Journal of Behavioral Decision Making

WILEY

RESEARCH ARTICLE OPEN ACCESS

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

Cillian McHugh | Eric R. Igou

University of Limerick, Limerick, Ireland

Correspondence: Cillian McHugh (cillian.mchugh@ul.ie)

Received: 17 July 2024 | **Revised:** 18 July 2025 | **Accepted:** 21 July 2025

Keywords: dilution effect | MJAC | moral categorization | moral judgment | representativeness heuristic | typicality

ABSTRACT

It is reasonable to expect that when making a judgment, we only consider the relevant (or diagnostic) information and that nonrelevant (nondiagnostic) information should not, and thus does not, influence our judgments. Previous research has shown that this is not always the case and that the inclusion of nondiagnostic information can reduce the impact of diagnostic information in judgments. This phenomenon is known as the dilution effect, and it has been observed for a range of judgments, including product evaluations, probability judgments, and predictions relating to people's behavior. The dilution effect has been explained as a consequence of the representativeness heuristic, such that the inclusion of nondiagnostic information reduces the match between the target and a typical member of the category. Consistent with this notion and recent approaches to moral decision making, we predict that the dilution effect should be observed for judgments about morality. Across four studies (total $N = 2535$), 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 nondiagnostic information. We showed that this effect is not the result of humanization, and we 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.

Imagine your friend got mugged on holiday. Fortunately, a bystander saw and helped your friend afterwards. Now, imagine your friend describing the experience to you. Their description includes much detail, including nonrelevant information regarding both the mugger and their helper, such as "the mugger was wearing grey shoes" and "they [the helper] lived in 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 nonrelevant 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 nondiagnostic information leads to judgments that are less extreme than they would have been in the absence of nondiagnostic information (Nisbett et al. 1981; Zukier 1982). Applied to the above example, the presence of nondiagnostic information ("gray shoes"/"lived in 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 product evaluations (Igou and Bless 2005; Meyvis and Janiszewski 2002), probability judgments (LaBella and Koehler 2004), and predictions relating to people's behavior (Nisbett et al. 1981; Zukier 1982). However, to our knowledge, research has not directly tested whether the dilution effect occurs in 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 nondiagnostic 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").

One explanation of this effect is that it emerges as a consequence of the representativeness heuristic (Kahneman and 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 nondiagnostic 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 (nondiagnostic 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).

Although the representativeness heuristic is a popular explanation for the dilution effect, other approaches have put forward additional or entirely different explanations. For example, Igou and Bless (2005; see also Igou 2007; Schwarz 1994; Tetlock et al. 1996) suggested that conversational norms play a role in the emergence of dilution effects. Specifically, as in everyday conversations, communicated information ought to be relevant for the ongoing conversation (Grice 1975; see also Higgins 1981), even information with objectively low informational value might seem appropriate and necessary to complete the task at hand.

Sanborn et al. (2020) have put forward another approach. Specifically, they argue that the effect is not due to an error in combining diagnostic information with nondiagnostic information, but rather due to overestimation when interpreting only diagnostic information. Sanborn et al. developed a dilution effect task that included objectively correct answers. They found that when nondiagnostic information was present, participants were relatively accurate in their responding; however, when presented with diagnostic information only, participants appeared to overestimate the strength of evidence, appearing to fill in "missing" information in a biased fashion.

Regardless of whether the dilution effects are solely a result of the use of the representativeness heuristic, whether conversational rules play a role in the emergence of the effect, or whether the error is rooted in the overestimation of diagnostic or nondiagnostic information, all approaches predict that additional, nondiagnostic information reduces the impact of diagnostic information for the judgments or decision at hand.

1 | Predicting the Moral Dilution Effect

As moral judgments are crucial to human existence (Haidt 2012) and in everyday life (Gray and Graham 2018), it is important to know if dilution effects emerge in this domain. Despite the lack of previous work, there is good reason to expect the presence of a moral dilution. First, the original (Nisbett et al. 1981) study of the dilution effect (and indeed many subsequent studies, e.g., Peters and Rothbart 2000; Rempala and Geers 2011; Tetlock et al. 1996; Tetlock and Boettger 1989; Zukier 1982) involved judgments relating to personal traits, that is, the dilution effect appears to reliably influence person perception. Previous work has shown that various laws of person perception also apply to judgments of moral character (Johnson and Ahn 2021). As such, if the dilution effect impacts person perception, it should be observed for judgments of moral character.

Second, in light of recent work within moral psychology highlighting the dynamic and context-sensitive nature of moral judgments (e.g., Hester and Gray 2020; McHugh et al. 2022; Schein 2020), it is possible that the presence or absence of nondiagnostic information presents another possible source of variability in moral judgments, such that the dilution effect may be observed for moral judgments. Indeed, there are interesting parallels between the *representativeness heuristic* explanation of the effect described above (Nisbett et al. 1981) and recent theorizing about typicality variation in moral judgment (Gray and Keeney 2015; McHugh et al. 2022; Schein and Gray 2018).

Several authors have argued that some behaviors may be viewed as more typical (or representative) examples of *wrongness* or *rightness* than others (Gray and Keeney 2015; McHugh et al. 2022; Schein and Gray 2018). Gray and colleagues suggest that this variation is based on how closely a target aligns with a prototype of moral wrongness as "an intentional agent causing damage to a vulnerable patient" (Schein and Gray 2018, 33). It is plausible that including nondiagnostic information in a description of a target 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 and Gray 2018).

Typicality variability in moral judgment is also a core prediction of recent theorizing on moral judgment, specifically moral judgment as categorization (MJAC; McHugh et al. 2022). According to MJAC, when making a moral judgment, people are categorizing something as *morally right*, *morally wrong*, or *not morally relevant*. This can relate to actions and to actors/people (categorizing someone as a *good person* or as a *bad person*). One of the assumptions of this approach is that moral categorizations should show similar patterns of variability as nonmoral categorizations. Nonmoral categorization is known to show typicality variability (Barsalou 2003; McCloskey and Glucksberg 1978;

Oden 1977), and McHugh et al. (2022) argue that it should also be seen in the moral domain. However, they note that in the moral domain, typicality may be confounded with *severity*, posing a significant challenge to testing this prediction (McHugh et al. 2022). For instance, murder is likely a highly typical example of a member of the category *morally wrong*, whereas 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 and 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., nondiagnostic 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. We note that the emergence of the moral dilution effect in this way requires that there is a broad range of information that is societally agreed to be neutral or nondiagnostic. It is possible that as societies become more partisan and polarized, this suite of neutral/nondiagnostic information gets smaller (for discussion, see DellaPosta 2020) and smaller, undermining the possibility of a moral dilution effect.

2 | The Current Research

Based on the literature discussed above, we predict that the dilution effect should be observed for moral judgments. We present three studies where we test the dilution effect in judgments of moral character and a fourth study where we attempt to rule out an alternative explanation for our results. 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. All three Studies 1–3 were preregistered. A priori power analyses revealed that in order to detect a small effect ($f^2 = 0.01$) 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 Supporting Information, we report two pilot studies that informed the development of the stimulus materials used.[1] All three studies reported in the main text employ a within-subjects design. This approach is consistent with previous work (e.g., Peters and Rothbart 2000; Sanborn et al. 2020 report within-subjects designs only; Meyvis and Janiszewski 2002 report seven studies with a mixed design that closely resemble the within-subjects design reported here and elsewhere). We note that previous research has also examined the dilution effect using between-subjects designs only (e.g., Igou and Bless 2005; Kemmelmeier 2004; Rempala and Geers 2011; Tetlock et al. 1996; Tetlock and Boettger 1989) and that the earliest studies showing the effect demonstrated it across both between-subjects and within-subjects designs (Nisbett et al. 1981; Zukier 1982). Here, we chose to employ within-subjects designs in order to (i) minimize the potential confounding influence of both participant-level and item-level variability and (ii) improve statistical power and relatedly maximize available resources. We note one exception to this approach with Study S3, reported in the Supporting Information. Study 4 also employed a within-subjects design and examined the possible roles of relatability and humanness in moral dilution.

