Supplement: Moral Dilution

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

[1] University of Limerick

Author Note

Department of Psychology, University of Limerick. 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. 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, Limerick, V94 T9PX. E-mail: cillian.mchugh@ul.ie

Abstract

Supplementary analysis to accompany the manuscript The Moral Dilution Effect: Irrelevant Information Influences Judgments of Moral Character.

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

Word count: 11,119

Supplement: Moral Dilution

Contents

Materials

Descriptions (Bad Characters)

Diagnostic Descriptions. Each moral description contains descriptive information relating to three different moral foundations as follows: *Sam*: care, fairness, loyalty; *Robin*: care, fairness, loyalty; *Francis*: purity, authority, fairness; *Alex*: care, fairness, authority.

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

Robin. 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.

Francis. 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.

Alex. 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.

Non-Diagnostic Descriptions.

Jackie. Imagine a person named Jackie. 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.

Descriptions (Good Characters)

Diagnostic Descriptions. Each moral description contains descriptive information relating to three different moral foundations as follows: *Sam*: care, fairness, loyalty; *Robin*: care, fairness, loyalty; *Francis*: purity, authority, fairness; *Alex*: care, fairness, authority.

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 show compassion and empathy for others, act with a sense of fairness and justice, and, never to break their word.

Francis. 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.

Alex. 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.

Non-Diagnostic

Jackie. Imagine a person named Jackie. They have dark hair, go for a jog twice a week, and their favorite color is blue.

Charlie. Imagine a person named Charlie. They have blue eyes, drink coffee in the morning, and their favorite color is green.

Descriptions (Both Good and Bad)

Diagnostic Descriptions.

Sam (good). 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 (good). 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.

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

Francis (bad). 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.

Non Diagnostic Descriptions. They have red hair, play tennis four times a month, and have one older sibling and one younger sibling.

They are left-handed, drink tea in the morning, and have two older siblings and one younger sibling.

Measures

Four-item Moral Perception Scale (MPS-4). Please rate _____ along the following dimensions:

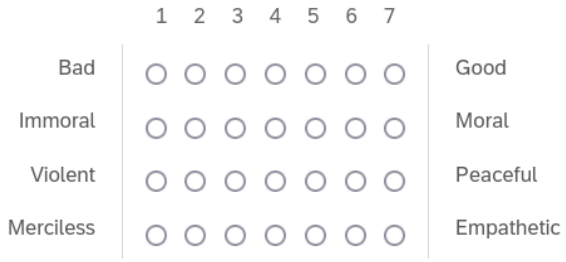

Figure S1. Screenshot of the MPS-4 items as presented to participants

Single-item Moral Perception Measure (MM-1). Please rate _____ according to immoral or moral you view them:

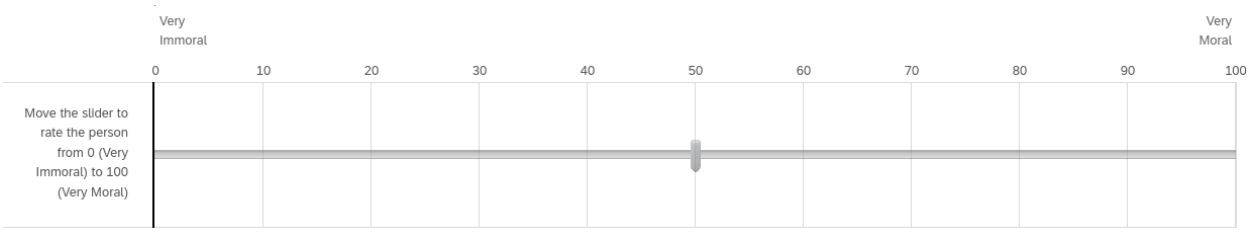

Figure S2. Screenshot of MM-1 as presented to participants

Study 1 (bad): Supplementary Analyses

Study 1: T-tests

A paired samples t-test indicated that there were significant differences in MPS-4 responses depending on information type $t(1601)$, = -6.27, $p < .001$ $d = 0.17$, 95% CI [-0.22, -0.11]. MPS-4 responses were higher in the non-diagnostic condition ($M = 2.60$, $SD = 0.96$) compared to the diagnostic condition ($M = 2.43$, $SD = 0.97$).

