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

Cillian McHugh & Eric R. Igou

University of Limerick

Author note

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. The authors declare that there are no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. All authors consented to the submission of this manuscript.

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

Abstract

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. The effect has been observed for a range of judgments, including judging products, probability judgments, and predictions relating to people's behavior. The dilution effect has been explained as emerging as a consequence of the representativeness heuristic, such that the inclusion of non-diagnostic information reduces the match between the target and a typical member of the category. In line with this, and drawing on the recent moral judgment as categorization theory of moral judgment (MJAC), we predict that the dilution effect should be observed for judgments about morality. Across three studies (total $N = 2485$), we tested for the dilution effect on judgments of morally bad actors and morally good actors. Overall, our results showed a dilution effect for judgments of both good and bad actors. People's moral evaluations of both good and bad actors were less extreme when the descriptions included non-diagnostic information. We also found that the robustness of the effect appears to be moderated by valence, with a more robust effect for bad actors. Our results highlight avenues for future research.

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

Word count: 6,987

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 bystander 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 et al., 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 bright. 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 their grade point average (GPA; 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, namely information that is usually

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 a 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, but the absence of any non-diagnostic information means there is nothing to suggest any differences between the target and a typical (or stereotypical) category member. 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 the 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 et al., 2022). Examples of known

contextual influences on people's moral judgments include emotions (Cameron et al., 2013; Giner-Sorolla, 2018), intentionality and evitability (Christensen et al., 2014; Christensen & Gomila, 2012), and how 'up close and personal' an action is (Greene et al., 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). The more closely a target aligns with this representative description, the more typical it is perceived to be. 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 et al., 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 stationery 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]

Ethical Declarations

All procedures performed in studies involving human participants were approved by the 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 Actors

The aim of Study 1 is to test if the dilution effect exists in the moral domain. Participants were presented with descriptions of four actors; two descriptions contained diagnostic information (morally relevant information), and two additionally contained non-diagnostic

[1] We also report in the supplementary materials three additional studies with similar designs to Studies 2 (Study S1) and 3 (Study S2), and a between-subjects version of Study 3 (Study S3). These Studies S1-S3 were conducted on MTurk and we observed irregularities with the quality of the data suggesting that participants were not engaging properly with the tasks despite passing the attention checks (e.g., bimodal distributions for responses to *bad* actors in both Studies S2 and S3). Because we are not confident in the quality of the Mturk data, these studies are not reported in the main text. With the exception of a small sub-sample in Study 2 all participants are drawn from convenience/snowball samples from the student body and wider community associated with the University of Limerick.

information (non-morally relevant information) along with the diagnostic information. We hypothesized that moral perceptions of the diagnostic-only descriptions would be more severe than those for the descriptions that also contain non-diagnostic information.

Methods

Participants and Design

Study 1 was a within-subjects design. The independent variable was the condition with two levels, diagnostic information only (diagnostic) and non-diagnostic information additionally included (non-diagnostic). We used two dependent variables, the four-item moral perception scale and the single-item moral perception measure. Both dependent variables were adapted from Walker et al. (2021).

A total sample of 851 (302 female, 526 male, 14 non-binary, 5 other; 3 prefer not to say, M_{age} = 26.16, min = 18, max = 76, *SD* = 10.14) completed the survey. Participants were recruited from the student population at the University of Limerick. Participants who failed both manipulation checks were removed (n = 50), leaving a total sample of 801 participants (283 female, 496 male, 14 non-binary, 5 other, 3 prefer not to say; M_{age} = 26.25, min = 18, max = 76, *SD* = 10.20).

Procedure and Materials

Data were collected using an online questionnaire presented with Qualtrics (www.qualtrics.com). Participants were presented with four descriptions of actors (*Sam*, *Alex*, *Francis*, *Robin*). These descriptions were developed by adapting items from the extended character morality questionnaire (Grizzard et al., 2020). All descriptions included diagnostic information relating to three moral foundations and read as follows:

Imagine a person named Sam. Throughout their life they have been known to be cruel, act unfairly, and to betray their own group.

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.

Imagine a person named Francis. Throughout their life they have been known to violate the standards of purity and decency, show lack of respect for authority, and treat people unequally.

Imagine a person named Alex. Throughout their life they have been known to cause others to suffer emotionally, to deny others their rights, and to cause chaos or disorder.

All participants were presented with all four descriptions. For each participant two descriptions additionally included non-diagnostic information. The inclusion of non-diagnostic information was fully randomized across the four descriptions. The non-diagnostic information read as follows: (i) *They have red hair, play tennis four times a month, and have one older sibling and one younger sibling*; (ii) *They are left-handed, drink tea in the morning, and have two older siblings and one younger sibling*. Piloting confirmed that the content of the diagnostic descriptions was rated as significantly more morally wrong than that of the non-diagnostic descriptions (See Pilot Study 1).

There were two dependent variables. The first dependent variable was the four-item moral perception scale (henceforth MPS-4). Participants were asked to "*Please rate the person along the following dimensions*": Bad/Good, Immoral/Moral, Violent/Peaceful, Merciless/Empathetic (each on 7-point Likert scales). This measure showed good reliability, $\alpha = 0.83$. The second dependent variable was a single-item Moral Perception Measure (henceforth MM-1). Participants were asked to "*Please rate the person according to how immoral or moral you view them*" with a slider ranging from 0 (*Very Immoral*) to 100 (*Very Moral*).

