are summarized in Table 1 and Figures 1 and 2. We centered participants' ratings around the overall mean of all scores (53.5), so that negative scores represent ratings beyond the

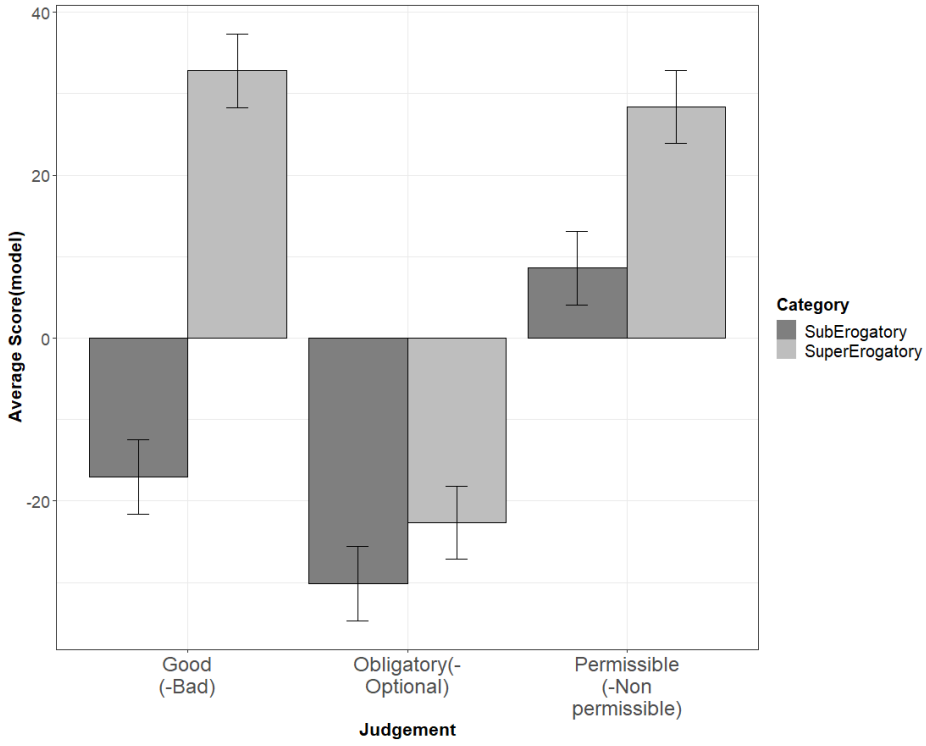
indifference point along each dimension pole. (e.g., negative permissibility scores indicate judgments of impermissibility, while positive permissibility scores indicate judgments of permissibility).

People distinguished between badness and permissibility. Suberogatory behaviors were, on average, considered bad (M_{Good} = -17.05, 95% CI [-24.32, -9.78]) and permissible ($M_{Permissible}$ = 8.58, 95% CI [1.32, 15.85]). Supererogatory behaviors, on the other hand, were considered good (M_{Good} = -32.80, 95% CI [25.39, 40.21]) and permissible (M_{Good} = 28.36, 95% CI [20.95, 35.78]). Supererogatory behaviors were rated as more permissible than suberogatory behaviors ($t(1369)$ =-13.10, $p < .001$, M_{diff} = -19.78). Evidence for similar dissociations among badness and permissibility did not emerge when participants judged sacrificial dilemmas (see Supplementary Materials §2).

Both sub- and supererogatory behaviors were considered similarly non-obligatory (Suberogatory M_{Oblig} = -22.65, 95% CI [-30.07, -15.24] and Supererogatory M_{Oblig} = - 30.12, 95% CI [-37.39, -22.86]) The effect of manipulating the type of response (action vs omission) was small and only significant for the Good/Bad dimension (See Table 1).

Figure 1

Mean Scores by Judgment Type and Erogation condition.



Note: Average scores by Erogation condition only. Error bars represent 95% confidence intervals.

Table 1.

Estimates and 95% Confidence intervals for fixed effects for Experiments 1a, 1b, and 2

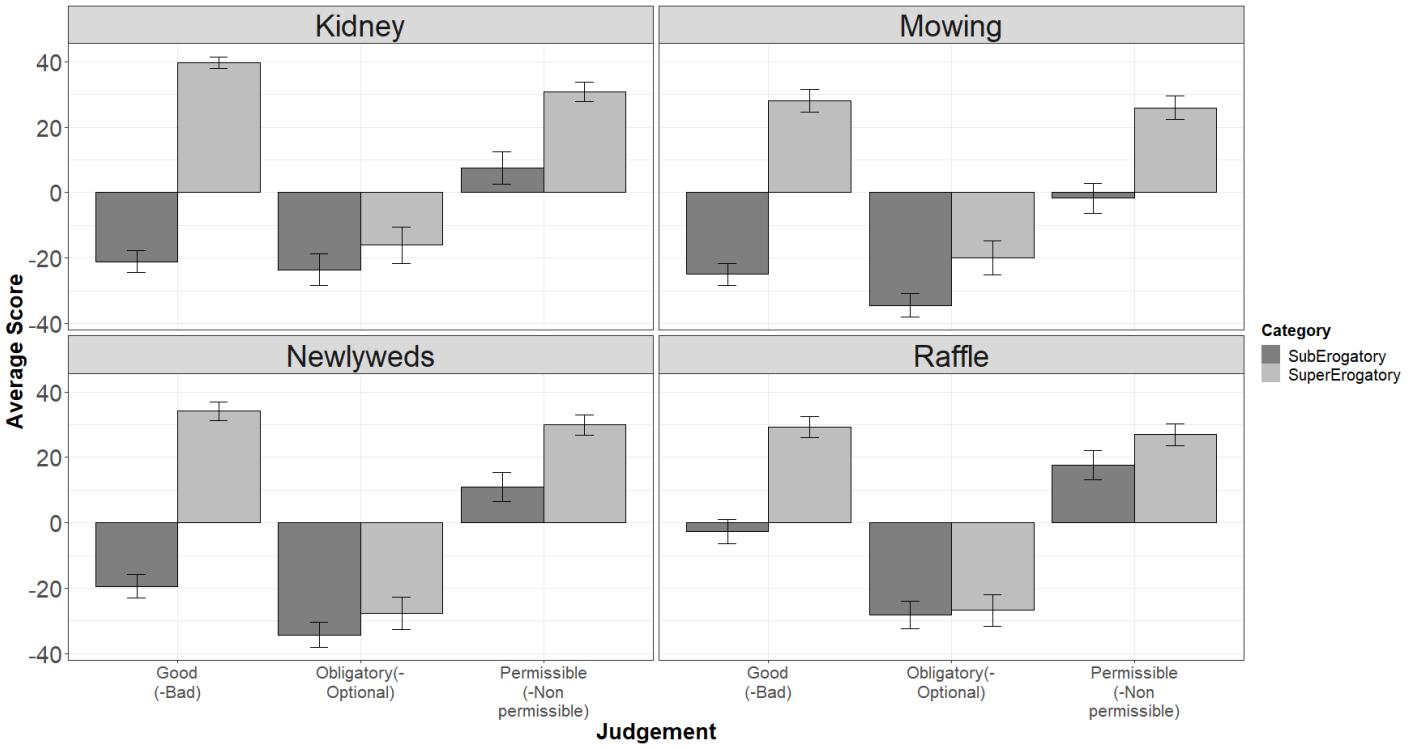
		Experiment 1a		Experiment 1b		Experiment 2	
Supererogatory vs Suberogatory	Good (Right)	49.85 47.06 – 52.64	**	29.15 26.46 – 31.84		47.58 45.06 – 50.10	**
	Permissible	19.78 16.94 – 22.62	**	15.76 13.21 – 18.31	**	21.53 18.89 – 24.17	**
	Obligatory	7.47 4.47 – 10.46	**	6.93 3.96 – 9.89	**	-0.3 -3.33 – -2.72	
Actions vs Omission	Good (Right)	7.33 5.07 – 9.59	**	9.25 6.81 –11.69	**	-- --	**
	Permissible	1.29 -2.49 – -5.06		0.25 -3.19 – 3.69		-- --	*
	Obligatory	-0.39 -4.84 – -4.06		-0.82 -4.72 – 3.08		-- --	**

Note: Interaction terms were fitted for all models with two fixed effects but are not reported since none of them were significant. Full estimates are reported in the supplementary materials for each Experiment.
*** p <0.001; *p<0.05; -- Parameter not estimated. Fixed effects represent overall difference between conditions.*

Responses varied across vignettes (see Figure 2). For example, while donating a kidney to a cousin is considered better and more permissible than not donating a kidney, the same pattern does not hold in the raffle scenario. In this case, both options are equally permissible, but donating the raffle

prize is better than not. Supererogatory responses elicit more positive evaluations, but the degree of difference might be a function of the local norms for each situation.