A paired samples t-test indicated that there were significant differences in MM1 responses depending on information type $t(1601)$, = -6.24, $p < .001$ $d = 0.14$, 95% CI [-3.24, -1.69]. MM-1 responses were higher in the non-diagnostic condition ($M = 25.03$, $SD = 17.15$) compared to the diagnostic condition ($M = 22.57$, $SD = 17.53$).

Study 1: Descriptives for Each Scenario

The means and standard deviations for MPS-4 for each scenario are as follows: *Sam*, $M_{\text{MPS-4}} = 2.55$, $SD_{\text{MPS-4}} = 0.86$, *Francis*, $M_{\text{MPS-4}} = 3.05$, $SD_{\text{MPS-4}} = 0.97$, *Alex*, $M_{\text{MPS-4}} = 2.32$, $SD_{\text{MPS-4}} = 0.88$, *Robin*, $M_{\text{MPS-4}} = 2.13$, $SD_{\text{MPS-4}} = 0.91$. 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 characters (all ps $< .001$), while *Robin* was rated as less moral than each of the other characters (all ps $< .001$), while *Sam* was rated more favorably than *Alex* ($p < .001$).

The means and standard deviations for MM-1 for each scenario are as follows: *Sam*, $M_{\text{MM-1}} = 23.94$, $SD_{\text{MM-1}} = 16.18$; *Francis*, $M_{\text{MM-1}} = 30.12$, $SD_{\text{MM-1}} = 17.86$; *Alex*, $M_{\text{MM-1}} = 20.55$, $SD_{\text{MM-1}} = 16.65$; *Robin*, $M_{\text{MM-1}} = 20.60$, $SD_{\text{MM-1}} = 17.06$. 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 characters ($p < .001$), *Sam* was the next most favorably rated character, rated significantly more favorably than both *Alex* and

Robin (*p*s < .001), there was no difference between *Alex* and *Robin* (*p* = = .953).

Study 1: Combined Measure

We developed a combined moral perception measure by calculating the mean of the combined mean-centered scores for MPS-4 and MM-1, and mean-centering this result. Below we report the analyses for this combined measure.

The means and standard deviations for the combined measure for each scenario are as follows: *Sam*, $M = 0.02$, $SD = 0.89$, *Francis*, $M = 0.48$, $SD = 1.00$, *Alex*, $M = -0.21$, $SD = 0.92$, *Robin*, $M = -0.32$, $SD = 0.94$. There was significant variation depending on the description, $F(3,2255) = 269.01$, $p < .001$, partial $\eta^2 = 0.10$. *Francis* appeared to be rated as the most favorable, followed by *Sam*, then *Alex* and finally *Robin* as the least favorable (all *p*s < .001).

We conducted a linear-mixed-effects model to test if condition influenced moral perception. Our outcome measure was the combined moral perception measure, 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, $\chi^2(8) = 762.31$, $p < .001$. Condition significantly influenced responses to the MPS-4, $F(1, 799.66) = 57.93$, $p < .001$; and was a significant predictor in the model when controlling for scenario, $b = -0.08$, $t(2,501.32) = -3.42$, $p < .001$, with the non-diagnostic descriptions being rated as more moral than the diagnostic (morally relevant) descriptions of immoral characters Figure S3.