We programmed our survey to randomly present non-diagnostic information along with two of the descriptions participants read (this was done through blocking, see

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

Results and Discussion

We first provide basic descriptive statistics relating to the different scenarios, before describing the main hypothesis tests. The means and standard deviations for MPS-4 for each scenario are as follows: *Sam*, $M_{MPS\text{-}4} = 2.55$, $SD_{MPS\text{-}4} = 0.86$, *Francis*, $M_{MPS\text{-}4} = 3.05$, $SD_{MPS\text{-}4} = 0.97$, *Alex*, $M_{MPS\text{-}4} = 2.32$, $SD_{MPS\text{-}4} = 0.88$, *Robin*, $M_{MPS\text{-}4} = 2.13$, $SD_{MPS\text{-}4} = 0.91$. There was significant variation depending on the description, $F(3,2280) = 297.82$, $p < .001$, partial $\eta^2 = 0.13$. *Francis* appeared to be rated as more moral than each of the other actors (all *p*s < .001), *Sam* was the next most favorable, followed by *Alex*, while *Robin* was rated as less moral than each of the other actors (all *p*s < .001).

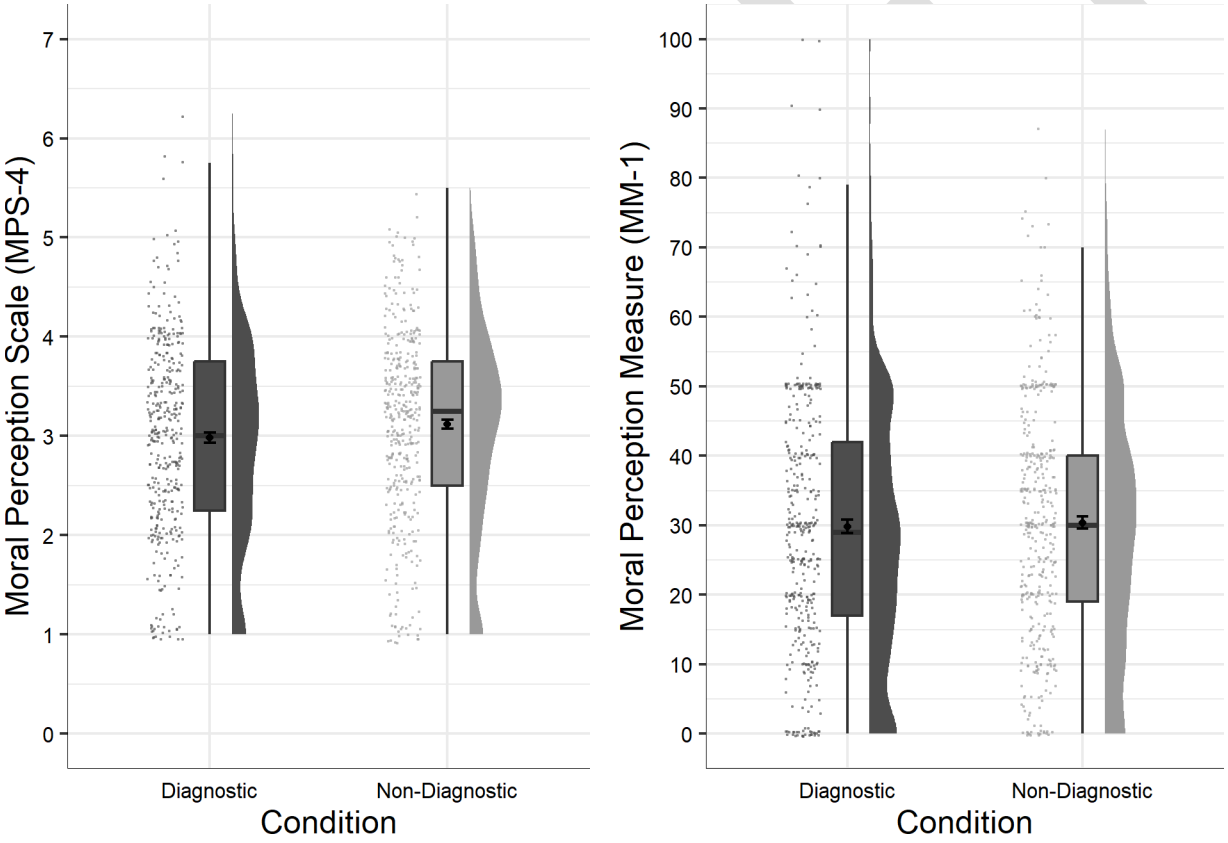
The means and standard deviations for MM-1 for each scenario are as follows: *Sam*, $M_{MM\text{-}1} = 23.94$, $SD_{MM\text{-}1} = 16.18$; *Francis*, $M_{MM\text{-}1} = 30.12$, $SD_{MM\text{-}1} = 17.86$; *Alex*, $M_{MM\text{-}1} = 20.55$, $SD_{MM\text{-}1} = 16.65$; *Robin*, $M_{MM\text{-}1} = 20.60$, $SD_{MM\text{-}1} = 17.06$. There was significant variation depending on the description, $F(3,2253) = 154.08$, $p < .001$, partial $\eta^2 = 0.05$. *Francis* was rated more favorably than all other actors ($p < .001$), *Sam* was the next most favorably rated actor, rated significantly more favorably than both *Alex* and *Robin* (*p*s < .001), there was no difference between *Alex* and *Robin* ($p = .953$).