2.1 | Ethical Considerations

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.

3 | 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 only (morally relevant information), and two additionally contained nondiagnostic information (nonmorally 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 nondiagnostic information.

3.1 | Methods

3.1.1 | Participants and Design

Study 1 was a within-subjects design. The independent variable was information type with two levels, diagnostic information only (diagnostic) and nondiagnostic information additionally included (nondiagnostic). 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 nonbinary, 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 nonbinary, 5 other, 3 prefer not to say; $M_{age} = 26.25$, min = 18, max = 76, SD = 10.20).

3.1.2 | 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*, and *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 nondiagnostic information. The inclusion of nondiagnostic information was fully randomized across the four descriptions. The nondiagnostic 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 nondiagnostic descriptions (see Pilot Study 1 and Figure S9).

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); see Figure S1. 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*); see Figure S2.

We programmed our survey to randomly present nondiagnostic 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 nondiagnostic information. We hypothesized that the descriptions including nondiagnostic information would be rated as less severe than the diagnostic-only descriptions. Study 1 was preregistered at https://aspredicted.org/DVY_QN3. We note a deviation from our preregistered analysis in the analysis reported below. Our preregistration stated that we would test for differences using *t*-tests; however, this did not account for the actual data structure, specifically, for each participant, there were two observations in each condition, and the planned *t*-tests were not suitable. These are still reported in the Supporting Information (though they should be regarded with caution); however, in the analysis below, we describe linear-mixed-effects models instead, which allow us to better control for participant-level effects (due to multiple observations).

3.2 | Results and Discussion

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 preregistered exclusion criteria (failing both attention checks), and we proceed with the analyses as planned.

To test our hypothesis, we conducted a linear-mixed-effects model to test if information type 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 information type (with contrasts coded as $+1 = \textit{nondiagnostic}$ and $-1 = \textit{diagnostic}$); we allowed intercepts and the effect of information type 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, $\chi^2(8) = 816.91$, $p < 0.001$. Information type significantly influenced responses to the MPS-4, $F(1, 799.42) = 51.47$, $p < 0.001$, and was a significant predictor in the model when controlling for scenario, $b = -0.08$ ($\beta = -0.09$), $t(799.42) = -7.17$, $p < 0.001$, with the diagnostic descriptions being rated as more immoral than the nondiagnostic descriptions ($d = -0.16$); see Figure 1 and Table 1 (see also Figure S3 for combined MPS-4 and MM-1 results and Figure S4 for results for each character individually).

For our second analysis, we conducted a linear-mixed-effects model to test if information type influenced MM-1 responses. Our outcome measure was MM-1, and our predictor variable was information type (with contrasts coded as $+1 = \textit{nondiagnostic}$ and $-1 = \textit{diagnostic}$); we allowed intercepts and the effect of

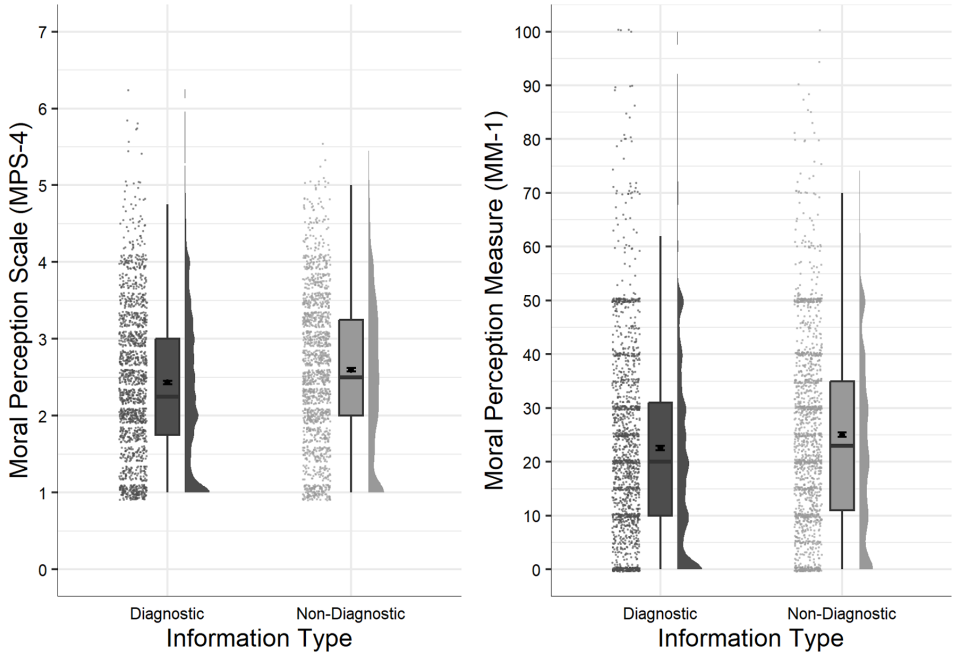

FIGURE 1 | Study 1: Differences in moral perception depending on information type.

information type to vary across participants. Overall, the model significantly predicted participants' responses and provided a better fit for the data than the baseline model, $\chi^2(8) = 475.52$, $p < 0.001$. Information type significantly predicted MM-1 responses, $F(1, 799.71) = 44.39$, $p < 0.001$, and, when controlling for scenario, was a significant predictor in the model, $b = -1.22$ ($\beta = -0.07$), $t(799.71) = -6.66$, $p < 0.001$, with the diagnostic descriptions being rated as more immoral than the nondiagnostic descriptions ($d = 0.16$); see Figure 1. (In the Supporting Information, we report the effect of information type 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 nondiagnostic information appears to reduce the match between a target and what is normally regarded as *representative* of *morally wrong*.

4 | 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 nondiagnostic information (nonmorally 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 nondiagnostic information.

4.1 | Methods

4.1.1 | Participants and Design

Study 2 was a within-subjects design. The independent variable was information type with two levels, diagnostic and nondiagnostic. 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 nonbinary, 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 the University of Limerick: $n = 533$ (female = 370, male = 147, nonbinary/other = 14, prefer not to say = 3, $M_{age} = 25.50$, SD = 9.60). In order to reach our preregistered target sample size, we recruited additional participants from Prolific: $n = 287$ (female = 96, male = 190, nonbinary/other = 1, prefer not to say = 1, $M_{age} = 35.70$, SD = 10.10). Participants from Prolific were paid $0.40 for their participation.

4.1.2 | 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

TABLE 1 | Differences in ratings of characters depending on information type for each study.