Figure 2. *Mean Scores by Judgment type and scenario.*



Note: Error bars represent 95% confidence intervals

2.3. ***Discussion***

We found quantifiable differences between distinct evaluative categories employed in moral judgment. Judgments of permissibility and badness dissociate for suberogatory behavior. While both types of behavior are considered permissible, supererogatory behaviors are considered good while suberogatory behaviors are considered bad. These behaviors are also non-obligatory, and superogatory behavior was evaluated more positively than suberogatory behavior was negatively.

3. Experiment 1b

In Experiment 1a, we found evidence that judgments of badness dissociate from judgments of permissibility. But it might be doubted whether these are moral judgments. Badness can apply to many different undesirable things, but *wrongness* implies the violation of a moral norm (Malle, 2021). To rule out this possibility, we conducted another study asking participants to evaluate wrongness.

3.1.Methods

3.1.1. Participants

320 participants were recruited using the same sample size rationale as Experiment 1a. 318 participants completed the task through Academic Prolific (M_{age} = 33.1, SD_{age} = 11.2, 61% female).

3.1.2. Materials and Procedure

Materials and procedure were identical to Experiment 1a with one exception: participants rated behaviors in terms of rightness or wrongness rather than goodness or badness.

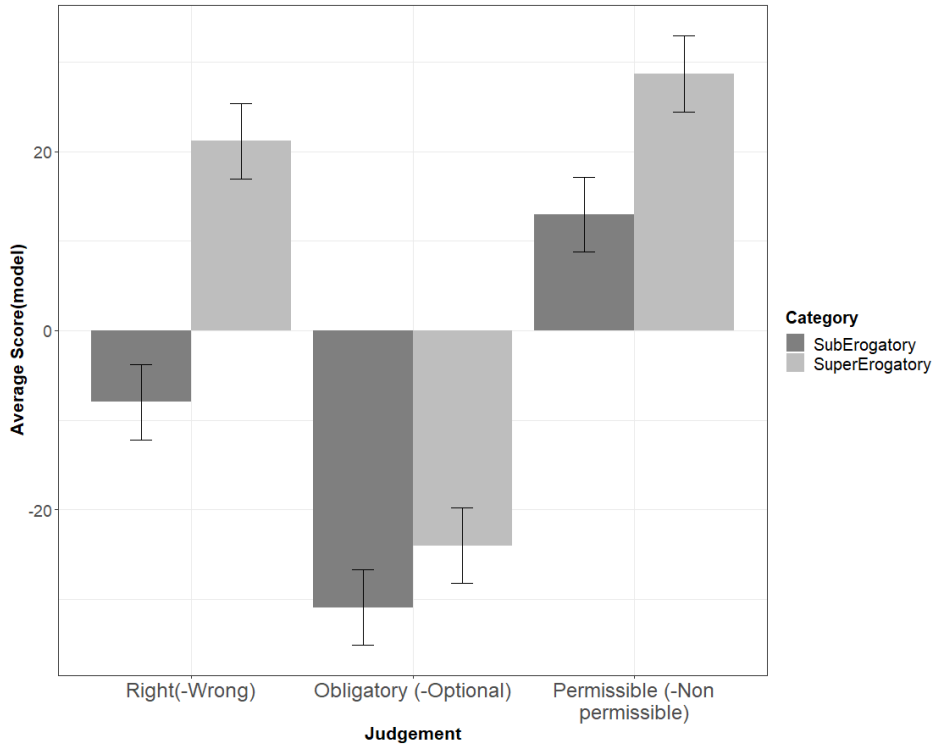
3.2.Results

Participants clearly distinguished between wrongness and permissibility for suberogatory behaviors. Suberogatory behaviors were judged to be wrong (Suberogatory M_{Right} = -7.98, 95% CI [-12.1, -3.84]), though participants also considered them permissible ($M_{\text{Permissible}}$ = 13.0, 95% CI [2.01, 23.9]); see Table 1 and Figure 4). Supererogatory behaviors were rated as significantly more permissible than suberogatory behaviors ($t(316)$ =-12.96, $p < .001$, M_{diff} = -15.8, CI [-15.50, -10.41]).

Response patterns resembled Experiment 1a. Both behaviors were considered optional to a similar degree (Suberogatory M_{Oblig} = -30.9, *SE* = 3.09, 95% CI [-42.4, -19.5] and. Supererogatory M_{Oblig} = - 24.0, *SE* = 3.21, 95% CI [-34.7, -13.3])). Actions were considered more right than omissions (See Table 1) and scenario variation was nearly identical (see supplementary materials).

Figure 3

Mean Scores by Judgment Type and Erogation condition.



Note: Error bars represent 95% confidence intervals

3.3. Discussion

In Experiment 1a, participants distinguished between the badness and permissibility of some behavior. In Experiment 1b, participants distinguished between the wrongness and permissibility

of those same behaviors. The results of these experiments suggest that supererogatory and suberogatory behaviors are both considered permissible, though supererogatory behaviors are considered right, while suberogatory behaviors are considered wrong.

But do people distinguish wrongness and permissibility for moral behaviors? Our situations might seem to pit self-interest against prosocial behavior or prudence against convenience, but they do not obviously involve harm. If moral transgressions imply that harm is caused (Gray & Schein, 2015), then perhaps suberogatory behaviors reflect prudential or conventional wrongness rather than moral wrongness.

To address this criticism, we conducted another study with two modifications. First, we used alternative vignettes that plausibly involve causing harm. Second, we included measures of praise and blame, which are prototypically treated as measures of *moral* judgments (Malle et al., 2014). If people tend to attribute blame to suberogatory behavior and praise to supererogatory behavior, then people likely view these as moral behaviors.

4. Experiment 2

4.1. *Methods*

4.1.1. Participants

We recruited 316 participants (M_{age} = 33.30, SD_{age} = 10.8, 51% female) from Academic Prolific. Sample size was set to reproduce results from Experiments 1a and 1b using a within-subjects design.

4.1.2. Materials and procedure

We constructed two new scenarios based on thought experiments from Thomson (1971) and Nozick (1974). The scenarios are described below (suberogatory version in brackets):

> **Violinist**[2]: 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 to 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 included the Newlyweds scenario from Experiments 1a and 1b. Participants saw each vignette (presented in random order). Each participant was randomly assigned to see either the supererogatory or suberogatory condition. Participants completed items used in Experiment 1a along with an item about blameworthiness (0 = praiseworthy, 50 = neither praiseworthy nor blameworthy, 100 =

[2] We used "Violinist" as a nod to Thomson's (1971) original case, which involved kidnapping someone to sustain an injured violinist. We removed references to violinists because of its well-known connection to debates about abortion.

blameworthy). Participants offered open responses to explain their ratings (as in Christensen et al., 2014), though these responses were not analyzed in the current study.