To test our hypothesis, we conducted a linear-mixed-effects model to test if condition influenced judgments of the morality of the actors. We conducted separate analyses for each measure. For our first analysis, our outcome measure was MPS-4, and 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, $b = -0.08$, $t(799.42) = -7.17$, $p < .001$, with the diagnostic descriptions being rated as more immoral than the non-diagnostic descriptions Figure 1.

Figure 1

Study 1: Differences in Moral Perception Depending on Condition



For our second analysis, 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, $\chi^2(8) = 475.52$, $p < .001$. 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 $b = -1.22$, $t(799.71) = -6.66$, $p < .001$, with the diagnostic descriptions being rated as more immoral than the non-diagnostic descriptions, see Figure 1. (In the supplementary analyses we report the effect of condition on moral perception for each description individually, as well as for a combined moral perception measure).

In line with our hypothesis, we found evidence for a dilution effect involving descriptions of immoral exemplars across both our dependent measures. This suggests that judgments of immoral actors are sensitive to variations in typicality and that this typicality variation is independent of the severity of the action. The inclusion of non-diagnostic information appears to reduce the match between a target and what is normally regarded as *representative* of *morally wrong*.

Study 2 - Good Actors

The aim of Study 2 is to test if the dilution effect exists in the moral domain for judgments of morally *good* actors. Participants were presented with descriptions of four actors; two descriptions contained diagnostic information only (morally relevant information), and two additionally contained non-diagnostic information (non-morally relevant information) along with the diagnostic information. We hypothesized that moral perceptions of the diagnostic-only descriptions would be more extreme (more moral) than for the descriptions that also contain non-diagnostic information.

Methods

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 Study 1, 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 (418 female, 557 male, 13 non-binary, 2 other; 4 prefer not to say, $M_{age} = 29.04$, min = 18, max = 74, $SD = 10.66$) started the survey. Participants who failed both manipulation checks were removed ($n = 248$), leaving a total sample of 820 participants (337 female, 466 male, 2 other, 2 prefer not to say; $M_{age} = 29.03$, min = 18, max = 74, $SD = 10.92$).

The majority of participants were from the student population at University of Limerick: $n = 533$, (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.

Procedure and Materials

As in Study 1, data were collected using an online questionnaire presented with Qualtrics (www.qualtrics.com). Participants were presented with four descriptions of actors that read as follows:

Imagine a person named Sam. Throughout their life they have been known to always help and care for others, treat everyone fairly and equally, and show a strong sense of loyalty to others.

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.

Imagine a person named Francis. Throughout their life they have been known to uphold the standards of purity and decency, show respect for authority, and to always act honestly and fairly.

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 were randomly programmed to additionally include non-diagnostic information (this was randomized through blocking, see https://osf.io/mdnpv/?view_only=77883e3fbc3d45f1a35fe92d5318cb67). The non-diagnostic information read as follows: (i) *They have dark hair, go for a jog twice a week, and their favorite color is blue*; (ii) *They have blue eyes, drink coffee in the morning, and their favorite color is green*. Piloting confirmed that the diagnostic descriptive material was rated as more moral than the non-diagnostic material (See Pilot Study 2). Study 2 was pre-registered at https://aspredicted.org/NX2_HN6

Results and Discussion

The means and standard deviations for MPS-4 for each scenario are as follows: *Sam*, $M_{MPS\text{-}4} = 6.12$, $SD_{MPS\text{-}4} = 0.97$, *Francis*, $M_{MPS\text{-}4} = 5.86$, $SD_{MPS\text{-}4} = 1.07$, *Alex*, $M_{MPS\text{-}4} = 6.13$, $SD_{MPS\text{-}4} = 0.99$, *Robin*, $M_{MPS\text{-}4} = 6.10$, $SD_{MPS\text{-}4} = 0.99$. There was significant variation depending on the description, $F(3,2356) = 54.47$, $p < .001$, partial $\eta^2 = 0.01$. *Francis* appeared to be rated as less moral than each of the other actors (all *p*s < .001), and there were no differences between *Alex*, *Robin*, and *Sam* (all *p*s > .5).