			Diagnostic		Nondiagnostic			95% confidence			Clustered variances			
Study	Val.	Measure	*M*	SD	*M*	SD	*d*	Upper	Lower	ID	ID × cond.	ID × val.	Residual	ICC
Study 1	Bad	MPS-4	2.4	1.0	2.6	1.0	−0.16	−0.09	−0.23	0.39	0.01	—	0.41	0.43
	Bad	MM-1	22.6	17.5	25	17.2	−0.16	−0.09	−0.23	181.12	2.65	—	101.56	0.61
Study 2	Good	MPS-4	6.1	1.0	6.1	1.0	0.00	0.07	−0.07	0.77	0.001	—	0.25	0.75
	Good	MM-1	84.3	14.8	84	14.7	0.02	0.09	−0.05	147.62	0.00	—	67.64	0.75
Study 3	Both	MPS-4	—	—	—	—	—	—	—	0.22	0.02	0.12	0.30	0.58
	Both	MM-1	—	—	—	—	—	—	—	62.17	3.21	82.75	84.16	0.58
	Bad	MPS-4	2.1	0.8	2.3	0.9	−0.23	−0.13	−0.33	0.37	—	—	0.33	0.52
	Bad	MM-1	20.3	17.1	22.6	17.1	−0.15	−0.05	−0.25	180.31	—	—	110.09	0.61
	Good	MPS-4	6.3	0.7	6.2	0.8	0.12	0.22	0.02	0.33	—	—	0.29	0.52
	Good	MM-1	88.1	11.9	86.7	13.9	0.13	0.22	0.03	107.12	—	—	59.38	0.64

> 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 nondiagnostic information (this was randomized through blocking; see https://osf.io/mdnpv/?view_only=77883e3fbc3d45f1a35fe92d5318cb67). The nondiagnostic 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 nondiagnostic material (see Pilot Study 2 and Figure S10). Study 2 was preregistered at https://aspredicted.org/NX2_HN6. As in Study 1, our analysis below includes a deviation from the preregistered planned analysis. Again, our preregistration stated that we would test for differences using *t*-tests; however, this was not suitable for the same reason as in Study 1 (for each participant, there were two observations in each condition). *T*-tests are reported in the Supporting Information (though they should be regarded with caution); however, in the analysis below, we describe the linear-mixed-effects models instead.

4.2 | Results and Discussion

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, although this is a potential source of error, based on the low proportions, we chose not to introduce any additional exclusions.

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

For our first analysis, our outcome measure was MPS-4, and our predictor variable was information type (with contrasts coded as $+1 = nondiagnostic$ and $-1 = diagnostic$); we allowed intercepts and the effect of information type to vary across participants, and scenario was also included in the model. Overall, the model

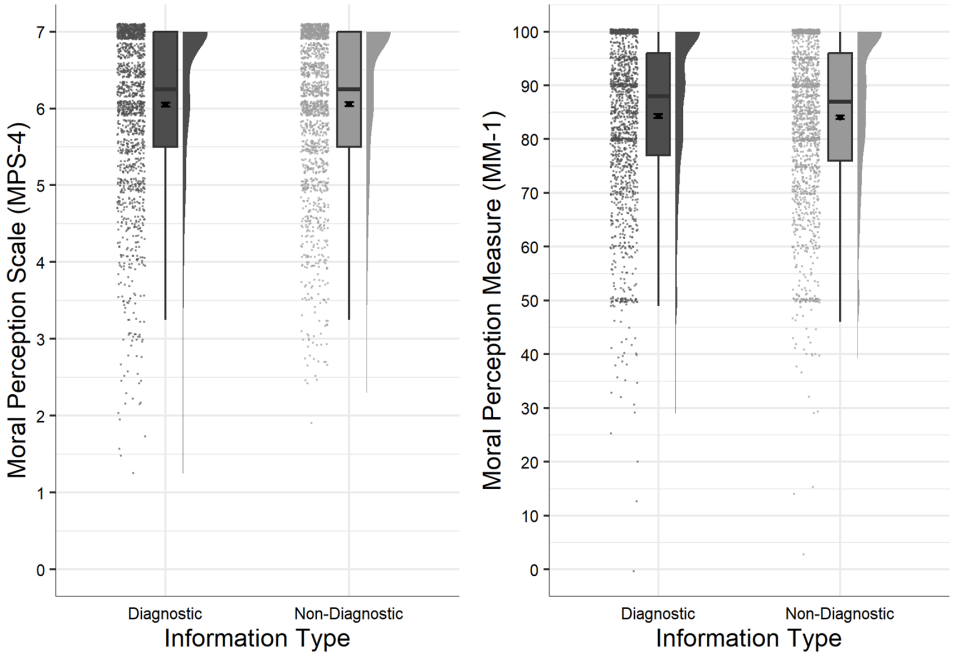

FIGURE 2 | Study 2: Differences in moral perception depending on information type.

significantly predicted participants' responses and provided a better fit for the data than the baseline model, $\chi^2(8) = 160.00$, $p < 0.001$. Information type did not influence responses to the MPS-4, $F(1, 838.12) = 0.24$, $p = 0.624$, and was not a significant predictor in the model when controlling for scenario, $b = 0.00$ ($\beta = 0.00$), $t(838) = 0.49$, $p = 0.624$ ($d = 0.00$); see Figure 2 and Table 1 (see also Figure S5 for combined MPS-4 and MM-1 results and Figure S6 for results for each character individually).

We conducted a linear-mixed-effects model to test if information type influenced MM-1 responses. Our outcome measure was MM-1, and our predictor variable was information type (with contrasts coded as +1 = *nondiagnostic* and −1 = *diagnostic*); we allowed intercepts and the effect of information type to vary across participants. Overall, the model significantly predicted participants' responses and provided a better fit for the data than the baseline model, $\chi^2(8) = 75.69$, $p < 0.001$. Information type did not influence MM-1 responses, $F(1, 2453) = 1.23$, $p = 0.267$, and was not a significant predictor in the model $b = 0.16$ ($\beta = 0.01$), $t(2453) = 1.11$, $p = 0.267$ ($d = 0.02$); see Figure 2. In the Supporting Information, we report the effect of information type 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 and 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. While developing a series of studies to test this explanation for the observed asymmetry is beyond the scope of the current paper, we test the replicability of this asymmetry in Study 3.

5 | Study 3—Good and Bad Actors

The aim of Study 3 was to test the replicability of the *good-bad* asymmetry in the incidence of the moral dilution effect. 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. We hypothesized that valence (good vs. bad) would interact with information type in producing a dilution effect, such that the dilution effect would be observed for bad actors but not for good actors. Study 3 was preregistered at https://aspredicted.org/QDF_XT1.

5.1 | Methods

5.1.1 | Participants and Design

Study 3 was a 2×2 within-subjects factorial design. The first independent variable was information type, with two levels, diagnostic and nondiagnostic. The second independent variable was the 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 nonbinary, 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 the 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 or did not complete all measures were removed ($n = 541$), 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).

5.1.2 | Procedure and Materials

Again, data were collected using an online questionnaire presented with Qualtrics (www.qualtrics.com). Participants were presented with four descriptions of 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 nondiagnostic 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 of the *good* and *bad* actors was randomly assigned to include nondiagnostic 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 preregistered at https://aspredicted.org/QDF_XT1.

5.2 | Results and Discussion

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. Again, these provide a possible source of error; however, the proportions remain relatively low.

In order to test the information type × valence interaction effect in a single model, we 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 and direction.

We note that using these recoded variables marks a deviation from the analysis described in the preregistration. However, this deviation was necessary to test our preregistered research question, that valence would moderate the dilution effect. If we used the raw scores of our dependent measures, our results would be confounded by the direction associated with valence, potentially leading to false positive results. Consider a case where, for both good and bad actors, there is a dilution effect, and it is of the same magnitude for both good and bad actors. Using the raw scores would suggest these dilution effects are different, because they present in opposite directions, and testing for an information type × valence interaction would likely yield a significant result. However, this apparent interaction is simply an artifact of the opposing directions, leading us to the false conclusion that the dilution effect is different for good and bad actors. As such, to provide an accurate test of our research question, we recoded the dependent measures.

