The Moral Dilution Effect: Irrelevant Information Influences Judgments of Moral Character

Cillian McHugh[1] & Eric R. Igou[1]

[1] University of Limerick

Author Note

Correspondence concerning this article should be addressed to Cillian McHugh, University of Limerick, Limerick, Ireland, V94 T9PX. E-mail: cillian.mchugh@.ul.ie

Abstract

Across five studies we investigated the moral dilution effect

Keywords: Moral judgment; moral categorization; dilution effect; representativeness heuristic; typicality; MJAC

Word count: TBC

The Moral Dilution Effect: Irrelevant Information Influences Judgments of Moral Character

Imagine a friend is telling you about a time they were mugged on holiday. Fortunately, a by-stander saw the incident and was very helpful to your friend afterwards. Their description of the experience includes much detail, including non-relevant information regarding both the mugger and their helper, such as, "the mugger was wearing grey shoes", "they [the helper] lived on the South of the City". As you listen to the story you will likely form an impression of the moral character of both the mugger and the helper. Conventional wisdom suggests that this non-relevant information should not impact your evaluation of either the mugger or the helper, however research suggests this may not be the case.

The dilution effect occurs when the presence of non-diagnostic information leads to judgments that are less extreme than they would have been in the absence of non-diagnostic information (Nisbett, Zukier, & Lemley, 1981; Zukier, 1982). Applied to the above example, the presence of non-diagnostic information ("grey shoes" / "lived on the South of the City") could lead judgments of the mugger to be less harsh, and judgments of the helper to be less positive. The effect has been observed for a range of judgments, including judging products (Igou & Bless, 2005; Meyvis & Janiszewski, 2002), probability judgments (LaBella & Koehler, 2004), and predictions relating to people's behavior (Nisbett et al., 1981; Zukier, 1982), however, to our knowledge, research has not directly tested if the dilution effect occurs for moral judgments.

In a classic demonstration of the dilution effect participants were presented with descriptions of target students and asked to estimate the grade point average (GPA) of these target students (Zukier, 1982). Descriptions that included non-diagnostic information (i.e., information that was not correlated with GPA – e.g., "has 1 brother and 2 sisters"), produced less extreme GPA estimates than descriptions that contained only diagnostic information i.e., information that is normally correlated with GPA. This finding held true

for descriptions suggesting low GPA (e.g., “He quite often starts things he doesn’t finish”) and for descriptions suggesting high GPA (e.g., “He never arrives late to appointments or meetings”).

The effect has been explained as emerging as a consequence of the representativeness heuristic (Kahneman & Tversky, 1972; Nisbett et al., 1981). According to this view, if all the information a person has about a target is information that is relevant to particular category membership (diagnostic information), the target will be perceived as being similar to what is representative or typical of that category. Crucially, not only is all the available information indicative of category membership, the absence of any non-diagnostic information means there is nothing to suggest any differences between the target and a typical (or stereotypical) member of the category. As such, the target is perceived as highly representative, or typical, of category membership. When information that is not relevant for category membership (non-diagnostic information) is included, this reduces the match between a target and a typical member of the category, thus reducing the perceived representativeness of the target, and this leads to the dilution effect being observed. For example, a person described as having little interest in political or social issues, with hobbies that include home carpentry and mathematical puzzles, is more likely to be categorized as an engineer than as a lawyer (Nisbett et al., 1981). In contrast, being Catholic is not representative of membership of the category engineer, and the inclusion of this information in relation to a target would reduce the similarity between the target and a typical engineer, thus reducing the representativeness of the target and resulting in the dilution effect (Nisbett et al., 1981).

Predicting the Moral Dilution Effect

Research in moral psychology increasingly highlights the need to understand the dynamic and context-sensitive nature of moral judgments (McHugh, McGann, Igou, & Kinsella, 2022). Examples of known contextual influences on people’s moral judgments

include emotions (Cameron, Payne, & Doris, 2013; Giner-Sorolla, 2018), intentionality and evitability (Christensen, Flexas, Calabrese, Gut, & Gomila, 2014; Christensen & Gomila, 2012), and how 'up close and personal' an action is (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001). We propose the presence or absence of non-diagnostic information presents another possible source of variability in moral judgments, such that the dilution effect may be observed for moral judgments. Indeed, there is a strong theoretical case for predicting that the dilution effect should be observed in the moral domain.

Previous work suggests that typicality is an important consideration when people make judgments about moral issues, that is, some behaviors are more typical (or representative) examples of wrongness or rightness than others (Gray & Keeney, 2015; McHugh et al., 2022; Schein & Gray, 2018). Some authors have attempted to identify the content of what is representative of moral wrongness, arguing that this prototype, or representative essence of moral wrongness involves "an intentional agent causing damage to a vulnerable patient" (Schein & Gray, 2018), p. 33]. While not directly discussed by proponents of this approach it is plausible that the inclusion of non-diagnostic information may reduce the match between a target and this prototype leading to the dilution effect being observed for moral character judgments (Gray, Waytz, & Young, 2012; Schein & Gray, 2018).

More recently, the theory of moral judgment as categorization (MJAC, McHugh et al., 2022) presents a more dynamic and context sensitive approach to understanding moral judgments. Two core predictions of MJAC are (i) that moral judgments are sensitive to a range of contextual factors, and (ii) that judgments of both moral rightness and moral wrongness will vary according to typicality. While context effects in moral judgment have been widely shown, McHugh et al. (2022) note that in the moral domain typicality may be confounded with severity posing a significant challenge to testing this prediction. For instance, murder is likely a highly typical example of a member of the category morally wrong, while stealing stationary is a less typical example, however this variation in

typicality cannot be separated from the difference in the severity of the actions. The dilution effect paradigm provides a means to test for this variability in typicality in moral judgments, while avoiding the confound of severity. Applying the same reasoning as described above in relation to the representative heuristic (Kahneman & Tversky, 1972; Nisbett et al., 1981), suggests that for moral categorizations, the presence of information that reduces the similarity between a target and an action/actor that is prototypically right or prototypically wrong (i.e., non-diagnostic information), should lead to the target being evaluated as less typical (or less representative), and this should lead to less extreme evaluations of the target. Thus, a moral dilution effect should be observed.