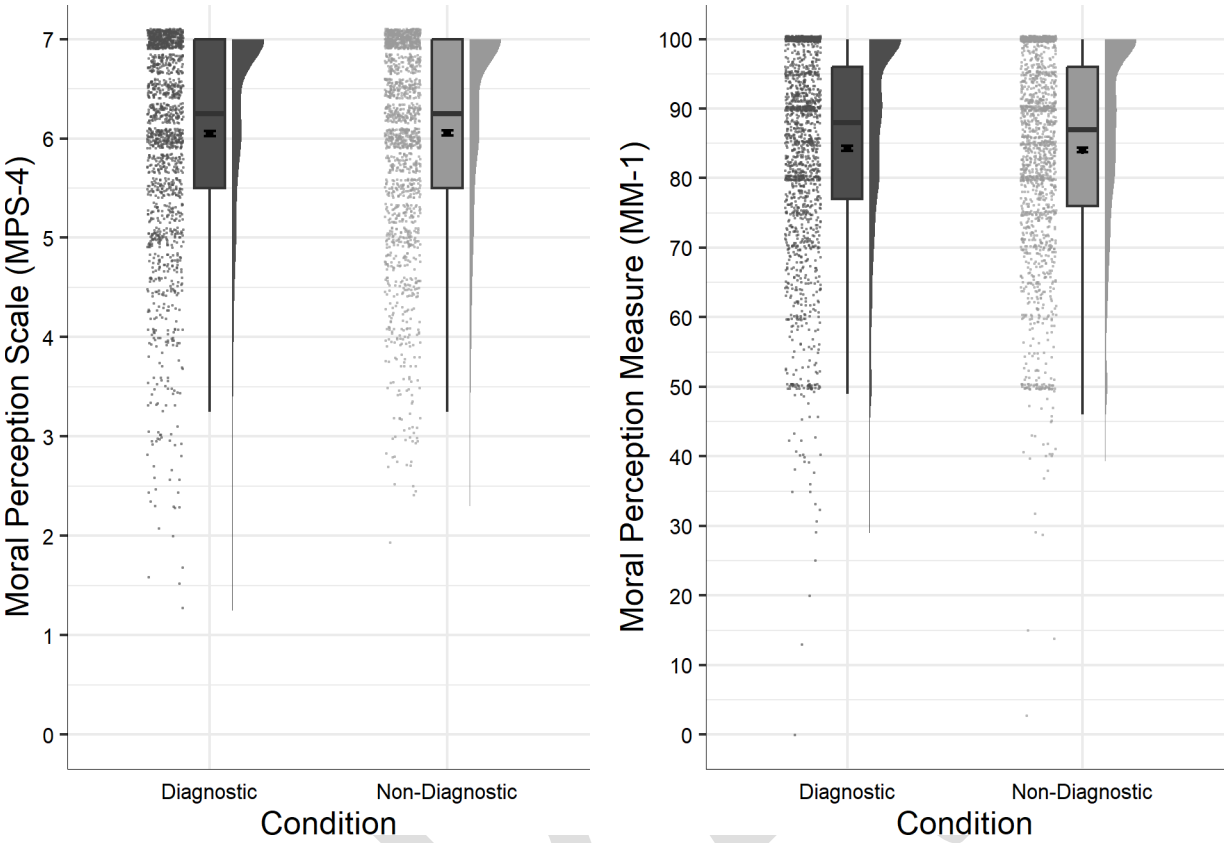
The means and standard deviations for MM-1 for each scenario are as follows: *Sam* (diagnostic/moral), $M_{MM\text{-}1} = 84.60$, $SD_{MM\text{-}1} = 14.47$; *Francis* (diagnostic/moral), $M_{MM\text{-}1} = 82.05$, $SD_{MM\text{-}1} = 15.24$; *Alex* (diagnostic/moral), $M_{MM\text{-}1} = 85.02$, $SD_{MM\text{-}1} = 15.01$; *Robin* (diagnostic/moral), $M_{MM\text{-}1} = 84.95$, $SD_{MM\text{-}1} = 13.94$. There was significant variation depending on the description, $F(3,2387) = 24.20$, $p < .001$, partial $\eta^2 = 0.007$. *Francis* was rated less favorably than all other actors (all ps $< .001$), there were no differences between *Alex*, *Robin*, and *Sam* (all ps $> .5$).

To test our hypothesis, we conducted a linear-mixed-effects model to test if condition influenced MPS-4 and MM-1 responses.

For our first analysis, 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, $b = 0.00$, $t(838) = 0.49$, $p = .624$, see Figure 2.

Figure 2

Study 2: Differences in moral perception depending on condition



We conducted a linear-mixed-effects model to test if the 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 $b = 0.16$, $t(2453) = 1.11$, $p = .267$, see Figure 2. In the supplementary analyses we report the effect of the condition on moral perception for each description individually.

In contrast to Study 1, Study 2 did not provide support for our hypothesis, and we did not find evidence for a dilution effect in judgments of morally good actors. These results may have

revealed something different about the way in which people think about *good* actors versus *bad* actors. It is possible that for bad actors, there is a more clearly defined *representative* prototype (in line with Schein & Gray, 2018). In contrast, the category *good actors* may represent a more varied set of exemplars, with a less clear prototype that is *representative* of the category. This would mean that the presence of the moral dilution effect may depend on whether people are judging good actors or bad actors. We test this possibility in Study 3.

Study 3 - Good and Bad Actors

The aim of Study 3 was to test one potential explanation of the contrasting results between Studies 1 and 2. It is possible that the presence of the moral dilution effect depends on whether people judge *good* actors or *bad* actors, reflecting differences in the ways people think about *good* versus *bad* actors. In this case, valence (good vs. bad) would moderate the dilution effect. Study 3 was designed to test for this potential moderation 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 actors but not for good actors. Study 3 was pre-registered at https://aspredicted.org/QDF_XT1.