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, $F(1, 813) = 48.53$, $p < 0.001$, partial $\eta^2 = 0.06$, 95% CI [0.03, 0.09], and a main effect for valence, $F(1, 813) = 3383.29$, $p < 0.001$, partial $\eta^2 = 0.81$, 95% CI [0.79, 0.82]. There was a significant condition × valence interaction effect, $F(1, 813) = 6.35$, $p = 0.012$, 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, $t(813) = 6.59$, $p < 0.001$ ($p_{adjusted} < 0.001$), $d = 0.23$, 95% CI [0.13, 0.25]. MPS-4 responses were higher in the nondiagnostic condition ($M = 2.33$, SD = 0.86) compared with the diagnostic condition ($M = 2.14$, SD = 0.81).

Similarly, for good characters, there were significant differences in MPS-4 responses depending on information type $t(813) = -3.43$, $p < 0.001$ ($p_{adjusted} = 0.003$), $d = 0.12$, 95% CI [−0.15, −0.04]. MPS-4 responses were lower in the nondiagnostic condition ($M = 6.21$, SD = 0.84) compared with 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, $p < 0.001$, partial $\eta^2 = 0.04$, 95% CI [0.01, 0.06], and a main effect for valence, $F(1, 813) = 258.78$, $p < 0.001$, partial $\eta^2 = 0.24$, 95% CI [0.19, 0.29]. There was no condition × valence interaction effect, $F(1, 813) = 1.78$, $p = 0.183$, partial $\eta^2 = 0.00$, 95% CI [0, 0.01], 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, $t(813) = 4.27$, $p < 0.001$ ($p_{adjusted} = < 0.001$), $d = 0.15$, 95% CI [1.22, 3.29]. MM-1 responses were higher in the nondiagnostic condition ($M = 22.60$, SD = 17.06) compared with the diagnostic condition ($M = 20.35$, SD = 17.13).

Similarly, for good characters, there were significant differences in MM-1 responses depending on information type, $t(813) = -3.60$, $p < 0.001$ ($p_{adjusted} = 0.001$), $d = 0.13$, 95% CI [−2.16, −0.64]. MM-1 responses were lower in the nondiagnostic condition ($M = 86.68$, SD = 13.86) compared with the diagnostic condition ($M = 88.08$, SD = 11.94).

We conducted a linear-mixed-effects model to test the effect of information type and valence on MPS-4R responses. Our outcome measure was the extremity of MPS-4R responses, and our predictor variables were information type (with contrasts coded as +1 = *nondiagnostic* and −1 = *diagnostic*) and valence (with contrasts coded as +1 = *good* and −1 = *bad*); we allowed intercepts and the effects of information type 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, $\chi^2(5) = 2695.60$, $p < 0.001$. Overall, there was a significant main effect for information type, $F(1, 813) = 48.53$, $p < 0.001$; valence significantly predicted responses, $F(1, 813) = 3383.29$, $p < 0.001$; and there was a significant information type × valence interaction, $F(1, 813) = 6.35$, $p = 0.012$.

We conducted a linear-mixed-effects model to test the effect of information type and valence on 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 responses. As above, our predictor variables were information type and valence; we allowed intercepts and the effects of information type 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, $\chi^2(5) = 749.17$, $p < 0.001$. Overall there was a main effect for information type, $F(1, 813) = 29.92$, $p < 0.001$; valence significantly predicted responses, $F(1, 813) = 258.78$, $p < 0.001$; and there was no significant information type × valence interaction, $F(1, 813) = 1.78$, $p = 0.183$. Below, we test for the presence of a dilution effect for bad and good characters separately.

5.2.1 | Differences in the "Bad" Descriptions

For the bad descriptions, we conducted a linear-mixed-effects model to test if information type influenced MPS-4 responses. Our outcome measure was MPS-4, and our predictor variable was information type (with contrasts coded as +1 = *nondiagnostic* and −1 = *diagnostic*); we allowed intercepts and the effect of information type to vary across participants. Overall, the model significantly predicted participants' responses and provided a better fit for the data than the baseline model, $\chi^2(3) = 76.88$, $p < 0.001$. Information type significantly influenced MPS-4 responses, $F(1, 812.00) = 46.02$, $p < 0.001$, and was a significant predictor in the model, $b = -0.10$ ($\beta = -0.11$), $t(812.00) = -6.78$, $p < 0.001$ ($d = -0.23$); see Figure 3.

We conducted a linear-mixed-effects model to test if the information type influenced MM-1 responses. Our outcome measure was MM-1, and our predictor variable was information type (with contrasts coded as +1 = *nondiagnostic* and −1 = *diagnostic*); we allowed intercepts and the effect of information type to vary across participants. Overall, the model significantly predicted participants' responses and provided a better fit for the data than the baseline model, $\chi^2(3) = 46.32$, $p < 0.001$. Information type significantly influenced MM-1 responses, $F(1, 812.00) = 19.25$, $p < 0.001$, and was a significant predictor in the model, $b = -1.14$ ($\beta = -0.06$), $t(812.00) = -4.39$, $p < 0.001$ ($d = -0.15$); see Figure 3 and Table 1 (see also Figure S7 for combined MPS-4 and MM-1 results and Figure S8 for results for each character individually).

5.2.2 | Differences in the "Good" Descriptions

For the good descriptions, we conducted a linear-mixed-effects model to test if information type influenced MPS-4 responses. Our outcome measure was MPS-4, and our predictor variable was information type (with contrasts coded as +1 = *nondiagnostic* and −1 = *diagnostic*); we allowed intercepts and the effect of information type to vary across participants. Overall, the model significantly predicted participants' responses and provided a better fit for the data than the baseline model, $\chi^2(3) = 30.01$, $p < 0.001$. Information type significantly influenced MPS-4 responses, $F(1, 812.00) = 11.87$, $p < 0.001$, and was a significant predictor in the model, $b = 0.05$ ($\beta = 0.06$), $t(812.00) = 3.45$, $p < 0.001$ ($d = 0.12$); see Figure 3.

We conducted a linear-mixed-effects model to test if the information type influenced MM-1 responses. Our outcome measure was MM-1, and our predictor variable was information type (with contrasts coded as +1 = *nondiagnostic* and −1 = *diagnostic*); we allowed intercepts and the effect of information type to vary across participants. Overall, the model significantly predicted participants' responses and provided a better fit for the data than the baseline model, $\chi^2(3) = 41.07$, $p < 0.001$. Information type significantly influenced MM-1 responses, $F(1, 812) = 13.21$, $p < 0.001$, and was a significant predictor in the model, $b = 0.69$ ($\beta = 0.05$), $t(812) = 3.63$, $p < 0.001$ ($d = 0.13$); see Figure 3.

The aim of Study 3 was to test if the moral dilution effect was moderated by the valence of the description. Based on the results of Studies 1 and 2, we predicted an information type × 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, although we did observe an information type × valence interaction effect, we also observed the dilution effect for both *bad* actors and *good* actors. This was unexpected given the results of Study 2; however, the investigation of the effect sizes suggests that the dilution effect is indeed stronger for the *bad* actors than for the *good* actors, which is in line with our prediction.