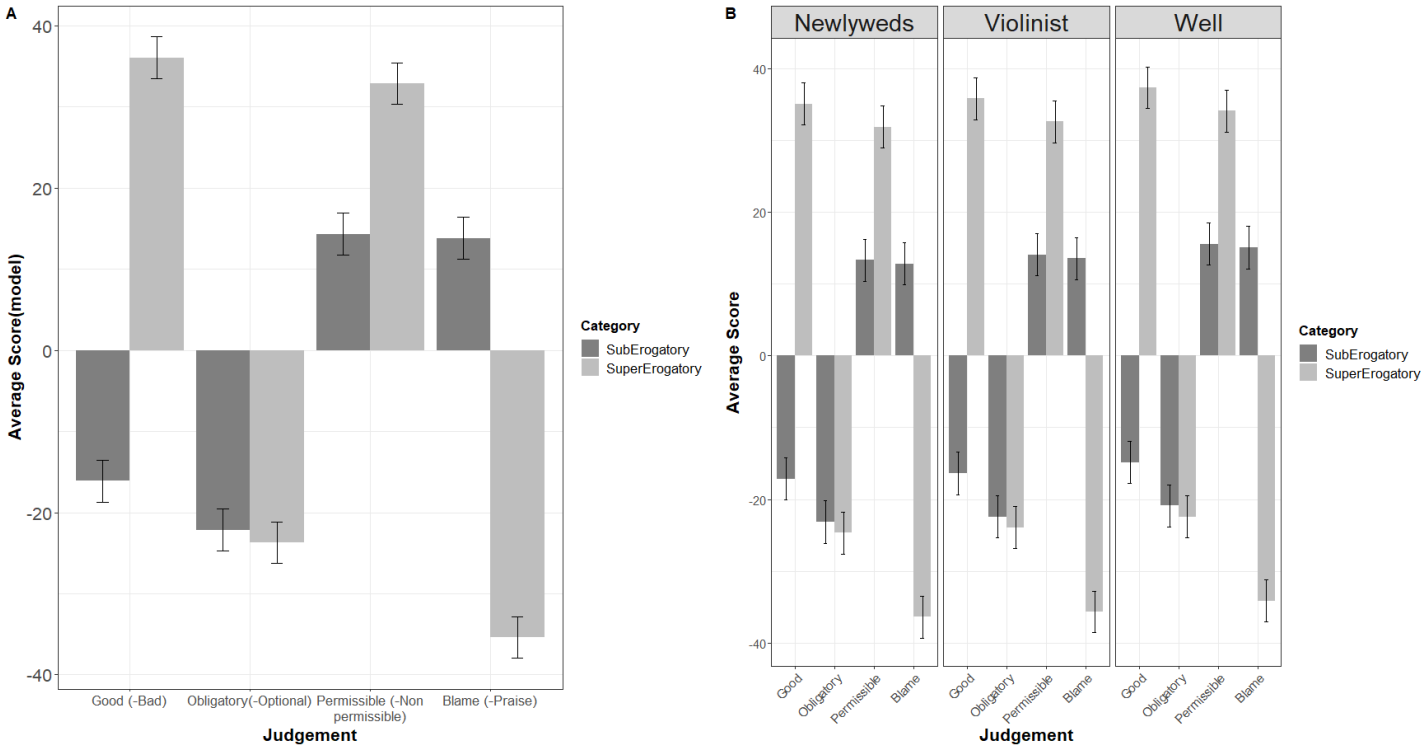
4.2.Results

Suberogatory behaviors were rated as bad (M = -16.12, 95% CI [-18.68, -13.15]) but also permissible (M = 14.33, 95% CI [11.76, 16.90]). Supererogatory behaviors were judged to be more permissible than suberogatory behaviors (M_{diff} = -18.85, $t(303)$ = -9.3, $p < .001$) but similarly non-obligatory (Supererogatory M = -22.1, 95% CI [-24.8, -19.5] vs Suberogatory M = -23.7, 95% CI [-26.3, -21.0]) (see Figure 4).

Participants also considered suberogatory behavior to be blameworthy (M = 13.82, 95% CI[11.25, 16.38]), while supererogatory responses deserved praise (M = -35.39, 95% CI [-37.93, 32.85]).

Figure 4

A. *Average scores by judgment type and condition*. B. *Average scores by Scenario*.



Note: Error bars represent 95% confidence intervals

Asymmetry for the Good/Bad scores was virtually identical for the Violinist and Well scenarios compared with the Newlyweds benchmark (Figure 4B). The same asymmetry is observed for the Blame/Praise evaluations.

5.3 Discussion

Participants again dissociated badness and permissibility even for behaviors that plausibly cause harm. Moreover, people attributed blame for suberogatory behavior and praise for supererogatory behavior, suggesting that engaging in these behaviors is viewed as warranting negative and positive personal evaluations.

5. Experiment 3

Experiments 1 and 2 show that people recognize that some wrong actions are morally permissible, indicating the dissociability of these categories. Experiment 3 examines some situational properties that might mediate this dissociation. Some moral philosophers who defend the possibility of suberogatory action suggest that such behavior reflects the morally problematic exercise of individual *rights* (Hurd, 1998). For example, people have a right to keep their seat, even when doing so is seen as rude (Driver, 1992). Likewise, people have a right to bodily autonomy, even if not relinquishing some of that autonomy would mean that another person dies (Thomson, 1971). From this, we predicted that judgments of permissibility would track judgments of individual rights.

We also found in Experiments 1 and 2 that suberogatory actions are considered blameworthy, while corresponding supererogatory actions tend to be regarded as praiseworthy. Some have argued that supererogatory behavior is morally exceptional, a positive deviation from what is frequent or more intense than expected (Lawn et al., 2022). If this is true, judgments of praise should predictably reflect underlying judgments about what we expect of others, where supererogatory behavior is considered uncommon. Conversely, suberogatory behaviors might merit blame because they fall short of our expectations.

6.1 Methods

6.1.1 Participants

We recruited 240 participants on Prolific (M_{age} = 38.65, SD_{age} = 13.9, 50% female). Because we tested new hypotheses, we did not have an estimate for effect sizes to determine sample size through a power analysis. Instead, we based sample size on our previous studies. We pre-registered our sample size before data collection and no data were analyzed prior to stopping data collection.

Per our pre-registered exclusion criteria, 3 participants were excluded for self-reported distraction during the task (N = 237). Given the number of participants and structure of the models used in our analyses, post-hoc sensitivity tests computed using G*Power software indicated that we achieved 95% power to detect medium-sized effects (f^2 = .11).

6.1.2 Materials and procedure

Materials were identical to Experiment 2. Participants viewed either the suberogatory or supererogatory version of each vignette. Situations were randomized across participants. For each situation, participants were asked to make 6 judgments using 100-pt. sliders anchored at the midpoint (midpoint = 'Unsure / Not a clear case'):

Wrong: To what extent do you consider [the behavior] to be morally right or wrong? (0 = Wrong, 100 = Right)

Obligatory: To what extent do you consider [the behavior] to be morally obligatory or optional? (0 = Obligatory, 100 = Optional)

Blame: To what extent do you consider [the behavior] to be morally praiseworthy or blameworthy? (0 = Blameworthy, 100 = Praiseworthy)

Permissibility: To what extent do you consider [the behavior] to be morally permissible or impermissible? (0 = Impermissible, 100 = Permissible)

Rights: To what extent does [person] have the right to [behave this way]? (0 = Definitely does NOT have the right, 100 = Definitely DOES have the right)

Expectation: Do you predict that people would [behave this way]? (0 = Definitely NOT, 100 = Definitely YES)

The first four questions were presented randomly and were followed by rights and expectations items. For each question, situation-relevant descriptions were provided. The *Expectation* judgment always referred to predictions about whether people would behave as the individual in the vignette does.

6.2 Results

Suberogatory behaviors were rated as wrong (M = -8.27, 95% CI[-11.6, -4.91) and permissible (M = 23.0, 95% CI[11.1, 34.8], though participants were unsure about whether these behaviors merited blame (M = 2.68, 95% CI[-5.18, 10.5]. Supererogatory behaviors were judged to be just as permissible as suberogatory behaviors (M_{diff} = 3.65, t(695) = 1.69, p = .09, d = 0.21, 95% CI[-0.03, 0.46]) and just as morally non-obligatory (M_{diff} = -1.49, t(660) = -0.69, p = .49, d = -0.06, 95% CI[-0.22, 0.16]) (see Figure 5).