Methods

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 (α = 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_{age} = 29.67, min = 0.36, max = 70, SD = 8.97) started the survey. Participants were

primarily recruited from the student body of University of Limerick, through an invitation circulated on the internal email. Additional participants were recruited by posting the survey on the department's social media accounts. 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_{age} = 26.03, min = 11, max = 70, SD = 9.53).

Procedure and materials.

Again, data were collected using an online questionnaire presented with Qualtrics (www.qualtrics.com). Participants were presented with four descriptions of actors 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 actor names as in previous studies. The *good* actors were *Sam* and *Robin*, that read as follows:

Imagine a person named Sam. Throughout their life they have been known to always help and care for others, treat everyone fairly and equally, and show a strong sense of loyalty to others.

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.

The *bad* actors were *Alex* and *Francis*, and the descriptions read as follows:

Imagine a person named Alex. Throughout their life they have been known to be cruel, act unfairly, and to betray their own group.

Imagine a person named Francis. Throughout their life they have been known to physically hurt others, treat some people differently to others, and show lack of loyalty.

The non-diagnostic descriptions read as follows: (i) *They have red hair, play tennis four times a month, and have one older sibling and one younger sibling*; (ii) *They are left-handed,*

drink tea in the morning, and have two older siblings and one younger sibling. One description for each the *good* and *bad* actors was randomly assigned to include non-diagnostic information for each participant thus all participants were exposed to all conditions (for details of the randomization blocks see https://osf.io/mdnpv/?view_only=77883e3fbc3d45f1a35fe92d5318cb67). Study 3 was pre-registered at https://aspredicted.org/QDF_XT1

Results and Discussion

The means and standard deviations for MPS-4 for each scenario are as follows: *Sam* (good), $M_{MPS\text{-}4} = 6.20$, $SD_{MPS\text{-}4} = 0.82$, *Francis* (bad), $M_{MPS\text{-}4} = 2.15$, $SD_{MPS\text{-}4} = 0.79$, *Alex* (bad), $M_{MPS\text{-}4} = 2.32$, $SD_{MPS\text{-}4} = 0.88$, *Robin* (good), $M_{MPS\text{-}4} = 6.32$, $SD_{MPS\text{-}4} = 0.76$. There was significant variation depending on the description, $F(2,1515) = 6{,}251.52$, $p < .001$, partial $\eta^2 = 0.86$. Both the *good* characters (*Robin* and *Sam*) were rated significantly more favorably than both the *bad* characters (*Alex* and *Francis*; all *p*s < .001). In addition, *Robin* was viewed more favorably than *Sam* (*good*: $p = .003$), and *Alex* more favorably than *Francis* (*bad*; $p < .001$).

The means and standard deviations for MM-1 for each scenario are as follows: *Sam* (good), $M_{MM\text{-}1} = 86.36$, $SD_{MM\text{-}1} = 13.71$; *Francis* (bad), $M_{MM\text{-}1} = 20.13$, $SD_{MM\text{-}1} = 16.87$; *Alex* (bad), $M_{MM\text{-}1} = 22.83$, $SD_{MM\text{-}1} = 17.29$; *Robin* (good), $M_{MM\text{-}1} = 88.41$, $SD_{MM\text{-}1} = 12.07$. There was significant variation depending on the description, $F(2,1380) = 5{,}282.47$, $p < .001$, partial $\eta^2 = 0.826$. Again, the *good* characters (*Robin* and *Sam*) were rated significantly more favorably than the *bad* characters (*Alex* and *Francis*; all *p*s < .001 In addition, *Robin* was viewed more favorably than *Sam* (*good*: $p = .001$), and *Alex* more favorably than *Francis* (*bad*; $p < .001$).

In order to test for the condition × valence interaction effect, we created a new variable to represent the extremity of participants' judgments for both MPS-4 and MM-1. For each measure,

we calculated the absolute value of the difference between each response to the midpoint of the scale. Higher scores are more extreme, and lower scores were less extreme.

We conducted a linear-mixed-effects model to test if our predictors influenced the extremity of MPS-4 responses. Our outcome measure was the extremity of MPS-4 responses, our predictor variables were condition and valence; we allowed intercepts and the effects of condition and valence to vary across participants. Overall, the model significantly predicted participants' responses and provided a better fit for the data than the baseline model, $\chi^2(5) = 755.71$, $p < .001$. Overall, there was a significant main effect for condition, $F(1, 825) = 56.42$, $p < .001$; valence significantly predicted responses, $F(1, 833) = 481.30$, $p < .001$; and there was a significant condition × valence interaction, $F(1, 826) = 12.64$, $p < .001$.

We conducted a linear-mixed-effects model to test if our predictors influenced MM-1 responses. The model was the same as the previous model, with a change to the outcome measure. Our outcome measure for this model was the extremity of MM-1 responses. As above, our predictor variables were condition and valence; we allowed intercepts and the effects of condition and valence to vary across participants. Overall, the model significantly predicted participants responses, and provided a better fit for the data than the baseline model, $\chi^2(5) = 840.14$, $p < .001$. Overall there was a main effect for condition, $F(1, 828) = 43.64$, $p < .001$; valence significantly predicted responses, $F(1, 838) = 389.21$, $p < .001$; and there was a significant condition × valence interaction, $F(1, 829) = 8.77$, $p = .003$.