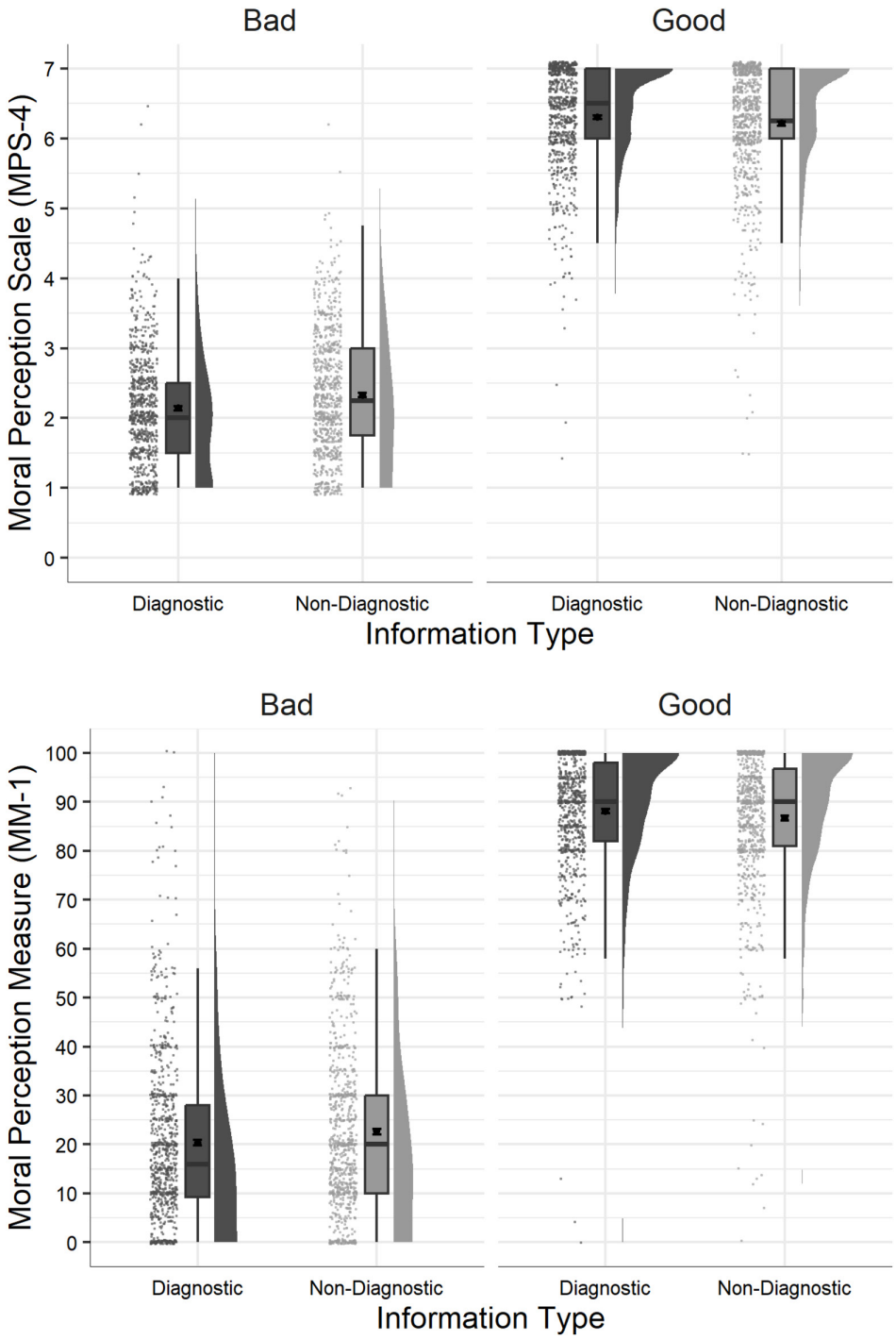

FIGURE 3 | Study 3: Differences in moral perception depending on information type.

6 | Meta-Analysis

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, whereas the second set of meta-analyses additionally includes the studies only reported in the Supporting Information. 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.

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).

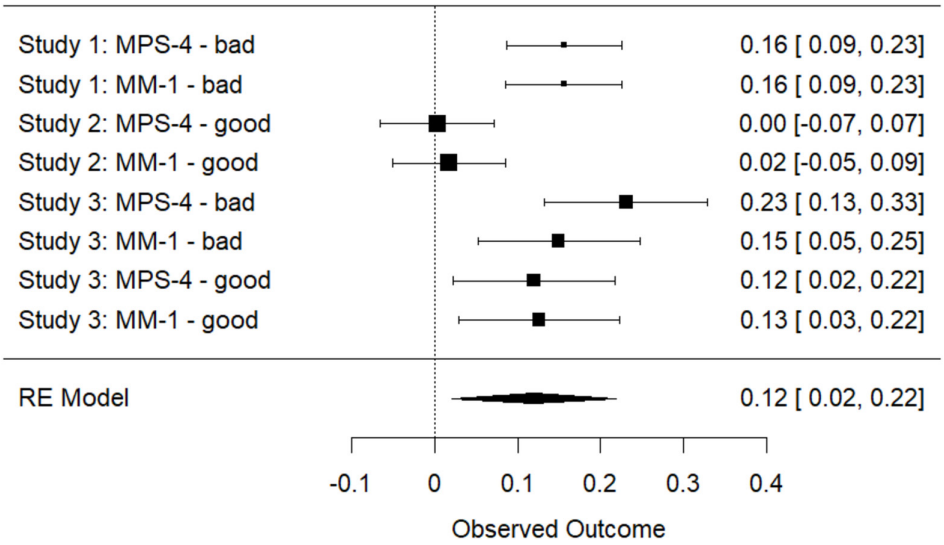

FIGURE 4 | Forest plot showing effect sizes for each study and pooled effect size.

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, $z = 2.38$, $p = 0.017$, 95% CI [0.02, 0.22]; see Figures 4 and S17.

Next, we 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 each measure. Overall, there was a significant dilution for bad characters across both studies for both measures, $d_{pooled} = -0.17$, SE = 0.02, $z = -7.97$, $p < 0.001$, 95% CI [−0.22, −0.13]; see Figure S18.

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 each measure. Overall, there was no significant dilution for good characters across both studies for either measure, $d_{pooled} = 0.06$, SE = 0.06, $z = 1.12$, $p = 0.263$, 95% CI [−0.05, 0.18]; see Figure S19.

We then proceeded to rerun the above meta-analyses but to additionally include the studies reported in the Supporting Information. 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, $d_{pooled} = 0.12$, SE = 0.03, $z = 3.99$, $p < 0.001$, 95% CI [0.06, 0.17]; see Figure S20.

Overall, there was a significant dilution for bad characters across all studies for both measures, $d_{pooled} = -0.12$, SE = 0.02, $z = -4.91$, $p < 0.001$, 95% CI [−0.17, −0.07]; see Figure S21. And, overall, there was a significant dilution for good characters, across all studies and for both measures, $d_{pooled} = 0.10$, SE = 0.04, $z = 2.91$, $p = 0.004$, 95% CI [0.03, 0.17]; see Figure S22.

These results show that the inclusion of nondiagnostic information leads to less extreme evaluations of moral character, suggesting that we have found evidence for a moral dilution effect. However, it is possible that our findings may be due to a humanizing effect, that is, the nondiagnostic information made the targets appear more *human*, and that this is what was driving our results. Study 4 was designed to test this possibility.

7 | Study 4—Humanizing or Diluting

To examine the possibility that the observed dilution effect may have been confounded by a humanizing effect of the nondiagnostic information rather than as a result of its dilution, we conducted a follow-up study with additional measures of both humanness and relatability. The assumption is that if the nondiagnostic information is having a humanizing effect, the humanness and relatability ratings for the nondiagnostic descriptions should be higher than for diagnostic descriptions (for both *bad* and *good* actors).

7.1 | Methods

7.1.1 | Participants and Design

Study 4 was a within-subjects design. The independent variable was information type with three levels: *good*, *bad*, and *neutral* (nondiagnostic). There were three dependent variables: morality ratings, humanness ratings, and relatability ratings.

For morality ratings, we predicted that morality ratings would be highest for the *good* descriptions and lowest for the *bad* descriptions, with the *neutral* descriptions being rated lower than *good* and higher than *bad*.