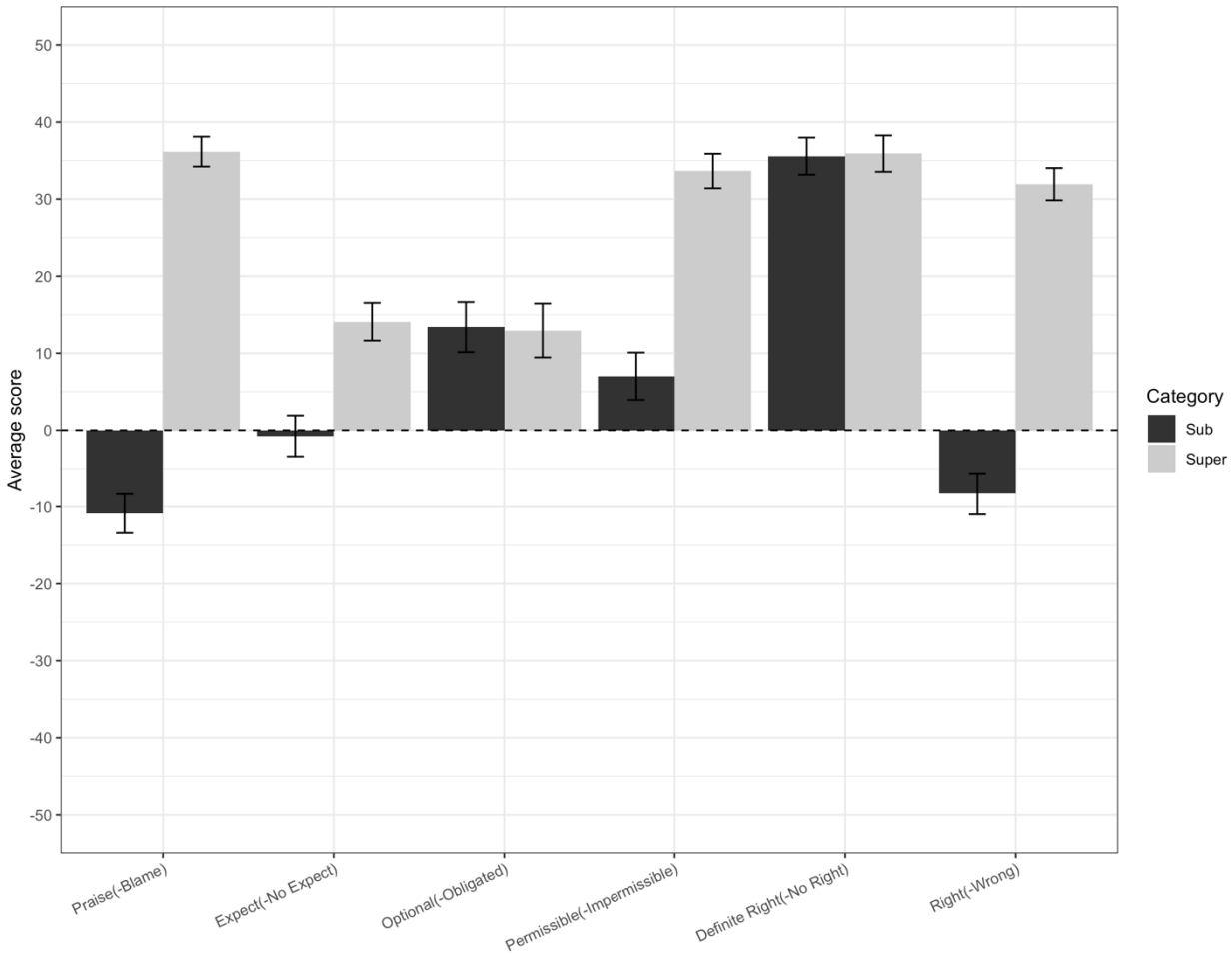


Figure 5. *Judgments by condition in Experiment 3. Error bars represent 95% confidence intervals.*

We found no evidence that participants attributed greater rights to individuals who engaged in suberogatory compared to supererogatory behavior (M_{diff} = 0.15, $t(649)$ = -0.09, p = .92, d = 0.01, 95% *CI*[-0.16, 0.17]). However, while participants were uncertain about whether others will engage in suberogatory behavior (M = -0.82, SE = 2.13, 95% *CI*[-7.54, 5.91]), they tended to expect that others would engage in supererogatory behavior (M = 14.15, SE = 2.11, 95% *CI*[7.36, 20.94]).

To assess the relationship between judgments of rights and judgments of permissibility, we computed hierarchical linear regressions to predict judgments of permissibility from judgments of rights across condition (suberogatory vs. supererogatory). The model also included a term for wrongness, blame, and their interaction, as well as interactions between blame, rights, and behavioral type. Participants and vignettes were coded as random effects.

Judgments of rights had significant partial effects in the model (β = 4.21, p < .001, 95% *CI*[2.75, 5.68], qualified by an interaction with behavioral type (β = -3.30, p = .02, 95% *CI*[-6.17, -0.43]) (see Figure 6). While increased judgments of rights predicted greater judgments of permissibility, the effect was stronger for suberogatory behavior compared to supererogatory behavior.

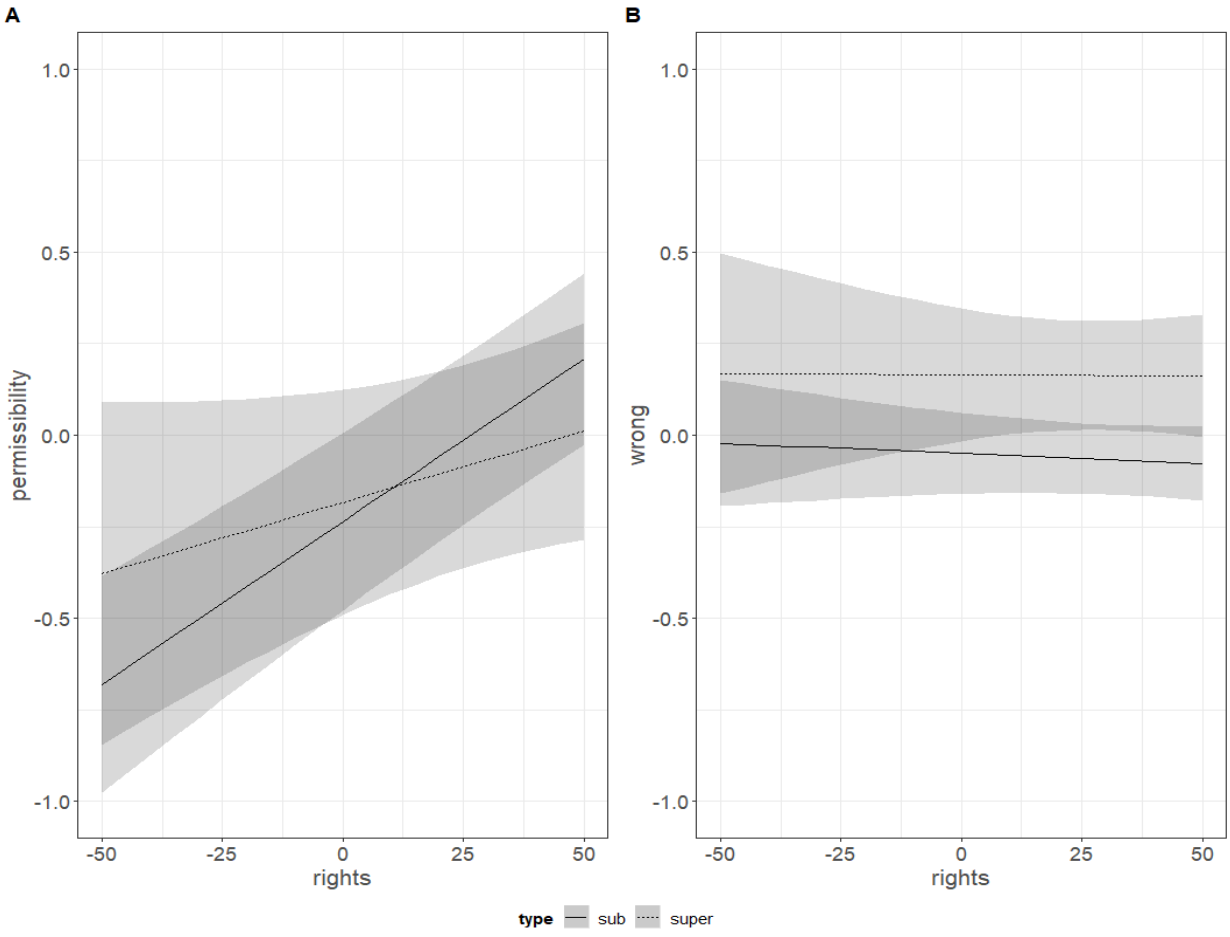


Figure 7. *Model estimates for permissibility (Panel A) and wrongness (Panel B) as a function of perceived rights for Experiment 4. Error bars represent 95% confidence intervals.*

We also computed a hierarchical linear regression to predict judgments of praise and blame from predictions of how others would behave across different behavioral types. Based on our

preregistered analysis plan, we also included terms for permissibility, individual rights, and obligation, as well as the interaction between obligation and condition.