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.

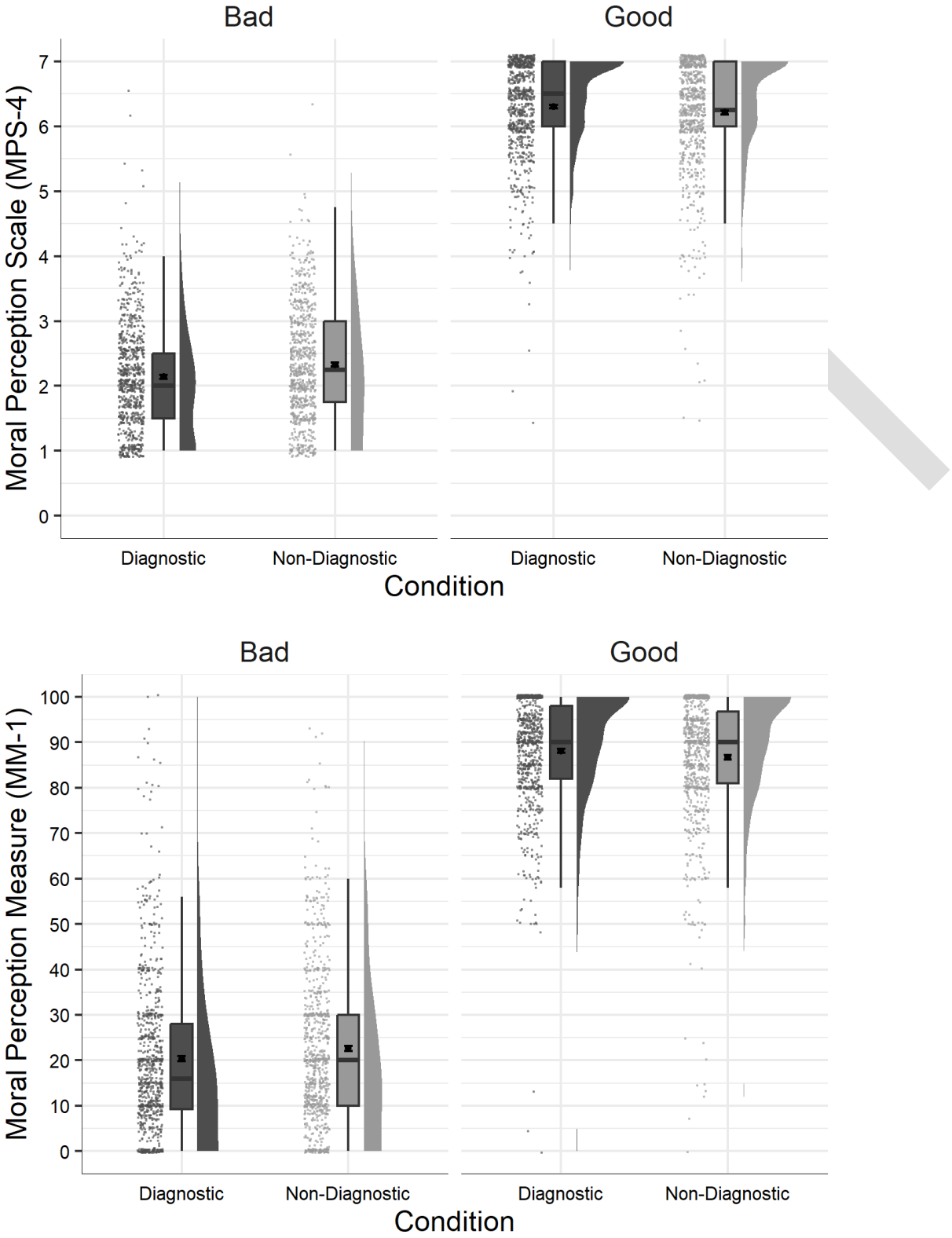
Differences in the 'Bad' Descriptions

For the bad descriptions, we conducted a linear-mixed-effects model to test if the condition influenced MPS-4 responses. Our outcome measure was MPS-4, and 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 $b = -0.10$, $t(812.00) = -6.78$, $p < .001$, see Figure 3.

We conducted a linear-mixed-effects model to test if the condition influenced MM-1 responses. Our outcome measure was MM-1, and 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 $b = -1.14$, $t(812.00) = -4.39$, $p < .001$, see Figure 3.

Figure 3

Study 3: Differences in Moral Perception Depending on Condition



Differences in the 'Good' Descriptions

For the good descriptions, we conducted a linear-mixed-effects model to test if the condition influenced MPS-4 responses. Our outcome measure was MPS-4, and 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 $b = 0.05$, $t(812.00) = 3.45$, $p < .001$, see Figure 3.

We conducted a linear-mixed-effects model to test if the condition influenced MM-1 responses. Our outcome measure was MM-1, and 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 $b = 0.69$, $t(812) = 3.63$, $p < .001$, see Figure 3.