If the dilution effect is being driven by a humanizing effect of nondiagnostic information, this would predict a different pattern for humanness and relatability ratings. Specifically, humanness and relatability ratings would be expected to be highest for the *neutral* (nondiagnostic) conditions compared with the *bad* and *good* conditions. If this pattern is not observed, it suggests that our findings may be evidence of a dilution effect rather than a humanizing effect.

A total sample of 100 (41 female, 59 male; $M_{age} = 40.87$, min = 19, max = 70, SD = 12.65) completed the survey. Participants were recruited from Prolific and paid £0.40 for their participation.

7.1.2 | Procedure and Materials

The materials were similar to those used in Study 3. Participants read six descriptions, two *good*, two *bad*, and two *neutral* (nondiagnostic). The content of these descriptions was taken directly from Study 3. For each description, participants rated the character's morality, humanness, and relatability. The morality rating was the same as the MM-1 rating used in previous studies. The humanness and relatability ratings were adapted from Martin and Mason (2022). Participants read the following instructions: "The following scale represents humanness levels. Please rate the humanness of X", 0 = *not at all human*, 100 = *very human*; "The following scale represents relatableness levels. Please rate the relatableness of X", 0 = *not at all Relatable*, 100 = *very relatable*.

Sam

> 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.

Robin

> Imagine a person named Robin. 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.

Francis

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

Alex

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

Jordan

> Imagine a person named Jordan. They have red hair, play tennis four times a month, and have one older sibling and one younger sibling.

Charlie

> Imagine a person named Charlie. They are left-handed, drink tea in the morning, and have two older siblings and one younger sibling.

7.2 | Results and Discussion

7.2.1 | Morality

We conducted a linear-mixed-effects model to test if information type influenced morality ratings. Our outcome measure was morality rating; our predictor variable was valence/information-type; we allowed intercepts to vary across participants. Overall, the model significantly predicted participants' responses and provided a better fit for the data than the baseline model, $\chi^2(2) = 945$, $p < 0.001$. Information type significantly influenced morality ratings, $F(2, 498) = 1217$, $p < 0.001$. Tukey's post hoc pairwise comparisons indicated that the highest rated descriptions were the good descriptions ($M = 89$, SD = 13), and these were significantly higher ($p < 0.001$) than the bad descriptions ($M = 14$, SD = 18) and significantly higher ($p < 0.001$) than the neutral/nondiagnostic descriptions ($M = 67$, SD = 17). The neutral/nondiagnostic descriptions were also significantly higher ($p < 0.001$) than the bad descriptions.

This pattern aligns with how the materials were designed (bad character viewed as least moral, good character viewed as most moral) and is consistent with the results of Studies 1–3. Interestingly, the morality of the neutral descriptions is rated closer to the good descriptions than to the bad descriptions. It is possible that this may partially explain the asymmetry observed in the occurrence of the dilution effect.

7.2.2 | Humanness and Relatability

We conducted a linear-mixed-effects model to test if information type influenced humanness ratings. Our outcome measure was humanness rating; our predictor variable was valence/information-type; we allowed intercepts to vary across participants. Overall, the model significantly predicted

participants' responses and provided a better fit for the data than the baseline model, $\chi^2(2) = 616$, $p < 0.001$. Information type significantly influenced humanness ratings, $F(2, 498) = 602$, $p < 0.001$. Tukey's post hoc pairwise comparisons indicated that the highest rated descriptions were the good descriptions ($M = 89$, SD = 13), and these were significantly higher ($p < 0.001$) than the bad descriptions ($M = 30$, SD = 29) and significantly higher ($p < 0.001$) than the neutral/nondiagnostic descriptions ($M = 79$, SD = 18). The neutral/nondiagnostic descriptions were also significantly higher ($p < 0.001$) than the bad descriptions.

We conducted a linear-mixed-effects model to test if information type influenced relatability ratings. Our outcome measure was relatability, our predictor variable was valence/information-type; we allowed intercepts across participants. Overall, the model significantly predicted participants' responses and provided a better fit for the data than the baseline model, $\chi^2(2) = 655$, $p < 0.001$. Information type significantly influenced relatability ratings, $F(2, 498) = 666$, Tukey's post hoc pairwise comparisons indicated that the highest rated descriptions were the good descriptions ($M = 80$, SD = 18), and these were significantly higher ($p < 0.001$) than the bad descriptions ($M = 19$, SD = 23) and significantly higher ($p < 0.001$) than the neutral/nondiagnostic descriptions ($M = 69$, SD = 19). The neutral/nondiagnostic descriptions were also significantly higher ($p < 0.001$) than the bad descriptions.

For both humanness and relatability ratings, the pattern of results suggests that the observed dilution effect cannot be entirely attributed to a humanizing effect. We predicted that if the nondiagnostic information has a humanizing effect, the humanness and relatability ratings for the nondiagnostic descriptions should be higher than for the diagnostic descriptions (for both *bad* and *good* actors). We found that humanness and relatability ratings for the nondiagnostic characters were higher than for the *bad* actors, but lower than for the *good* actors. This suggests that for the *bad* actors, it is possible that the observed dilution effect may be confounded by a humanizing effect; however, for the *good* actors, this is not necessarily the case. Interestingly, we did find an asymmetry in the reliability of the dilution effect in *good* versus *bad* actors, with a more reliable dilution effect occurring for *bad* actors than for *good* actors. This asymmetry may be (at least partially) attributed to a humanizing effect for the *bad* actors.

The results of Study 4 suggest that the moral dilution effect cannot be attributed to a humanizing effect; however, we note that there is another potential interpretation of our results. It is possible that participants see humanness/relatability as a cue or a proxy for morality. In Study 4, for all three DVs, the *good* condition was the highest, the *bad* condition was the lowest, and the *neutral/nondiagnostic* condition was in the middle (and closer to the *good* condition than to the *bad* condition). When integrating descriptions that include *good*/*bad* information with *neutral/nondiagnostic* information into a summary judgment, participants may form an "average" positive/negative rating, combining the ratings for the different parts of the description without necessarily distinguishing between diagnostic versus nondiagnostic information. That is, the diluting effect of nondiagnostic information would not be due to its nondiagnosticity but due to its modest humanness and relatability, implying modest morality. Although our results cannot rule out this possibility, previous research on the dilution effect suggests that this is unlikely to be the case. For example, under this interpretation, it is not clear why manipulations of accountability would influence the strength of the dilution effect, as has been shown in previous studies (Tetlock and Boettger 1989). Furthermore, previous work has shown that in some situations, participants can identify and even disregard nondiagnostic information (Igou and Bless 2005). Future research is needed to test if these effects replicate for *moral* dilution and to better understand the possible role of humanization in the moral dilution effect.

8 | General Discussion

Across three studies, we provide evidence for a moral dilution effect—the presence of nondiagnostic 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. Our findings are consistent with categorization approaches to moral judgment that predict that moral judgments vary in their typicality (McHugh et al. 2022; Schein and Gray 2018) or their representativeness—the inclusion of nondiagnostic information in a description of a moral actor reduces the representativeness of the target, leading to less extreme judgments of the actors (Kahneman and Tversky 1972; McHugh et al. 2022; Nisbett et al. 1981; Schein and Gray 2018). It is also consistent with predictions from research in person perception, whereby known effects within person perception are predicted to also apply to perceptions of moral character (Johnson and Ahn 2021).

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 comparable effects across moral and nonmoral domains (for discussion, see McHugh et al. 2022). In line with this, MJAC and the theory of dyadic morality (TDM) predict 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* (McHugh et al. 2022; Schein and Gray 2018). The studies presented here provide empirical evidence consistent with this prediction. In addition, the paradigm developed here can be applied and developed further to examine the moral dilution effect in future research, for example, 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 and Gray 2018).