The Current Research

Informed by recent work that explains moral judgment as occurring through the same cognitive processes as categorization more generally (McHugh et al., 2022), we predict that the dilution effect should be observed for moral judgments. We present three studies where we test for the dilution effect in judgments of moral character. In Study 1 we investigate descriptions of bad actors. In Study 2, we investigate descriptions of good actors, and in Study 3 we investigate descriptions of both good and bad actors. The contribution of the current work is twofold. First, we are (to our knowledge) the first to test for the dilution effect in the moral domain. Second, we provide an empirical test of a core hypothesis of the recently proposed MJAC theory of moral judgment.

All three studies 1-3 were pre-registered. A-priori power analyses revealed that in order to detect a small effect (f 2 = .01) for Studies 1-3, a minimum sample of N = 785 was required. As such, for each Study our target minimum sample size was N = 800. In the supplementary materials we report two pilot studies that informed the development of the stimulus materials used.[1]

[1] We also report in the supplementary materials three additional studies with similar designs to Studies 2 (Study S1) and 3 (Study S2), there is also a between-subjects version of Study 3 (Study S3). These Studies

Ethical Declarations

All procedures performed in studies involving human participants were approved by institutional research ethics committee and conducted in accordance with the Code of Professional Ethics of the Psychological Society of Ireland, and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. All studies were approved by the ethics committee of the Faculty of Education and Health Sciences at the University of Limerick (Education and Health Sciences Research Ethics Committee: EHSREC), and the project approval number is 2020_12_06_EHS. Informed consent was obtained from all individual participants included in the study.

Study 1 - Bad Characters

The aim of Study 1 is to test if the dilution effect exists in the moral domain. Participants were presented with descriptions of four characters, two descriptions will only contain diagnostic information (morally relevant information) and two will additionally contain non-diagnostic information (non morally relevant information) along with the diagnostic information. We hypothesize that moral perceptions of the diagnostic only descriptions will be more severe than for the descriptions that also contain non-diagnostic information.

Study 1: Method

Study 1: Participants and design

Study 1 was a within-subjects design. The independent variable was condition with two levels, diagnostic information only (diagnostic), and non-diagnostic information additionally included (non-diagnostic). We used the same two dependent variables as in

---

S1-S3 were conducted on MTurk and due to irregularities with the quality of the data they are not reported in the main text. Because of this, for the studies reported in the main text our samples are primarily drawn from student populations.

Pilot Study 1, the four item moral perception scale (MPS-4) which showed good reliability, $\alpha = 0.83$, and the single item moral perception measure MM-1.

A total sample of 851 (526 female, 303 male, 14 non-binary, 5 other; 3 prefer not to say, $M_{\text{age}} = 26.11$, min = 18, max = 76, $SD = 10.14$) started the survey. Participants were recruited from the student population at University of [BLINDED].

Participants who failed both manipulation checks were removed ($n = 249$), leaving a total sample of 801 participants (496 female, 283 male, 14 non-binary, 5 other, 3 prefer not to say; $M_{\text{age}} = 26.25$, min = 18, max = 76, $SD = 10.20$).

Study 1: Procedure and materials

As in the pilot study, data were collected using an online questionnaire presented with Qualtrics (www.qualtrics.com). Participants were presented with four descriptions of characters (*Sam*, *Alex*, *Francis*, *Robin* from Pilot Study 1). All descriptions included diagnostic information relating to three moral foundations, e.g., *Imagine a person named Robin. Throughout their life they have been known to physically hurt others, treat some people differently to others, and show lack of loyalty.* We programmed our survey to randomly present non-diagnostic information along with two of the descriptions participants read (this was done through blocking, for details on the blocks see full materials at https://osf.io/mdnpv/?view_only=77883e3fbc3d45f1a35fe92d5318cb67. This meant that all participants read two descriptions containing diagnostic information only, and two descriptions that additionally included non-diagnostic information. We hypothesized that the descriptions including non-diagnostic information would be rated as less severe than the diagnostic-only descriptions. Study 1 was pre-registered at https://aspredicted.org/DVY_QN3

Study 1: Results

Prior to conducting the main analysis, we conducted a preliminary test of the data quality, examining the extent to which participants' ratings of the characters deviated from what would be expected based on the character descriptions. All characters were described as ostensibly *bad* characters, and therefore responses suggesting these characters are *good* (as measured by scoring above the midpoint of either measure) would be surprising, and may indicate measurement error or inattentiveness. All $N = 801$ participants responded to four descriptions resulting in a total of 3204 responses for each measure. For MPS-4, 142 (4.43%) responses were above the midpoint, and for MM-1, 154 (4.81%) were above the midpoint. These responses present a potential source of error (e.g., inattentiveness), however the observed proportions are relatively low, and do not justify deviating from our pre-registered exclusion criteria (failing both attention checks), and we proceed with the analyses as planned.

We conducted a linear-mixed-effects model to test if condition influenced MPS-4 responses. Our outcome measure was MPS-4, our predictor variable was condition; we allowed intercepts and the effect of condition to vary across participants, and scenario was also included in the model. Overall, the model significantly predicted participants responses, and provided a better fit for the data than the baseline model, $\chi^2(8) = 816.91$, $p < .001$. Condition significantly influenced responses to the MPS-4, $F(1, 799.42) = 51.47$, $p < .001$; and was a significant predictor in the model when controlling for scenario, $b = -0.08$ ($b_{\text{standardized}} = -0.09$), $t(799.42) = -7.17$, $p < .001$, with the diagnostic descriptions being rated as more immoral than the non-diagnostic descriptions ($d = -0.16$), see Figure 1.

We conducted a linear-mixed-effects model to test if condition influenced MM-1 responses. Our outcome measure was MM-1, our predictor variable was condition; we allowed intercepts and the effect of condition to vary across participants. Overall, the model significantly predicted participants responses, and provided a better fit for the data