The aim of Study 3 was to test if the moral dilution effect was moderated by the valence of description. Based on the results of Studies 1 and 2 we predicted a condition × valence interaction effect, whereby we hypothesized that a dilution effect would be observed for judgments of *bad* actors, but not for judgments of *good* actors. Interestingly, while we did observe a condition × valence interaction effect, we also observed the dilution effect for both *bad* actors, and for *good* actors. This was unexpected given the results of Study 2; however, the investigation of the b values suggests that the dilution effect is indeed stronger for the *bad* actors than for the *good* actors, which is in line with our prediction.

Meta-Analysis

In order to examine whether the observed effects held across studies, we conducted a series of mini meta-analyses. A mini meta-analysis found a significant dilution effect across both good and bad character descriptions across all studies $\chi^2(16) = 114.11$, $p < .001$ (Fisher's method); this remained when weighting for sample size, $z = 7.995$, $p < .001$ (Stouffer's Z-score method). A mini meta-analysis also found a significant dilution effect for bad character descriptions across all studies $\chi^2(8) = 73.68$, $p < .001$ (Fisher's method); or when weighting for sample size, $z = 7.438$, $p < .001$ (Stouffer's Z-score method). Finally, a mini meta-analysis found a significant dilution effect for good character descriptions across all studies $\chi^2(8) = 40.43$, $p < .001$ (Fisher's method); or when weighting for sample size, $z = 3.871$, $p < .001$ (Stouffer's Z-score method).

General Discussion

Across three studies, we provide evidence for a moral dilution effect – the presence of non-diagnostic information resulted in less extreme judgments of both good and bad actors. To our knowledge, this is the first demonstration of the dilution effect on judgments of moral character. In line with the representativeness heuristic, some actors are more representative examples of a *bad actor* or a *good actor*, and the inclusion of non-diagnostic information in a description of a moral actor reduces the representativeness of the target, leading to less extreme judgments of the actors (Kahneman & Tversky, 1972; Nisbett et al., 1981).

By extending this well-established finding to include judgments in the moral domain, our findings contribute both (a) to scholarship on the dilution effect (demonstrating additional areas in which it is observed), and (b) to ongoing theorizing about moral judgment. Regarding (b) our findings demonstrate that moral judgments can be subject to the same kinds of influences as judgments outside the moral domain, adding to a growing body of evidence demonstrating these

comparable effects across moral and non-moral domains (for discussion, see McHugh et al., 2022). More specifically, moral judgment as categorization (MJAC) predicts that moral judgments and judgments of moral actors should vary in their typicality or in their representativeness as members of the category *morally wrong* or *morally right* (we note this prediction is not unique to MJAC, see Schein & Gray, 2018). The moral dilution effect observed here provides clear evidence for the variability in typicality predicted by MJAC (McHugh et al., 2022). The inclusion of non-diagnostic information in a description creates a mismatch between a target and a *typical* member of the relevant category. This mismatch leads to less extreme ratings of the actors when compared to descriptions that have only diagnostic information – the *moral* dilution effect. Thus, our findings are consistent with and provide preliminary support for MJAC as an approach to understanding moral judgment. Future research should apply the dilution paradigm to derive competing predictions between approaches to moral judgment to further advance our understanding of how people make moral judgments or moral categorizations (e.g., providing a direct test between MJAC and Dyadic Morality; McHugh et al., 2022; Schein & Gray, 2018).

Interestingly, the dilution effect appears to be stronger for judgments of *bad* actors, as compared to judgments of *good* actors (Study 3), and the effect also appears to be more robust for judgments of *bad* characters (no effect observed in Study 2). It is possible that this variability sheds light on differences in how categorizations of moral rightness and moral wrongness are organized. Specifically, it may suggest that there is greater variation regarding what is considered representative of morally *right* compared to what is representative of morally *wrong*. Future research should explore this possibility in more detail.

We note that the validity of our conclusions is strengthened by the recruitment of relatively large samples across all studies. Our stimulus materials, measures, and methods were

derived from existing work, and we conducted two separate pilot studies to test the appropriateness of the materials (one for the *bad* descriptions and one for the *good* descriptions, see Supplementary Materials). We also note that despite the observed variability across studies, the results of the meta-analysis provide strong support for our conclusions.

Limitations and Future Directions

Despite finding overall evidence for the moral dilution effect, our results also showed some interesting variability. Specifically, the effect was not observed for descriptions of *good* actors in Study 2; however, when participants were additionally presented with descriptions of *bad* actors (Study 3), the dilution effect was reliably observed for the *good* actors. These results may provide an insight into the different ways people think about *good* actors versus *bad* actors. One well established finding in the literature is that *bad* (information/events/actors/emotions) is more salient, attention-grabbing, and is processed mor thoroughly than *good* (Baumeister et al., 2001; Pratto & John, 1991). We found that in the presence of *bad* actors, people appear to readily differentiate between different levels of *good* actors. It is possible that the presence of *bad* actors provides an anchor or a contrast case to which the good actors can be evaluated. The presence of this contrast case results in the different actors being rated in relation to each other thus any differences between the different good actors may become more salient, with actors that better match a typical *good* prototype being rated as *more* good than actors that diverge from this prototype (through the presence of non-diagnostic information). Future research should replicate and extend these findings, investigating this in more detail.

We also note that the way in which our participants made their judgments in these studies is not necessarily representative of how people make moral judgments in everyday life. Future research should investigate more varied descriptions and attempt to investigate the effect in more real world settings. Our participants were taken largely from student populations and participants

who are connected to the university through social media. Future research should investigate this effect in more diverse populations.