Interestingly, the dilution effect appears to be stronger for judgments of *bad* actors, as compared with judgments of *good* actors (Study 3), and the effect also appears to be more robust

for judgments of *bad* characters than *good* characters (no effect observed in Study 2). We suggest three possible explanations for this asymmetry. First, in Study 4, we found that the morality ratings for the neutral descriptions were closer to the *good* descriptions than to *bad* descriptions. This results in a larger difference in ratings between *bad* descriptions and nondiagnostic descriptions, such that the effect of nondiagnostic information on *bad* descriptions is stronger than on *good* descriptions. Second, one interpretation of the results of Study 4 suggests that the nondiagnostic information leads to a humanizing effect for the *bad* descriptions, but less so for the *good* descriptions. Third, 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 with what is representative of morally *wrong*. Future research should explore these possibilities 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 Supporting Information). We also note that despite the observed variability across studies, the results of the meta-analysis provide strong support for our conclusions.

9 | 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 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 more thoroughly than *good* (Baumeister et al. 2001; Pratto and 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 nondiagnostic information). Future research should replicate and extend these findings, investigating this in more detail.

We also note that in all studies, there were instances where participants provided responses that were inconsistent with the descriptions provided (i.e., rating *bad* characters as *good* and vice-versa). Although overall, the proportions were relatively low, this does present a potential limitation with the current studies, which should be addressed in future research.

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.

A further limitation may be that our materials are based on moral foundations, such that each description included three of the five foundations. People vary in the degree to which they regard the different foundations as important, for example, liberals show greater endorsement of the harm and fairness foundations (often referred to as individualizing foundations), whereas conservatives additionally endorse the authority, loyalty, and purity foundations (often referred to as binding foundations), tending to endorse all five foundations more equally (see Graham et al. 2009). As such, it is possible that some participants (liberal participants in particular) did not view some descriptions as morally relevant. However, there is some evidence to suggest that this variation between participants is a variation in *strength* of endorsement, rather than variation in actual endorsement (i.e., whether or not a foundation matters at all). For example, although liberals tend to place less emphasis on the binding foundations, they still endorse them to some degree—a secondary analysis of publicly available data across 30 countries from Klein et al. (2017) shows differences between liberals and conservatives in the strength of endorsement of different foundations; however, for both liberals and conservatives, the mean score for all foundations is greater than 3 (where 3 = *slightly relevant*). Furthermore, in designing our materials, we included at least one individualizing and at least one binding foundation in each description, with the aim that all descriptions would contain at least one foundation that holds moral relevance for all participants. Future research should examine if the dilution effect exists for a wider range of moral descriptions, as well as any possible interactions with political ideology (we note that we did not include a measure of participants' political ideology in our studies).

Our findings provide some insight into the role of humanization in the moral dilution effect. We showed that in the case of morally bad characters, the dilution effect could be partially attributed to a humanizing effect; however, this was not the case for morally good characters. We also found that ratings of morality, humanness, and relatability followed similar patterns. This similarity is interesting, and future research should examine it further to better understand what it means for the moral dilution effect.

Another area of future research could examine the dilution effect in the context of more concrete moral characters and whether or not the effect occurs equally for *perpetrators*, *victims*, and *nonvictims* across different transgressions. For example, previous research has shown that victims are perceived as more moral than nonvictims who commit the same acts (Jordan and Kouchaki 2021). It is possible that perceiving victims as more moral leads to more stable evaluations of victims that are less susceptible to the influence of nondiagnostic information (i.e., reduced incidence of the dilution effect for victims). Relatedly, the asymmetry we found in our studies suggests that the dilution effect may occur more readily for perpetrators. These are questions that can be tested empirically with future research.

10 | 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, and 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 and 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 and 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).

11 | 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 are consistent with a categorization approach to understanding moral judgment (contributing to theory building) and also have practical implications across a range of real-world settings.

Author Contributions

Cillian McHugh: project conception, study design, data collection, analysis and results, interpretation of findings, drafting manuscript. **Eric R. Igou:** project conception, study design, interpretation of findings, review of manuscript.

Ethics Statement

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. All authors consented to the submission of this manuscript.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are openly available in Moral Dilution at https://osf.io/mdnpv/, reference number DOI 10.17605/OSF.IO/MDNPV.

Endnotes

[1] We also report in the *Supporting Information* 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). See Figures S11–S16 for results. 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 subsample from Prolific in Study 2 all participants are drawn from convenience/snowball samples from the student body and wider community associated with the University of Limerick.

References

Barsalou, L. W. 2003. "Situated Simulation in the Human Conceptual System." *Language & Cognitive Processes* 18, no. 5–6: 513–562. https://doi.org/10.1080/01690960344000026.

Baumeister, R. F., E. Bratslavsky, C. Finkenauer, and K. D. Vohs. 2001. "Bad Is Stronger Than Good." *Review of General Psychology* 5, no. 4: 323–370. https://doi.org/10.1037/1089-2680.5.4.323.

DellaPosta, D. 2020. "Pluralistic Collapse: The "Oil Spill" Model of Mass Opinion Polarization." *American Sociological Review* 85, no. 3: 507–536. https://doi.org/10.1177/0003122420922989.

Freeman, H. 2018. "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., and E. L. Paluck. 2018. "What Is the Psychological Appeal of the Serial Rapist Model? Worldviews Predicting Endorsement." Behavioral Public Policy. https://papers.ssrn.com/abstract=3190670.

Graham, J., J. Haidt, and B. A. Nosek. 2009. "Liberals and Conservatives Rely on Different Sets of Moral Foundations." *Journal of Personality and Social Psychology* 96, no. 5: 1029–1046. https://doi.org/10.1037/a0015141.

Gray, K. J., and J. Graham, eds. 2018. *Atlas of Moral Psychology*. Guilford Press.

Gray, K. J., and J. E. Keeney. 2015. "Impure or Just Weird? Scenario Sampling Bias Raises Questions About the Foundation of Morality."

Social Psychological and Personality Science 6, no. 8: 859–868. https://doi.org/10.1177/1948550615592241.

Gray, K. J., A. Waytz, and L. Young. 2012. "The Moral Dyad: A Fundamental Template Unifying Moral Judgment." *Psychological Inquiry* 23, no. 2: 206–215. https://doi.org/10.1080/1047840X.2012.686247.

Grice, H. P. 1975. "Logic and Conversation." In *Syntax and Semantics 3: Speech Arts*, edited by C. Peter and J. Morgan, 41–58. Academic Press. https://cir.nii.ac.jp/crid/1571135649252606592.

Grizzard, M., K. Fitzgerald, C. J. Francemone, et al. 2020. "Validating the Extended Character Morality Questionnaire." *Media Psychology* 23, no. 1: 107–130. https://doi.org/10.1080/15213269.2019.1572523.

Haidt, J. 2012. *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Vintage.

Hester, N., and K. Gray. 2020. "The Moral Psychology of Raceless, Genderless Strangers." *Perspectives on Psychological Science* 15, no. 2: 216–230. https://doi.org/10.1177/1745691619885840.

Higgins, E. T. 1981. "The "Communication Game": Implications for Social Cognition and Persuasion." In *Social Cognition*, edited by E. T. Higgins, C. P. Herman, and M. P. Zanna, 343–392. Routledge.

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. 2007. "Additional Thoughts on Conversational and Motivational Sources of the Dilution Effect." *Journal of Language and Social Psychology* 26, no. 1: 61–68. https://doi.org/10.1177/0261927X06296473.