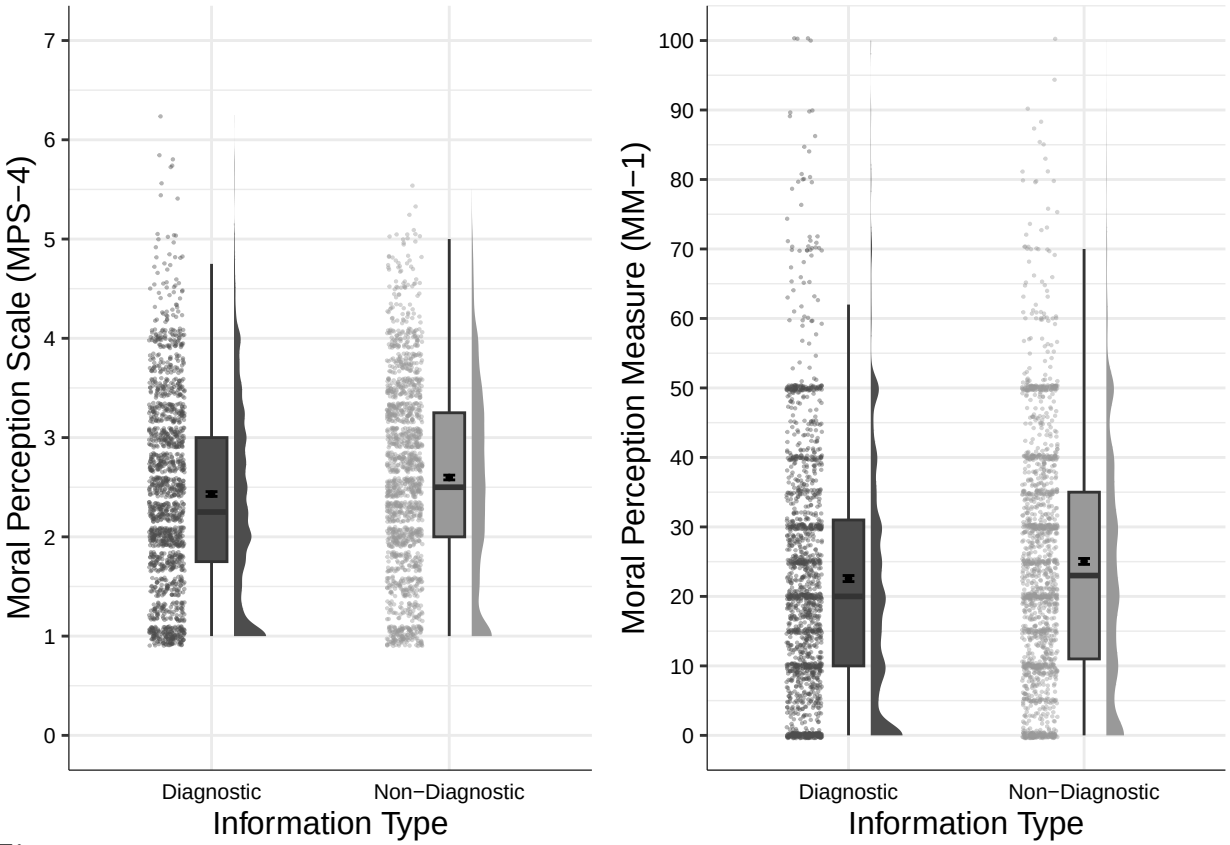

Figure 1

Study 1: Differences in moral perception depending on condition

than the baseline model, $\chi^2(8) = 475.52$, $p < .001$. Condition significantly predicted MM-1 responses $F(1, 799.71) = 44.39$, $p < .001$, and when controlling for scenario was a significant predictor in the model b = -1.22 ($b_{\text{standardized}}$ = -0.07), $t(799.71)$ = -6.66, $p <$ .001, with the diagnostic descriptions being rated as more immoral than the non-diagnostic descriptions ($d = 0.16$), see Figure 1.

In the supplementary analyses we report the effect of condition on moral perception for each description individually.

Study 2 - Good Characters

The aim of Study 2 is to test if the dilution effect exists in the moral domain for judgments of morally *good* characters. Participants were presented with descriptions of four

characters, two descriptions contain diagnostic information only (morally relevant information) and two will additionally contain non-diagnostic information (non morally relevant information) along with the diagnostic information. We hypothesize that moral perceptions of the diagnostic only descriptions will be more extreme (more moral) than for the descriptions that also contain non-diagnostic information.

Study 2: Method

Study 2: Participants and design

Study 2 was a within-subjects design. The independent variable was condition with two levels, diagnostic and non-diagnostic. We used the same two dependent variables as in previous studies, the four item moral perception scale (MPS-4, $\alpha = 0.85$), and the single item moral perception measure MM-1.

A total sample of 1068 (557 female, 418 male, 13 non-binary, 2 other; 4 prefer not to say, $M_{\text{age}} = 29.04$, min = 18, max = 74, $SD = 10.66$) started the survey. Participants were recruited from the student population at University of [BLINDED].

The majority of participants were from the student body: $n = 533$, (female = 370, male = 147, non-binary/other = 14, prefer not to say 3, $M_{age} = 25.50$, $SD = 9.60$).

In order to reach our pre-registered target sample size we recruited additional participants from MTurk: $n = 287$, (female = 96, male = 190, non-binary/other = 1, prefer not to say 1, $M_{age} = 35.70$, $SD = 10.10$). Participants from MTurk were paid \$0.40 for their participation.

Participants who failed both manipulation checks were removed ($n = 248$), leaving a total sample of 820 participants (466 female, 337 male, 2 other, 2 prefer not to say; $M_{\text{age}} = 29.03$, min = 18, max = 74, $SD = 10.92$).

Study 2: Procedure and materials

Again, data were collected using an online questionnaire presented with Qualtrics (www.qualtrics.com). Participants were presented with four descriptions of characters (*Sam*, *Alex*, *Francis*, *Robin* from Pilot Study 2). All descriptions included diagnostic information relating to three moral foundations, e.g., *Imagine a person named Alex. Throughout their life they have been known to protect and provide shelter to the weak and vulnerable, uphold the rights of others, and show respect for authority.* For each participant, two descriptions additionally included non-diagnostic information (this was randomized through blocking, see https://osf.io/mdnpv/?view_only=77883e3fbc3d45f1a35fe92d5318cb67. Study 1 was pre-registered at https://aspredicted.org/NX2_HN6

Study 2: Results

Similar to Study 1 we conducted a preliminary test of the quality of the data, examining the extent to which participants rated these ostensibly *good* characters as *bad* (or as measured by responses falling below the midpoint of either measure). All $N = 820$ participants responded to four descriptions resulting in a total of 3280 responses for each measure. For MPS-4, 128 (3.90%) responses were below the midpoint, and for MM-1, 59 (1.80%) were below the midpoint. Again, while this is a potential source of error, based on the low proportions we chose not to introduce any additional exclusions.