There was no evidence for an effect of expectation on judgments of blame and praise (β = -1.15, p = .07, 95% *CI*[-2.38, 0.08]). Valence (β = 16.83, p < .001, 95% *CI*[14.90, 18.77]) and permissibility (β = 5.18, p < .001, 95% *CI*[3.43, 6.92]) both had significant partial effects in the model: as participants judged some action to be more right or permissible, they judged it to be more praiseworthy, while judging an action to be more wrong or impermissible predicted stronger judgments of blame. There was also an interaction between behavioral condition and obligation (β = 3.29, p = .01, 95% *CI*[0.80, 5.77]): As participants perceived supererogatory behavior to be more non-obligatory, they tended to attribute more praise, though judgments of blame did not change as a function of perceived obligatoriness or optionality of the behavior.

Because we failed to support the prediction that judgments of praise would be associated with varying levels of expectation about whether others would engage in supererogatory behavior, we wanted to explore further the relationship between expectation and other kinds of judgments, especially judgments of valence. The model to predict judgments of valence included all measures and interactions, with participants and vignettes coded as random effects. Expectation (β= 2.78, p < .001, 95% *CI*[1.57, 3.98]), responsibility (β = 16.42, p < .001, 95% *CI*[14.42, 18.43]), and permissibility (β = 2.78, p < .001, 95% *CI*[1.57, 3.98]) all had significant partial effects on valence: stronger expectation, greater praise, and increased judgments of permissibility all predicted stronger judgments of rightness (whereas lower expectations, greater blame, and lower judgments of permissibility all predicted stronger judgments of wrongness). There was no evidence that judgments of individual rights had significant partial effects on judgments of valence (β = -0.22,

325 p = .73, 95% *CI*[-1.44, 1.01]). There was also a significant interaction between vignette type and
326 obligation (β = -5.96, p < .001, 95% *CI*[-8.58, -3.35]).
327
328 ***6.3 Discussion***
329 Experiment 3 replicated the pattern of dissociations among judgments. These results also extend
330 the findings of Experiments 1 and 2 by providing evidence for a distinctive situational property—
331 individual rights—underlying judgments of permissibility but not judgments of valence or
332 praise/blame. Permissibility, then, is partly a function of what rights one seems to have. However,
333 these rights do not seem to inform judgments of valence or praise/blame. This provides part of a
334 sensible interpretation of what people mean when they judge that some behavior is permissible but
335 wrong: some behaviors are wrong despite it being within our rights to act in this way.
336 This leaves open the question of what people mean when they judge suberogatory
337 behaviors to be wrong and supererogatory behaviors to be right. In a preliminary experiment (see
338 Supplementary Materials §1), when participants provided open descriptions of suberogatory and
339 supererogatory behaviors, they often used character descriptions (selfish, rude, kind, generous,
340 etc.). We hypothesized that judgments of valence might track the degree to which some behavior
341 is seen as manifesting good or bad character. To test this, we conducted another experiment.
342 Our results also indicated that supererogatory behaviors were not considered to exceed
343 people's expectations. This shows that, although some morally exceptional behaviors might be
344 considered supererogatory, the supererogatory need not be regarded as exceptional.
345
346 ## *6. Experiment 4*
347 ***6.1 Methods***

348 **6.1.1 Participants**

349 260 participants were recruited on Academic Prolific. Sample size was computed using the
350 *mixedpower* package in R (Kumle, Võ, & Draschkow, 2021). Based on the coefficients of fixed
351 effects from models used in Experiment 3, we simulated 1000 models for 50, 90, 140, 180, 220,
352 260, and 300 participants. The simulation used a *t*-value of 2 as a threshold for significance. 260
353 participants provided 86% power to detect effect sizes that matched the smallest effects identified
354 in previous studies. 4 participants were excluded for failing a pre-registered attention check (N =
355 256; M_{age} = 37.52, SD_{age} = 13.5, 49% female).

356

357 **6.1.2 Materials and procedures**

358 Materials and procedure were identical to Experiment 3 with one exception: instead of an item
359 about perceived rights, participants responded to a question about character:

360 *Character*: Is [condition-specific behavior] the kind of thing a good or bad person would
361 do? 0 = A *bad person* would DEFINITELY do this, 50 = Unsure / Not a clear case;
362 100 = A *good person* would DEFINITELY do this)
363