Practical Implications

The moral dilution effect has implications for real-world character judgments more generally. For example, in the wake of a political scandal, it is not uncommon for politicians or public figures to attempt to present themselves as "just an ordinary, hard-working" person (Hussey, 2022; Valgarðsson et al., 2021). Such a strategy may reduce their similarity with a prototypical example of a *corrupt person (in politics, business, or other areas of public life)*, leading to more favorable evaluations. Outside of politics, these influences on character judgments can have a significant impact on how people are treated across a range of settings (e.g., legal settings, access to institutions / services / accommodation, hiring decisions).

One area where this impact is readily apparent is in the domain of sexual harassment and consent and in attempts to address these complex issues. For example, a widely acknowledged challenge to addressing the issue of sexual harassment is the perpetuation of what has been termed the serial rapist model (Gantman & Paluck, 2018). This is the belief that most instances of sexual misconduct are committed by a small number of males who are fundamentally different from their peers. This means that people only expect sexual harassment to be perpetrated by people who match the prototype of a serial rapist, and the actions of actors who do not match this prototype are less likely the be identified as harassment. This has clear legal implications, whereby the character of the accused (and indeed of the victim, e.g., Randall, 2010) comes under intense scrutiny in court cases involving sexual misconduct. Descriptions of the accused may include details that are inconsistent with the prototypical "serial rapist" (e.g., their standing in the community, their various achievements, see Levin, 2016; McKay, 2018). Relatedly, people have pre-existing beliefs about victims of sexual harassment, and if a victim does not match the

prototype of a "real" victim they are less likely to be believed (Randall, 2010). And as with the accused, the character of a victim also comes under intense scrutiny in court (Freeman, 2018; Levin, 2016).

Conclusion

Our moral judgments can be highly variable and sensitive to various contextual influences. We show that the presence of seemingly irrelevant information influences people's judgments of moral character. People's judgments of good actors and bad actors were less extreme when information that was not morally relevant was included in the description. Our findings support a categorization approach to understanding moral judgment (contributing to theory building), and also have practical implications across a range of real-world settings.

Accessibility Statement

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

Competing Interests

The authors declare that there are no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. All authors consented to the submission of this manuscript.

Author Contribution Statement

CMcH: project conception, study design, data collection, analysis and results, interpretation of findings, drafting manuscript

ERI: project conception, study design, interpretation of findings, review of manuscript

References

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is Stronger than Good. *Review of General Psychology*, *5*(4), 323–370. https://doi.org/10.1037/1089-2680.5.4.323

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

Freeman, H. (2018, April 4). What does the Belfast rape trial tell women? Make a complaint and you'll be vilified. *The Guardian*. https://www.theguardian.com/fashion/2018/apr/04/what-does-the-belfast-trial-tell-women-make-a-complaint-and-youll-be-vilified

Gantman, A., & Paluck, E. L. (2018). What is the Psychological Appeal of the Serial Rapist Model? Worldviews Predicting Endorsement. *Behavioral Public Policy*. https://papers.ssrn.com/abstract=3190670

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). 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

Grizzard, M., Fitzgerald, K., Francemone, C. J., Ahn, C., Huang, J., Walton, J., McAllister, C., & Eden, A. (2020). Validating the extended character morality questionnaire. *Media Psychology*, *23*(1), 107–130. https://doi.org/10.1080/15213269.2019.1572523

Hussey, S. (2022). *Robert Troy resigns from Minister of State role*. https://www.rte.ie/news/politics/2022/0824/1318499-bacik-says-troy-showed-careless-disregard-for-rules/

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

Levin, S. (2016, July 19). Stanford sexual assault victim faced personal questions at trial, records show. *The Guardian*. https://www.theguardian.com/us-news/2016/jul/19/stanford-sexual-assault-brock-turner-victim-personal-questions

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

McKay, S. (2018, December 4). How the 'rugby rape trial' divided Ireland. *The Guardian*. http://www.theguardian.com/news/2018/dec/04/rugby-rape-trial-ireland-belfast-case

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

Pratto, F., & John, O. P. (1991). Automatic vigilance: The attention-grabbing power of negative social information. *Journal of Personality and Social Psychology*, *61*(3), 380–391. https://doi.org/10.1037//0022-3514.61.3.380

Randall, M. (2010). Sexual Assault Law, Credibility, and "Ideal Victims": Consent, Resistance, and Victim Blaming. *Canadian Journal of Women and the Law*, *22*(2), 397–433. https://doi.org/10.3138/cjwl.22.2.397

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

Valgarðsson, V. O., Clarke, N., Jennings, W., & Stoker, G. (2021). The Good Politician and Political Trust: An Authenticity Gap in British Politics? *Political Studies*, *69*(4), 858–880. https://doi.org/10.1177/0032321720928257

Walker, A. C., Turpin, M. H., Fugelsang, J. A., & Białek, M. (2021). Better the two devils you know, than the one you don't: Predictability influences moral judgments of immoral actors. *Journal of Experimental Social Psychology*, *97*, 104220. https://doi.org/10.1016/j.jesp.2021.104220

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