We conducted a linear-mixed-effects model to test if condition influenced MPS-4 responses. Our outcome measure was MPS-4, our predictor variable was condition; we allowed intercepts and the effect of condition to vary across participants, and scenario was also included in the model. Overall, the model significantly predicted participants responses, and provided a better fit for the data than the baseline model, $\chi^2(8) = 160.00$, $p < .001$. Condition did not influence responses to the MPS-4, $F(1, 838.12) = 0.24$, $p = .624$; and was not a significant predictor in the model when controlling for scenario, $b = 0.00$

($b_{\text{standardized}} = 0.00$), $t(838.12) = 0.49$, $p = .624$, ($d = 0$), see Figure 2.

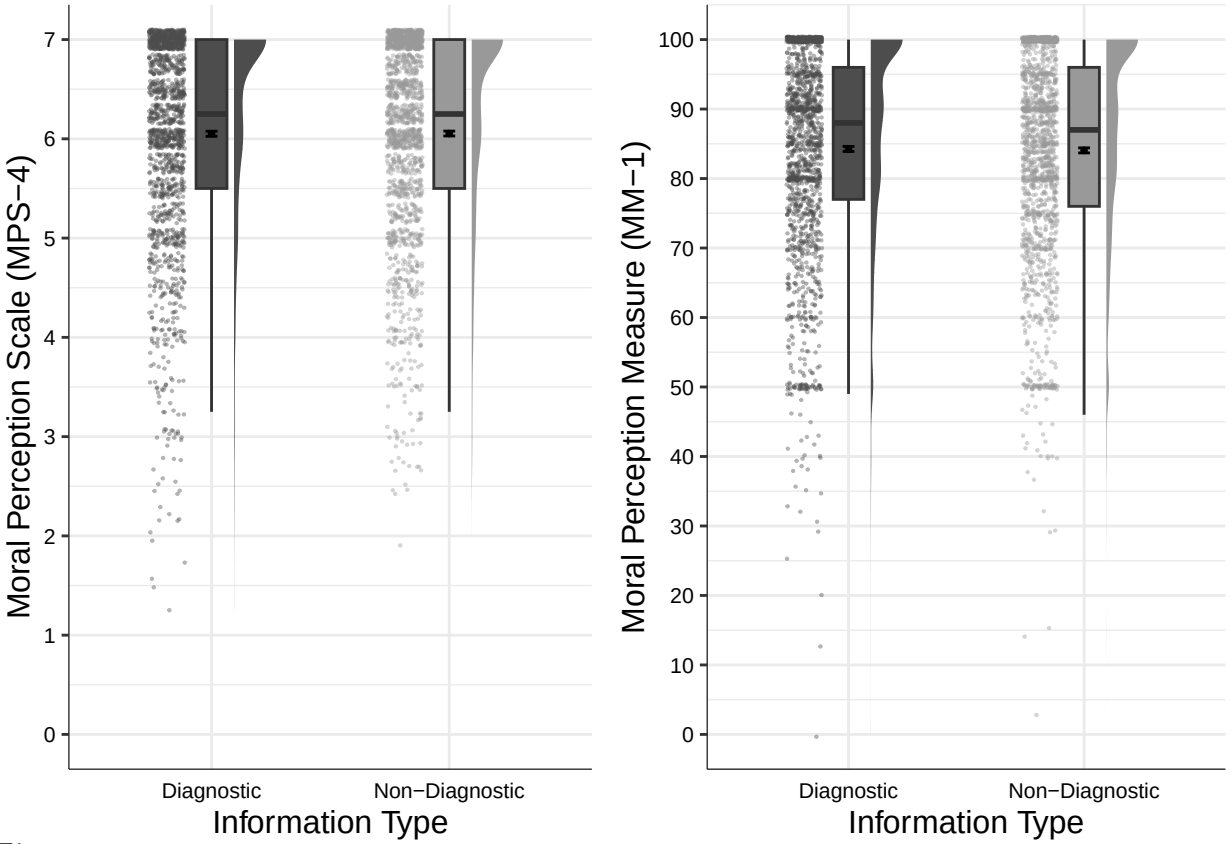

Figure 2

Study 2: Differences in moral perception depending on condition

We conducted a linear-mixed-effects model to test if condition influenced MM-1 responses. Our outcome measure was MM-1, our predictor variable was condition; we allowed intercepts and the effect of condition to vary across participants. Overall, the model significantly predicted participants responses, and provided a better fit for the data than the baseline model, $\chi^2(8) = 75.69$, $p < .001$. Condition did not influence MM-1 responses $F(1, 2453) = 1.23$, $p = .267$, and was not a significant predictor in the model $b = 0.16$ ($b_{\text{standardized}} = 0.01$), $t(2453) = 1.11$, $p = .267$, ($d = 0.02$), see Figure 2.

In the supplementary analyses we report the effect of condition on moral perception for each description individually.

Study 3 - Good and Bad Characters

In Study 1 we found evidence for the moral dilution effect for judgments of *bad* moral characters. In Study 2 we failed replicate this effect for judgments of *good* moral characters. The aim of Study 3 was to test if valence (good vs. bad) moderates the moral dilution effect. We hypothesized that valence (good vs bad) would interact with condition in producing a dilution effect, such that the dilution effect would be observed for bad characters but not for good characters. Study 3 was pre-registered at https://aspredicted.org/QDF_XT1.

Study 3: Method

Study 3: Participants and design

Study 3 was a 2 × 2 within-subjects factorial design. The first independent variable was condition with two levels, diagnostic and non-diagnostic. The second independent variable was valence of character description, with two levels morally good and morally bad. We used the same two dependent variables as in previous studies, the four item moral perception scale (MPS-4, $\alpha = 0.97$), and the single item moral perception measure MM-1.

A total sample of 1386 (535 female, 758 male, 10 non-binary, 2 other; 11 prefer not to say, $M_{\text{age}} = 29.67$, min = 0.36, max = 70, $SD = 8.97$) started the survey. Participants were recruited from Prolific Academic and paid $0.40 for their participation.

Participants who failed both manipulation checks ($n = 541$), or did not complete all measures were removed , leaving a total sample of 814 participants (462 female, 327 male, 2 other, 2 prefer not to say; $M_{\text{age}} = 26.03$, min = 11, max = 70, $SD = 9.53$).