364 ***6.2 Results***

365 Figure 7 summarizes judgments across all vignette types.

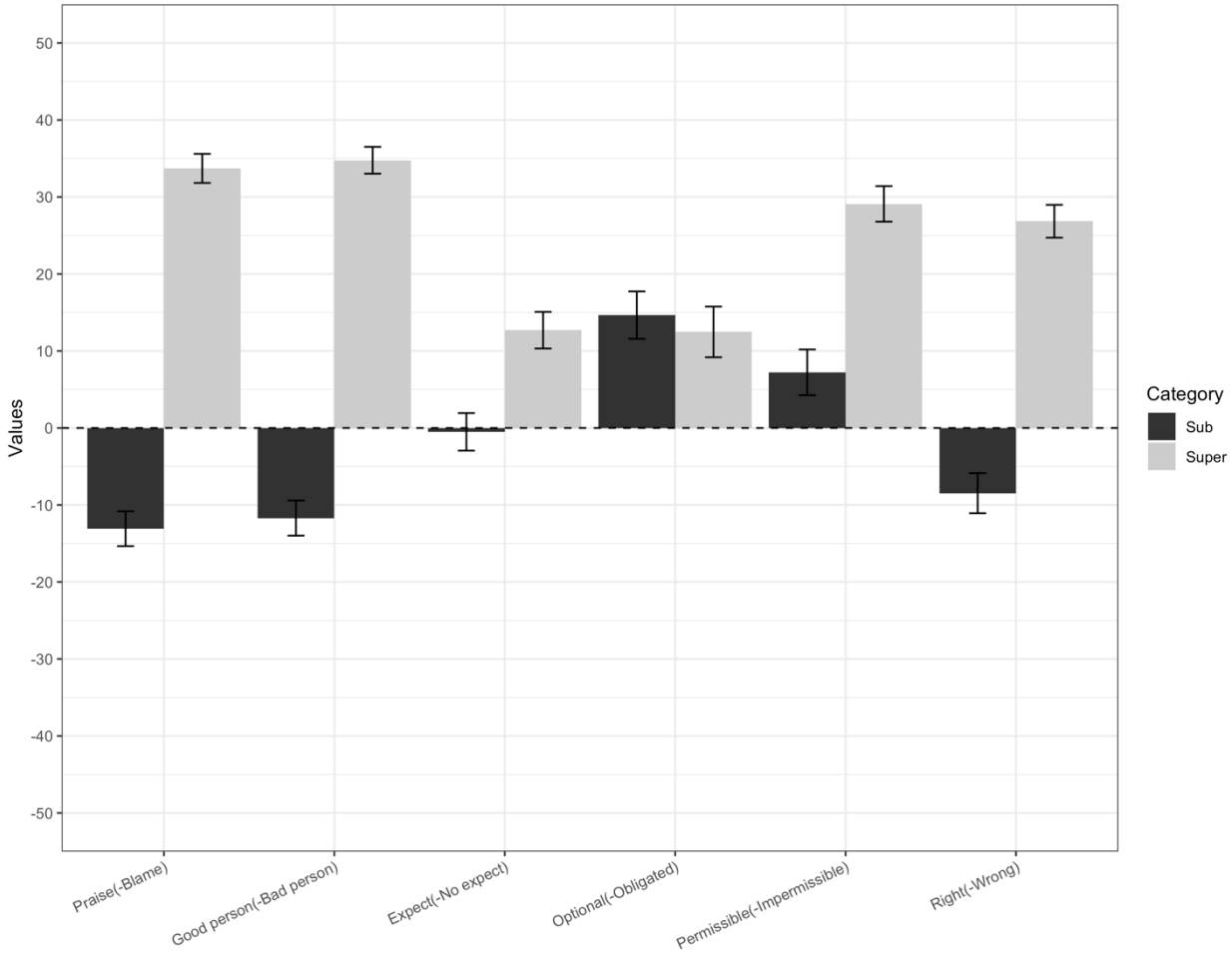


Figure 7. *Judgments by condition in Experiment 5. Error bars represent 95% confidence intervals.*

Participants judged suberogatory behaviors to be wrong, blameworthy, and permissible, while supererogatory behaviors were judged to be right, praiseworthy, and permissible. Notably, participants were more confident that supererogatory behavior is something a good person would do than that suberogatory behavior is something a bad person would do ($t(754) = -13.56$, $p < .001$, $d = -1.57$, 95% CI[-1.81, -1.33]).

To assess the relationship between judgments of character and judgments of valence, we computed hierarchical linear regressions to predict judgments of valence from judgments of character across behavioral types (suberogatory vs. supererogatory). The model also included a

term for obligation, praise/blame, permissibility, expectation, and the interactions between these and behavioral type. Participants and vignettes were coded as random effects.

Judgments of character had significant partial effects in the model ($\beta = 7.50$, $p < .001$, 95% *CI*[5.47, 9.52], qualified by an interaction with behavioral type ($\beta = 8.03$, $p < .001$, 95% *CI*[4.00, 12.05]) (see Figure 8). As people perceived behavior to be something a bad person would definitely do, the behavior was judged to be more wrong. As people perceived behavior to be something a good person would definitely do, the behavior was judged to be more right, though the effect was stronger for supererogatory behavior. Moreover, we found evidence that judgments of wrongness partially mediated the effect of character assessments on judgments of blame (see Supplementary Materials §3).

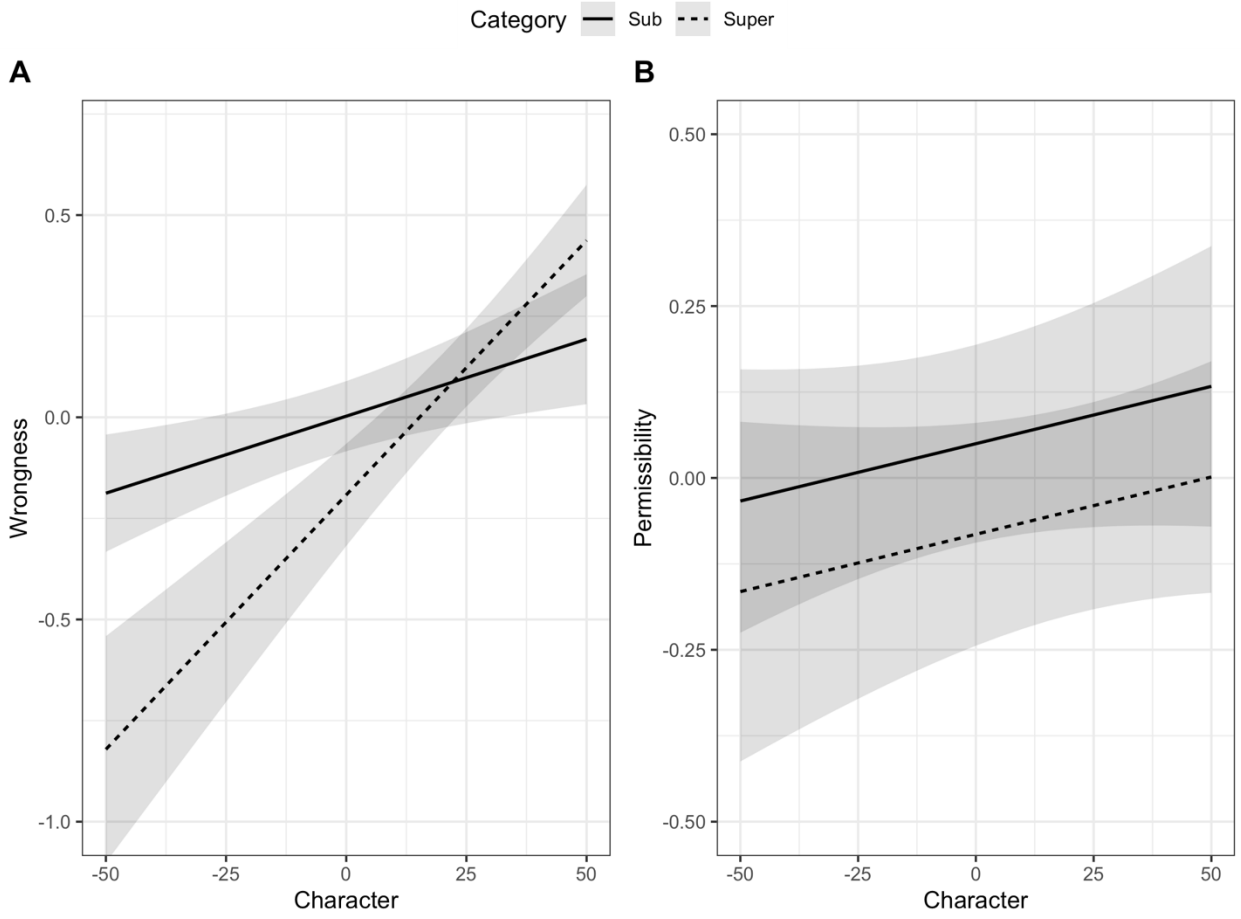


Figure 8. *Model estimates for wrongness as a function of character assessments across condition in Experiment 4. Error bars represent 95% confidence intervals.*

To assess how judgments of character are related to judgments of permissibility, we fitted a hierarchical linear model for predicting permissibility from character, behavioral type, obligation, valence, praise/blame, and expectation, as well as two-way interactions between these terms and behavioral type. Judgments of character did not have significant partial effects in the model (β = 1.48, p = .23, 95% *CI*[-0.94, 3.90]).

6.3 Discussion

Experiment 3 found evidence that judgments of individual rights predicted judgments of permissibility but not judgments of valence or praise/blame. Experiment 4 found evidence that judgments of character predicted judgments of valence but not judgments of permissibility. This partly explains why judgments of wrongness and permissibility might dissociate. Sometimes, we have the right to do things, but in exercising those rights we might manifest bad character.

7. General Discussion

To some, it seems intuitive that people are sometimes morally permitted to do what is morally wrong. Across 4 experiments, we found evidence that some situations allow for a variety of distinctions among relevant folk-conceptual categories of evaluation. These are situations in which the normative space of evaluation is defined not only by considerations about the moral valence of certain behaviors (i.e., their rightness or wrongness), but also by considerations regarding the things people have a right to do. Experiments 1 and 2 showed dissociations between permissibility,

obligatoriness, valence (rightness/wrongness), and praise/blame, even among actions with harmful consequences. Experiments 3 and 4 indicated that judgments of permissibility—but not valence—are related to perceived individual rights, while judgments of valence—but not permissibility—are related to character assessments associated with the behavior in question. This, in turn, makes sense of the responses observed in exploratory experiments (see Supplementary Materials §§1-2). People sometimes have the right to do things that manifest bad character. Having the right explains why the behavior is considered permissible, but the fact that this behavior manifests negative character traits explains why the behavior is considered wrong.