Igou, E. R., and H. Bless. 2005. "The Conversational Basis for the Dilution Effect." *Journal of Language and Social Psychology* 24, no. 1: 25–35. https://doi.org/10.1177/0261927X04273035.

Johnson, S. G. B., and J. Ahn. 2021. "Principles of Moral Accounting: How Our Intuitive Moral Sense Balances Rights and Wrongs." *Cognition* 206: 104467. https://doi.org/10.1016/j.cognition.2020.104467.

Jordan, J. J., and M. Kouchaki. 2021. "Virtuous Victims." *Science Advances* 7, no. 42: eabg5902. https://doi.org/10.1126/sciadv.abg5902.

Kahneman, D., and A. Tversky. 1972. "Subjective Probability: A Judgment of Representativeness." *Cognitive Psychology* 3, no. 3: 430–454. https://doi.org/10.1016/0010-0285(72)90016-3.

Kemmelmeier, M. 2004. "Separating the Wheat From the Chaff: Does Discriminating Between Diagnostic and Nondiagnostic Information Eliminate the Dilution Effect?" *Journal of Behavioral Decision Making* 17, no. 3: 231–243. https://doi.org/10.1002/bdm.473.

Klein, R. A., M. Vianello, F. Hasselman, et al. 2017. "Many Labs 2: Investigating Variation in Replicability Across Sample and Setting." https://osf.io/ux3eh/.

LaBella, C., and D. J. Koehler. 2004. "Dilution and Confirmation of Probability Judgments Based on Nondiagnostic Evidence." *Memory & Cognition* 32, no. 7: 1076–1089. https://doi.org/10.3758/BF03196883.

Levin, S. 2016. "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/.

Martin, A. E., and M. F. Mason. 2022. "What Does It Mean to Be (Seen as) Human? The Importance of Gender in Humanization." *Journal of Personality and Social Psychology* 123, no. 2: 292–315. https://doi.org/10.1037/pspa0000293.

McCloskey, M. E., and S. Glucksberg. 1978. "Natural Categories: Well Defined or Fuzzy Sets?" *Memory & Cognition* 6, no. 4: 462–472. https://doi.org/10.3758/BF03197480.

McHugh, C., M. McGann, E. R. Igou, and E. L. Kinsella. 2022. "Moral Judgment as Categorization (MJAC)." *Perspectives on Psychological Science* 17, no. 1: 131–152. https://doi.org/10.1177/1745691621990636.

McKay, S. 2018. "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., and C. Janiszewski. 2002. "Consumers' Beliefs About Product Benefits: The Effect of Obviously Irrelevant Product Information." *Journal of Consumer Research* 28, no. 4: 618–635. https://doi.org/10.1086/338205.

Nisbett, R. E., H. Zukier, and R. E. Lemley. 1981. "The Dilution Effect: Nondiagnostic Information Weakens the Implications of Diagnostic Information." *Cognitive Psychology* 13, no. 2: 248–277. https://doi.org/10.1016/0010-0285(81)90010-4.

Oden, G. C. 1977. "Fuzziness in Semantic Memory: Choosing Exemplars of Subjective Categories." *Memory & Cognition* 5, no. 2: 198–204. https://doi.org/10.3758/BF03197362.

Peters, E., and M. Rothbart. 2000. "Typicality Can Create, Eliminate, and Reverse the Dilution Effect." *Personality and Social Psychology Bulletin* 26, no. 2: 177–187. https://doi.org/10.1177/0146167200264005.

Pratto, F., and O. P. John. 1991. "Automatic Vigilance: The Attention-Grabbing Power of Negative Social Information." *Journal of Personality and Social Psychology* 61, no. 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, no. 2: 397–433. https://doi.org/10.3138/cjwl.22.2.397.

Rempala, D. M., and A. L. Geers. 2011. "The Influence of Nondiagnostic Information and Victim Stereotypes on Perceptions of Guilt." *Western Criminology Review* 12, no. 3: 90–105.

Sanborn, A. N., T. Noguchi, J. Tripp, and N. Stewart. 2020. "A Dilution Effect Without Dilution: When Missing Evidence, Not Non-Diagnostic Evidence, Is Judged Inaccurately." *Cognition* 196: 104110. https://doi.org/10.1016/j.cognition.2019.104110.

Schein, C. 2020. "The Importance of Context in Moral Judgments." *Perspectives on Psychological Science* 15, no. 2: 207–215. https://doi.org/10.1177/1745691620904083.

Schein, C., and K. J. Gray. 2018. "The Theory of Dyadic Morality: Reinventing Moral Judgment by Redefining Harm." *Personality and Social Psychology Review* 22, no. 1: 32–70. https://doi.org/10.1177/1088868317698288.

Schwarz, N. 1994. "Judgment in a Social Context: Biases, Shortcomings, and the Logic of Conversation." In *Advances in Experimental Social Psychology*, edited by M. P. Zanna, Vol. 26, 123–162. Academic Press. http://www.sciencedirect.com/science/article/pii/S0065260108601537.

Tetlock, P. E., and R. Boettger. 1989. "Accountability: A Social Magnifier of the Dilution Effect." *Journal of Personality and Social Psychology* 57, no. 3: 388–398. https://doi.org/10.1037/0022-3514.57.3.388.

Tetlock, P. E., J. S. Lerner, and R. Boettger. 1996. "The Dilution Effect: Judgmental Bias, Conversational Convention, or a Bit of Both?" *European Journal of Social Psychology* 26, no. 6: 915–934. https://doi.org/10.1002/(SICI)1099-0992(199611)26:6<915::AID-EJSP797>3.0.CO;2-W.

Valgarðsson, V. O., N. Clarke, W. Jennings, and G. Stoker. 2021. "The Good Politician and Political Trust: An Authenticity Gap in British Politics?" *Political Studies* 69, no. 4: 858–880. https://doi.org/10.1177/0032321720928257.

Walker, A. C., M. H. Turpin, J. A. Fugelsang, and M. Białek. 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, no. 6: 1163–1174. https://doi.org/10.1037/0022-3514.43.6.1163.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Figure S1.** Screenshot of the MPS-4 items as presented to participants. **Figure S2.** Screenshot of MM-1 as presented to participants. **Figure S3.** Study 1: Differences in combined measure depending on condition. **Figure S4.** Study 1: Differences in moral perception for each description. **Figure S5.** Study 2: Differences in combined measure depending on condition. **Figure S6.** Study 2: Differences in moral perception for each description. **Figure S7.** Study 3: Differences in the combined measure depending on condition. **Figure S8.** Study 3: Differences in moral perception for each description. **Figure S9.** Pilot Study 1: Differences in moral perception depending on condition. **Figure S10.** Pilot Study 2: Differences in moral perception depending on condition. **Figure S11.** Study S1: Responses to moral perception measures depending on condition. **Figure S12.** Study 2: Differences in moral perception for each description. **Figure S13.** Study S2: Differences in moral perception depending on condition. **Figure S14.** Study S2: Differences in moral perception for each description. **Figure S15.** Study S3: Differences in moral perception depending on condition. **Figure S16.** Study S3: Differences in moral perception for each description. **Figure S17.** Forest plot showing effects for Studies 1–3 and pooled effect. **Figure S18.** Forest plot showing effects for bad characters for Studies 1 and 3 and pooled effect. **Figure S19.** Forest plot showing effects for good characters Studies 2 and 3 and pooled effect. **Figure S20.**Forest plot showing effects for all studies and pooled effect. **Figure S21.**Forest plot showing effects for bad characters for all studies and pooled effect. **Figure S22.** Forest plot showing effects for good characters for all studies and pooled effect.