Study 3: Procedure and materials

Again, data were collected using an online questionnaire presented with Qualtrics (www.qualtrics.com). Participants were presented with four descriptions of characters taken from Studies 1 and 2. To ensure consistency across character judgments, we selected

descriptions that related to the same moral foundations (care, fairness, and loyalty). We used the same four character names as in previous studies. The *good* characters were *Sam* and *Robin*, and the *bad* characters were *Francis* and *Alex*, e.g., *Imagine a person named Robin. Throughout their life they have been known to show compassion and empathy for others, act with a sense of fairness and justice, and, never to break their word.* or, *Imagine a person named Alex. Throughout their life they have been known to be cruel, act unfairly, and to betray their own group.* Full descriptions for each character are in the supplementary materials. One description for each the *good* and *bad* characters was randomly assigned to include non-diagnostic information for each participant thus all participants were exposed to all conditions (see https://osf.io/mdnpv/?view_only=77883e3fbc3d45f1a35fe92d5318cb67for details of the randomization blocks). Study 3 was pre-registered at https://aspredicted.org/QDF_XT1

Study 3: Results

As in Studies 1 and 2 we assessed the quality of the data by examining the extent to which responses fell above / below the midpoint for the *bad* / *good* descriptions respectively. All $N = 814$ participants responded to two *bad* descriptions and 2 *good* descriptions, resulting in a total of 1628 responses for each measure for both *bad* and *good* descriptions. Taking the *bad* descriptions first, for MPS-4, 33 (2.03%) responses were above the midpoint, and for MM-1, 114 (7.00%) were above the midpoint.

Regarding the *good* descriptions, for MPS-4, 22 (1.35%) responses were below the midpoint, and for MM-1, 17 (1.04%) were below the midpoint.

In order to test for the information type $\times$ valence interaction effect in a single model recoded both MPS-4 and MM-1 into two new variables MPS-4R, and MM-1R. These recoded variables were the same as the original variables but the responses to the good characters were reverse coded. This allowed us to examine whether the dilution effect was different depending on whether participants were judging good characters or bad

characters without this analysis being confounded by valence.

First we conducted a within-subjects factorial ANOVA to test for differences in responses to MPS-4R depending on information type and valence. There was a main effect for condition, $F(1, 813) = 48.53$, $p < .001$, partial $\eta^2 = 0.06$, 95% CI [0.03, 0.09], and a main effect for valence, $F(1, 813) = 3{,}383.29$, $p < .001$, partial $\eta^2 = 0.81$, 95% CI [0.79, 0.82]. There was a significant × valence interaction effect, $F(1, 813) = 6.35$, $p = .012$, partial $\eta^2 = 0.01$, 95% CI [0.00, 0.02].

Follow-up pairwise t-tests indicated that for the *bad* characters there were significant differences in MPS-4 responses depending on information type $t(813)$, = 6.59, $p < .001$ ($p_{\text{adjusted}} < .001$), $d = 0.23$, 95% CI [0.13, 0.25]. MPS-4 responses were higher in the non-diagnostic condition ($M = 2.33$, $SD = 0.86$) compared to the diagnostic condition ($M = 2.14$, $SD = 0.81$).

Similarly, for *good* characters, were significant differences in MPS-4 responses depending on information type $t(813)$, = -3.43, $p < .001$ ($p_{\text{adjusted}} = .003$), d 0.12, 95% CI [-0.15, -0.04]. MPS-4 responses were lower in the non-diagnostic condition ($M = 6.21$, $SD = 0.84$) compared to the diagnostic condition ($M = 6.31$, $SD = 0.74$).

Next we conducted a within-subjects factorial ANOVA to test for differences in responses to MM-1R depending on information type and valence. There was a main effect for condition, $F(1, 813) = 29.92$, $p < .001$, partial $\eta^2 = 0.04$, 95% CI [0.01, 0.06], and a main effect for valence, $F(1, 813) = 258.78$, $p < .001$, partial $\eta^2 = 0.24$, 95% CI [0.19, 0.29]. There was no condition × valence interaction effect, $F(1, 813) = 1.78$, $p = .183$, partial $\eta^2 = 0.00$, 95% CI [0, 0.01].

Follow-up pairwise t-tests indicated that for the *bad* characters there were significant differences in MM-1 responses depending on information type $t(813)$, = 4.27, $p = < .001$ ($p_{\text{adjusted}} = < .001$), $d = 0.15$, 95% CI [1.22, 3.29]. MPS-4 responses were higher in the non-diagnostic condition ($M = 22.60$, $SD = 17.06$) compared to the diagnostic

condition (M = 20.35, SD = 17.13).

Similarly, for *good* characters, were significant differences in MPS-4 responses depending on information type $t(813)$, = -3.60, $p < .001$ ($p_{\text{adjusted}} = .001$), $d = 0.13$, 95% CI [-2.16, -0.64]. MPS-4 responses were lower in the non-diagnostic condition (M = 86.68, SD = 13.86) compared to the diagnostic condition (M = 88.08, SD = 11.94).

We conducted a linear-mixed-effects model to test if our predictors influenced MPS-4R responses. Our outcome measure was MPS-4R, our predictor variables were condition and valence; we allowed intercepts and the effects of condition and valence to vary across participants, and we included random effects for scenario. Overall, the model significantly predicted participants responses, and provided a better fit for the data than the baseline model, $\chi^2(5)$ = 2,695.60, $p < .001$. Overall, there was a significant main effect for condition, $F(1, 813) = 48.53$, $p < .001$; valence significantly predicted responses, $F(1, 813) = 3{,}383.29$, $p < .001$; and there was a significant condition × valence interaction, $F(1, 813) = 6.35$, $p = .012$.

We conducted a linear-mixed-effects model to test if our predictors influenced MM-1R responses. The model was the same as the previous model, with a change to the outcome measure, our outcome measure for this model was MM-1R. As above, our predictor variables were condition and valence; we allowed intercepts and the effects of condition and valence to vary across participants, and we included random effects for scenario. Overall, the model significantly predicted participants responses, and provided a better fit for the data than the baseline model, $\chi^2(5)$ = 749.17, $p < .001$. Overall there was a main effect for condition, $F(1, 813) = 29.92$, $p < .001$; valence significantly predicted responses, $F(1, 813) = 258.78$, $p < .001$; and there was no significant condition × valence interaction, $F(1, 813) = 1.78$, $p = .183$.

Interestingly, there was a consistent effect for both condition and valence, as well as a condition × valence interaction effect. To further examine this interaction effect, we

report separate analyses for the good and bad descriptions below.