One upshot of these results is that different categories of moral evaluation track different ways of appraising a situation. This cuts against a view that differences between categories are negligible (O'Hara et al., 2010) or linguistic variations on a common underlying construct (Hauser, 2006). Interestingly, we found that permissibility and wrongness are dissociable. This raises the question of what participants communicate in making a judgment of permissibility. Our results suggest that participants use permissibility judgments to acknowledge an entitlement to engage in the behavior. However, our results do not allow further interpretation of this result. Outstanding questions include whether universalization guides permissibility judgments (Levine et al., 2022) and whether permissibility or wrongness judgments are more psychologically fundamental. These questions should be pursued in future research.

The results of Experiments 3 and 4 suggest general conditions under which wrongness and permissibility dissociate. As we found, judgments of permissibility are a function of the rights people seem to have. However, people have the right to behave in certain ways, even if the right can be exercised in a selfish or otherwise vicious manner. In these situations, we expect that

judgments of permissibility and wrongness dissociate because each is tracking different moral aspects of the situations: what the person is entitled to do versus their character.

Many of the vignettes used in our experiments involve people making decisions about what they own (raffle winnings, water, body parts, etc.) and how to exercise rights of ownership (Nichols & Thrasher, 2023). But permissible wrongdoing extends beyond how people exercise rights of ownership. Potential examples include refusing to thank a server for bringing food, or not offering support to a colleague falsely accused of wrongdoing. We predict that these are cases of permissible wrongdoing, but it is unclear whether the underlying rights concern ownership. Instead, the common thread is that each person does something that is within their autonomy to do or not.

This points to a different implication of our results. The results reported above cut against the seemingly intuitive idea that whatever is morally wrong is morally impermissible. Instead, we found that some morally wrong behaviors are allowed. This indicates a constraint on permissibility and entitlement that does not extend to wrongness. That is, people consider something wrong to the extent that it manifests some objectionable trait or motive that reflects one's concern for others. But it is permissible for people to manifest some of these traits or motives insofar as we cannot demand that people alter these traits or motives without violating their autonomy. For instance, we do not think it is obligatory to be thankful or supportive, because the point of thanking and supporting others is to do it when there is no obligation. This does not mean that *any* demands are illegitimate. We can require people not to lie, steal, or kill in cold blood. For other things that we find wrong, we can *request* of people not to do them but cannot *require* them not to do them. In other words, when you refuse to switch seats with someone, they can find other ways to appeal to

you, but they cannot make legitimate demands that you switch. For instance, they can blame them for not doing it.

Thus, we speculate that suberogatory and superogatory behaviors reveal an underappreciated dimension of the moral life. We are sometimes placed in situations where multiple options are permissible, but some are better or worse, morally speaking. To respect individual autonomy, we cannot disallow the worse options despite recognizing that pursuing such options cultivates vice, or make the better alternatives obligatory. Importantly, situations of this kind and the behaviors they afford need not be (and typically are not) considered exceptional (Lawn et al., 2022). You do nothing extraordinary when you kindly switch seats with someone, even if it was not required of you.

Finally, we focused on situations of permissible wrongdoing. But the reverse is possible, where the right thing to do is impermissible (Uhlmann et al., 2013). This reveals a different facet of the moral life: we are sometimes placed in situations where the world forces a choice between two bad options, such as a trolley driver deciding whether to kill one person or let five people die or a hospital administrator deliberating about whether to divert resources to save a patient or purchase essential resources for future operations. People tend to think that in these dilemmas, the right thing to do is maximize benefits while minimizing costs (Rosas et al., 2023). However, nobody is entitled to kill a person or intentionally divert resources from those in need. Thus, there could be situations where the right thing to do is impermissible. Between these two dissociations, it seems that we can sometimes demand that people get their hands dirty although we cannot demand that people always keep their hands clean. Such is the paradox of autonomous agents attempting to get along with each other in an imperfect world.

Our goal in this paper was to begin sketching a more complex picture of how different moral categories interact to account for observed moral judgments. To that end, we systematically tested for dissociations among these categories across a wide range of situations. We also identified distinct situational properties that are related to different kinds of judgments, which explains when these judgments are likely to dissociate.

8.1 Methodology

As noted in the Introduction, Malle (2021) suggests that the *absence* of evidence for dissociability among different categories might be driven by two factors. First, judgments of wrongness and permissibility have distinctive prototypical temporal orientations, but researchers often ask participants to evaluate situations that have already occurred. Thus, participants end up interpreting valence and permissibility in terms of the same construct. Second, some moral concepts are binary, but researchers often provide measurements in terms of scales. Possibly, participants interpret questions about permissibility and obligation in terms of valence to interpret them in scalar terms.

Our results raise questions about these conjectures. Although the scenarios used in our experiments depicted actions that had already occurred, we identified differences between judgments of permissibility and valence. Moreover, these results were robust over several different experiments. We also provided scales for participants to register different judgments and identified significant differences between ratings of obligation, permissibility, and valence, some of which Malle claimed to be binary concepts. Thus, while we agree with Malle that dissociations between different evaluative categories should be explored more systematically, we disagree with his proposal as to why research on moral judgment has so far failed to consistently find interesting dissociations among a variety of moral judgments. Rather than being primarily an issue of

measurement, we think it is an issue of the normative structure of the situations thought to be relevant to study moral judgment, some of which obviously translates into the materials used.

8.2 Moral encounters

Identifying dissociations among categories of moral judgment seems to require different kinds of stimuli than those typically used in experimental moral psychology. Researchers often use sacrificial dilemmas, such as trolley dilemmas to evoke judgments of wrongness, blame, and permissibility. However, these dilemmas often consist in pitting categorical norms against each other. Hence, depending upon which norms are endorsed by participants, they will tend to regard some as good and permissible and others as wrong and impermissible. The dissociations we found here would consequently go unnoticed.

The stimuli used in our experiments are not dilemmas, though they do invoke conflicts of a different kind. The key feature of our stimuli is that they depict situations in which the options available are permissible because acting one way or other is within people's rights. But the valence of the available options differs; choosing one as opposed to the other manifests either good or bad character. The moral conflicts presented here, therefore, are not structured around choosing between two systems of norms (e.g. deonotology vs. consequentialism). They are instead structured around the morally problematic ways we can sometimes exercise our rights. Further work should attempt to systematically vary these features of situations to better understand the causal relationships between character inference, rights, permissibility, and wrongness.

8.3 Complexity

Some psychologists have mentioned the need for using new measures in studying moral judgment (Uhlmann et al., 2015), arguing that folk-psychological categories of judgment are fundamentally

directed at personal evaluation rather than behavior evaluation. Accordingly, they argue that the content of such judgments consists mainly in aretaic rather than deontic concepts.

Our results show the importance of expanding which measures are considered relevant to study the psychology of moral judgments. People show an interest in personal evaluation when making different kinds of moral judgments, where judgments of permissibility and valence seem anchored to distinct aspects of persons. This does not show that commonsense concepts of moral evaluation are primarily aretaic, but it does show that deontic and aretaic concepts are intertwined in the production of moral judgment. Providing a complete model of how these are related and how they affect different dimensions of moral evaluation is, obviously, a task for which more evidence is required.

Some researchers have attempted to identify scenarios that elicit other dissociations of moral evaluations. Behaviors that evoke disgust or violate norms of purity are sometimes claimed to dissociate judgments of harm from judgments of wrongness (Haidt et al., 1993; Horberg et al., 2009; Mooijman et al., 2018). These disgusting behaviors are commonly claimed to fall under a unified moral foundation of Purity or Sanctity (Graham et al., 2018). However, there has been substantial discussion about whether and to what extent purity forms a coherent moral category (Gray et al., 2022; Fitouchi et al., 2023). Some have argued that judgments in this domain are primarily driven by statistical abnormality and that, controlling for these abnormalities, one finds that purity violations are no longer considered morally wrong (Gray & Keeney, 2015). This is a dispute about whether considerations of harm (a causal upshot of behavior) explains most or all of what people find wrong about some actions. Our claim is different: we are not arguing about the explanatory relationship between judgments of wrongness and the considerations motivating such judgments; rather, we are arguing about the relationship between two different kinds of judgments.