Differences in the *Bad* Descriptions

We conducted a linear-mixed-effects model to test if condition influenced MPS-4 responses. Our outcome measure was MPS-4, our predictor variable was condition; we allowed intercepts and the effect of condition to vary across participants. Overall, the model significantly predicted participants responses, and provided a better fit for the data than the baseline model, $\chi^2(3) = 76.88$, $p < .001$. Condition significantly influenced MPS-4 responses $F(1, 812.00) = 46.02$, $p < .001$, and was a significant predictor in the model $b =$ -0.10 ($b_{\text{standardized}} =$ -0.11), $t(812.00) =$ -6.78, $p < .001$, ($d =$ -0.23), see Figure 3.

We conducted a linear-mixed-effects model to test if condition influenced MM-1 responses. Our outcome measure was MM-1, our predictor variable was condition; we allowed intercepts and the effect of condition to vary across participants. Overall, the model significantly predicted participants responses, and provided a better fit for the data than the baseline model, $\chi^2(3) = 46.32$, $p < .001$. Condition significantly influenced MM-1 responses $F(1, 812.00) = 19.25$, $p < .001$, and was a significant predictor in the model $b =$ -1.14 ($b_{\text{standardized}} =$ -0.06), $t(812.00) =$ -4.39, $p < .001$, ($d =$ -0.15), see Figure 3.

Differences in the *Good* Descriptions

We conducted a linear-mixed-effects model to test if condition influenced MPS-4 responses. Our outcome measure was MPS-4, our predictor variable was condition; we allowed intercepts and the effect of condition to vary across participants. Overall, the model significantly predicted participants responses, and provided a better fit for the data than the baseline model, $\chi^2(3) = 30.01$, $p < .001$. Condition significantly influenced MPS-4 responses $F(1, 812.00) = 11.87$, $p < .001$, and was a significant predictor in the model $b =$ 0.05 ($b_{\text{standardized}} =$ 0.06), $t(812.00) =$ 3.45, $p < .001$, ($d =$ 0.12), see Figure 3.

We conducted a linear-mixed-effects model to test if condition influenced MM-1

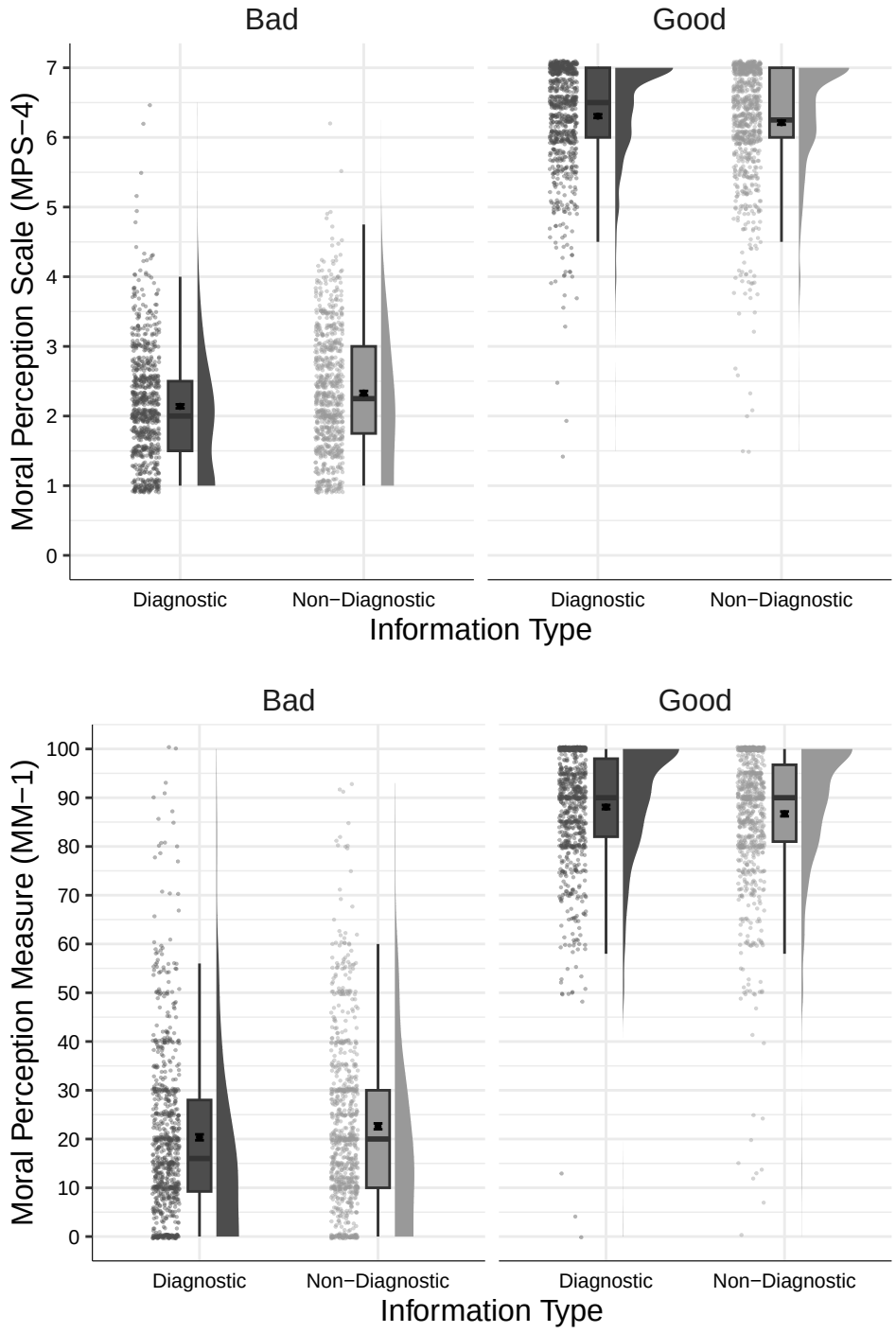

Figure 3

Study 3: Differences in moral perception depending on condition

responses. Our outcome measure was MM-1, our predictor variable was condition; we allowed intercepts and the effect of condition to vary across participants. Overall, the model significantly predicted participants responses, and provided a better fit for the data than the baseline model, $\chi^2(3) = 41.07$, $p < .001$. Condition significantly influenced MM-1 responses $F(1, 812) = 13.21$, $p < .001$, and was a significant predictor in the model $b = 0.69$ ($b_{\text{standardized}} = 0.05$), $t(812) = 3.63$, $p < .001$, ($d = 0.13$), see Figure 3.

Internal Meta Analyses

In order to examine whether the observed effects held across studies, we conducted a series of internal meta-analyses. The first set of meta-analyses includes only the studies reported in the main text of this manuscript, while the second set of meta-analyses additionally includes the studies only reported in the supplementary materials. In both cases, we report one overall meta-analysis, investigating the presence of the dilution effect for both bad characters and good characters together. Following this, we also report two additional meta-analyses testing for the presence of the dilution effect for bad characters and good characters separately.

Studies 1-3: Meta Analysis

Both

Our first meta-analysis examined Studies 1, 2, and 3. Testing for an overall dilution effect across both bad characters and good characters, and across both measures (MPS-4 and MM-1). We computed the absolute value for all effect sizes. In order to account for the nested structure of our data (i.e., multiple effect sizes being reported from each included study), we included random effects for Study (Study 1/2/3), Valence (good/bad), and Measure (MPS-4/MM-1). Overall, there was a significant dilution effect across studies, for both good and bad characters, and for both measures, $d_{pooled} = 0.12$, $SE = 0.05$, $z = 2.38$, $p = .017$, 95% CI [0.02, 0.22].

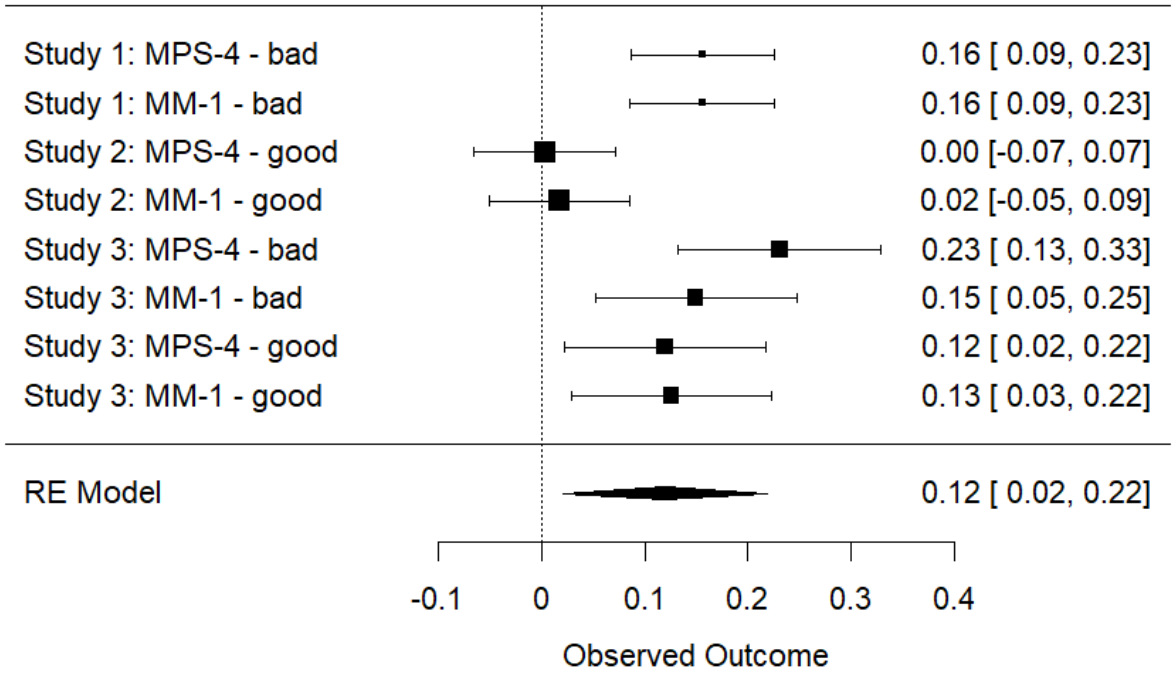

Figure 4

Forest plot showing effects for Studies 1-3 and pooled effect.

Bad

Next examined the presence of an overall dilution effect for bad characters across both measures for Studies 1 and 3. Again we included random effects for each study, as well as for measure. Overall, there was a significant dilution for bad characters, across both studies for both measures, d_{pooled} = -0.17, SE = 0.02, z = -7.97, $p < .001$, 95% CI [-0.22, -0.13].

Good

We then examined the presence of an overall dilution effect for good characters across both measures for Studies 1 and 3. Again we included random effects for each study, as well as for measure. Overall, there was no significant dilution for good characters, across both studies for either measures, $d_{pooled} = 0.06$, $SE = 0.06$, $z = 1.12$, $p = .263$, 95% CI [-0.05, 0.18].

All Studies

Both

We then proceeded to re-run the above meta-analyses but to additionally include the supplementary studies. Other than the inclusion of additional studies, the meta-analyses described below are the same as those already reported. We first included effect sizes for both bad characters and good characters, for both measures (MPS-4 and MM-1), with absolute values for all effect sizes and random effects for Study (Study 1/2/3/S1/S2/S3), Valence (good/bad), and Measure (MPS-4/MM-1). Second we test the effect for bad characters only, and third we test for the effect for good characters only. Overall, there was a significant dilution effect across all studies, for both good and bad characters, and for both measures, $d_{pooled} = 0.11$, $SE = 0.03$, $z = 3.80$, $p < .001$, 95% CI [0.05, 0.16].

Bad

Overall, there was a significant dilution for bad characters across all studies for both measures, $d_{pooled} = -0.12$, $SE = 0.02$, $z = -4.91$, $p < .001$, 95% CI [-0.17, -0.07].

Good

Overall, there was a significant dilution for good characters, across all studies and for both measures, $d_{pooled} = 0.09$, $SE = 0.04$, $z = 2.55$, $p = .011$, 95% CI [0.02, 0.17].