547 This is a distinct argument because one could plausibly identify suberogatory behavior under a
548 variety of different moral categories.

549 Finally, although folk conceptualizations of rights are part of commonsense morality, to
550 our knowledge there has not been any systematic attempt to explain how considerations of
551 individual rights, in particular of the rights of wrongdoers (as opposed to their victims), impact
552 moral evaluation. Our results show that once we expand the study of moral judgement to include
553 different kinds of moral encounters, these considerations might make a difference in the observed
554 judgements.

555

556 **References**

557 Archer, A. (2018). Supererogation. *Philosophy Compass* 13:3, e12476
558
559 Baayen, H., Davidson, D.J., and Bates, D.M. (2008). Mixed-effects modeling with crossed random effects for
560 subjects and itmes. *Journal of Memory and Language* 59:4, 390-412.
561
562 Barbosa, S. and Jiménez-Leal, W. (2017). It's not right but it's permitted: Wording effects in moral judgement.
563 *Judgment and Decision Making* 12:3, 308-313.
564
565 Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using
566 lme4. Journal of Statistical Software, 67(1), 1 - 48. doi:http://dx.doi.org/10.18637/jss.v067.i01
567
568 Bennis, W.M., Medin, D.L., and Bartels, D.M. (2010). The costs and benefits of calculation and moral rules.
569 *Perspectives on psychological Science* 5:2, 187-202.
570
571 Björklund, F. (2003). Differences in the justification of choices in moral dilemmas: Effects of gender, time
572 pressure and dilemma seriousness. *Scandinavian Journal of Psychology*, *44*(5), 459–466.
573
574 Brysbaert, M., & Stevens, M. (2018). Power Analysis and Effect Size in Mixed Effects Models: A
575 Tutorial. *Journal of Cognition*, *1*(1), 9. DOI: http://doi.org/10.5334/joc.10.
576
577 Chisholm, R. (1963). Supererogation and Offence: A conceptual scheme for ethics. *Ratio* 5, 1-14.
578
579 Christensen JF, Flexas A, Calabrese M, Gut NK and Gomila A (2014) Moral judgment reloaded: a moral
580 dilemma validation study. *Front. Psychol*. **5**:607. doi: 10.3389/fpsyg.2014.00607
581
582 Cushman, F. (2008). Crime and punishment: distinguishing the roles of causal and intentional analyses in
583 moral judgment. *Cognition* 108:2, 353-80.
584
585 Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral
586 judgment: Testing three principles of harm. *Psychological Science*, *17*(12), 1082–1089.

Dahl, A., Gross, R. L., & Siefert, C. (2020). Young Children's Judgments and Reasoning about Prosocial Acts: Impermissible, Suberogatory, Obligatory, or Supererogatory?. *Cognitive development*, *55*, 100908. https://doi.org/10.1016/j.cogdev.2020.100908.

Driver, J. (1992). The suberogatory. *Australasian Journal of Philosophy* 70, 286-95.

Fitouchi, L., André, J.-B., and Baumard, N. 2023. Moral disciplining: The cognitive and evolutionary foundations of puritanical morality. *Behavioral and Brain Sciences*.

Gray, K., & Keeney, J. E. (2015). Impure or Just Weird? Scenario Sampling Bias Raises Questions About the Foundation of Morality. *Social Psychological and Personality Science*, *6*(8), 859–868. https://doi.org/10.1177/1948550615592241

Gray, K., DiMaggio, N., Schein, C., & Kachanoff, F. (2022). The Problem of Purity in Moral Psychology. *Personality and Social Psychology Review*.

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., & Baron, J. (1996). Social roles and the moral judgement of acts and omissions. European Journal of Social Psychology, 26(2), 201–218. https://doi.org/10.1002/(SICI)1099-0992(199603)26:2<201::AID-EJSP745>3.0.CO;2-J

Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology*, *65*(4), 613.

Heyd, D. 1982. *Supererogation: Its status in ethical theory* (Cambridge: Cambridge University Press).

Horberg, E. J., Oveis, C., Keltner, D., & Cohen, A. B. (2009). Disgust and the moralization of purity. *Journal of Personality and Social Psychology*, *97*(6), 963–976. https://doi.org/10.1037/a0017423

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.

Hurd, H. (1998). Duties Beyond the Call of Duty. *Jahrbuch Für Recht Und Ethik / Annual Review of Law and Ethics, 6*, 3-39.

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological methods*, 15(4), 309.

Judd, C.M., Westfall, J., and Kenny, D.A. 2017. Experiments with more than one random factor: designs, analytic models, and statistical power. *Annual Review of Psychology* 68, 601-25.

Kneer, M. and Machery, E. 2019. No luck for moral luck. *Cognition* 182, 331-48.

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

Kumle, L., Võ, M. L., & Draschkow, D. (2021). Estimating power in (generalized) linear mixed models: an open introduction and tutorial in R. *Behav Res*. doi:10.3758/s13428-021-01546-0

Lawn, E. C., Smillie, L. D., Pacheco, L. B., & Laham, S. M. (2022). From ordinary to extraordinary: A roadmap for studying the psychology of moral exceptionality. *Current opinion in psychology*, *43*, 329-334.

Lenth, R.V. 2021. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.6.1. https://CRAN.R-project.org/package=emmeans

Malle, B.F. 2021. Moral judgments. *Annual Review of Psychology* 72, 293-318. Doi:10.1146/annurev-psych-072220-104358.

Malle, B.F., Guglielmo, S., and Monroe, A.E. (2014). A theory of blame. *Psychological Inquiry* 102:4, 661-84.

Malle, B.F., Scheutz, M., Arnold, T., Voiklis, J., and Cusimano, C. 2015. Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI'15*), pp. 117-24. New York: ACM.

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

Mooijman, M., Meindl, P., Oyserman, D., Monterosso, J., Dehghani, M., Doris, J. M., & Graham, J. (2018). Resisting temptation for the good of the group: Binding moral values and the moralization of self-control. *Journal of Personality and Social Psychology*, *115*(3), 585–599. https://doi.org/10.1037/pspp0000149

Muñoz, D. (2021). Three paradoxes of supererogation. *Noûs* 55:3, 699-716

Nichols, S., & Thrasher, J. (2023). Ownership and convention. *Cognition*, *237*, 105454. Advance online publication. https://doi.org/10.1016/j.cognition.2023.105454.

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

O'Hara, R.E., Sinnott-Armstrong, W., and Sinnott-Armstrong, N.A. (2010). Wording effects in moral judgments. *Judgment and Decision-Making* 5:7, 547-54.

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/

Singmann, H., & Kellen, D. (2019). An Introduction to Mixed Models for Experimental Psychology. In D. H. Spieler & E. Schumacher (Eds.), *New Methods in Cognitive Psychology* (pp. 4–31). Psychology Press.

Sinnott-Armstrong, W. (1988). *Moral dilemmas* (London: Blackwell).

Sinnott-Armstrong, W. (2016). The disunity of morality. In S.M. Liao (ed.) *Moral brains: The neuroscience of morality* (Oxford: Oxford University Press), 331-54.

Thomson, J.J. (1971). A defense of abortion. *Philosophy and Public Affairs* 1, 47-66.

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

Ullmann-Margalit, E. 2011. Considerateness. *Iyyun* 60, 205-44.

Voiklis, J., Kim, B., Cusimano, C. and Malle, B.F. 2016. "Moral judgments of human vs. robot agents," *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 775-780, doi: 10.1109/ROMAN.2016.7745207.

Westfall, J., Judd, C., and Kenny, D.A. 2015. Replicating studies in which samples of participants respond to samples of stimuli. *Perspectives on Psychological Science* 10:3, 390-99.