Discussion

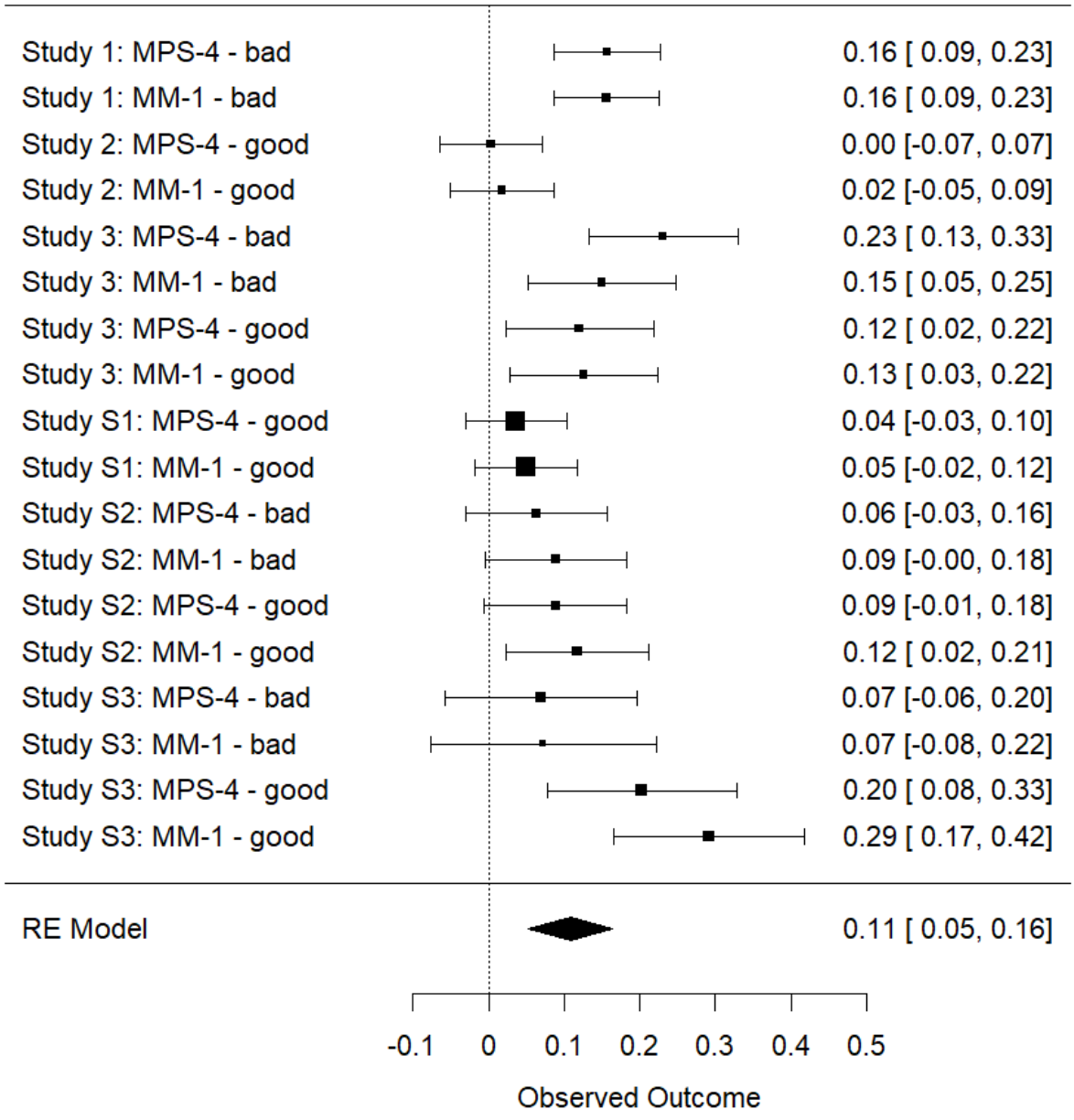

Figure 5

Forest plot showing effects for all studies and pooled effect.

Accessibility Statement

All data and analysis code are publicly available on this project's OSF page at https://osf.io/mdnpv/?view_only=77883e3fbc3d45f1a35fe92d5318cb67.

References

Cameron, C. D., Payne, B. K., & Doris, J. M. (2013). Morality in high definition: Emotion differentiation calibrates the influence of incidental disgust on moral judgments. *Journal of Experimental Social Psychology*, *49*(4), 719–725. https://doi.org/10.1016/j.jesp.2013.02.014

Christensen, J. F., Flexas, A., Calabrese, M., Gut, N. K., & Gomila, A. (2014). Moral judgment reloaded: A moral dilemma validation study. *Frontiers in Psychology*, *5*, 1–18. 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. https://doi.org/10.1016/j.neubiorev.2012.02.008

Giner-Sorolla, R. (2018). A Functional Conflict Theory of Moral Emotions. In K. J. Gray & J. Graham (Eds.), *Atlas of Moral Psychology* (pp. 81–87). New York, NY: The Guilford Press.

Gray, K. J., & 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. J., Waytz, A., & Young, L. (2012). The Moral Dyad: A Fundamental Template Unifying Moral Judgment. *Psychological Inquiry*, *23*(2), 206–215. https://doi.org/10.1080/1047840X.2012.686247

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 (New York, N.Y.)*, *293*(5537), 2105–2108. https://doi.org/10.1126/science.1062872

Igou, E. R., & Bless, H. (2005). The Conversational Basis for the Dilution Effect. *Journal of Language and Social Psychology*, *24*(1), 25–35. https://doi.org/10.1177/0261927X04273035

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of

representativeness. *Cognitive Psychology*, *3*(3), 430–454. https://doi.org/10.1016/0010-0285(72)90016-3

LaBella, C., & Koehler, D. J. (2004). Dilution and confirmation of probability judgments based on nondiagnostic evidence. *Memory and Cognition*, *32*(7), 1076–1089. https://doi.org/10.3758/BF03196883

McHugh, C., McGann, M., Igou, E. R., & Kinsella, E. L. (2022). Moral Judgment as Categorization (MJAC). *Perspectives on Psychological Science*, *17*(1), 131–152. https://doi.org/10.1177/1745691621990636

Meyvis, T., & Janiszewski, C. (2002). Consumers' Beliefs about Product Benefits: The Effect of Obviously Irrelevant Product Information. *Journal of Consumer Research*, *28*(4), 618–635. https://doi.org/10.1086/338205

Nisbett, R. E., Zukier, H., & Lemley, R. E. (1981). The dilution effect: Nondiagnostic information weakens the implications of diagnostic information. *Cognitive Psychology*, *13*(2), 248–277. https://doi.org/10.1016/0010-0285(81)90010-4

Schein, C., & Gray, K. J. (2018). The Theory of Dyadic Morality: Reinventing Moral Judgment by Redefining Harm. *Personality and Social Psychology Review*, *22*(1), 32–70. https://doi.org/10.1177/1088868317698288

Zukier, H. (1982). The dilution effect: The role of the correlation and the dispersion of predictor variables in the use of nondiagnostic information. *Journal of Personality and Social Psychology*, *43*(6), 1163–1174. https://doi.org/10.1037/0022-3514.43.6.1163