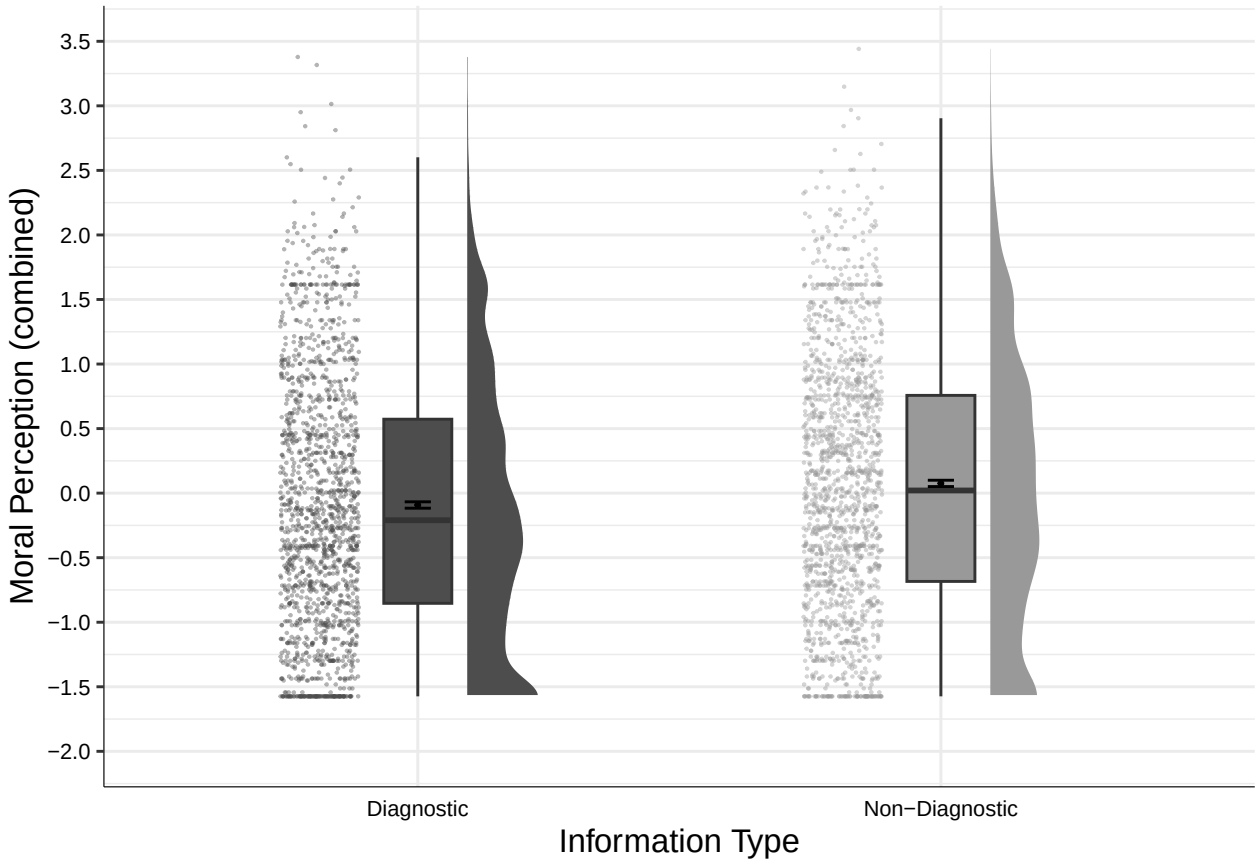

Figure S3. Study 1: Differences in combined measure depending on condition

Study 1: Alternative Model

In the analyses reported in the main text we included random effects for participant, as well as participant × condition random effects (allowing the effect of condition to vary across participants). When we attempted to additionally include random effects for scenario, the model failed to converge. Here we report alternative analyses where we replace the participant × condition random effects with random effects for scenario.

First we conducted a linear-mixed-effects model to test if condition influenced MPS-4 responses. Our outcome measure was MPS-4, our predictor variable was condition, and scenario was also included in the model; we allowed intercepts to vary across participants, and across scenarios.

Overall, the model significantly predicted participants responses, and provided a

better fit for the data than the baseline model, $\chi^2(8) = 816.21$, $p < .001$. Condition significantly influenced responses to the MPS-4, $F(1, 2{,}395.94) = 53.21$, $p < .001$; and was a significant predictor in the model when controlling for scenario, b = -0.08, $t(2{,}395.94)$ = -7.29, $p < .001$, with the diagnostic descriptions being rated as more immoral than the non-diagnostic descriptions (d = -0.16).

Next we conducted the same linear-mixed-effects model, this time to test if condition influenced MM-1 responses. Our outcome measure was MM-1, our predictor variable was condition, and scenario was also included in the model; we allowed intercepts to vary across participants, and across scenarios.

Overall, the model significantly predicted participants responses, and provided a better fit for the data than the baseline model, $\chi^2(8) = 474.81$, $p < .001$. Condition significantly influenced responses to the MM-1, $F(1, 2{,}396.05) = 45.91$, $p < .001$; and was a significant predictor in the model when controlling for scenario, b = -1.22, $t(2{,}396.05)$ = -6.78, $p < .001$, with the diagnostic descriptions being rated as more immoral than the non-diagnostic descriptions (d = -0.16).

Study 1: Differences between the Descriptions

We additionally conducted separate analyses for each scenario individually (for each dependent measure MPS-4, MM-1 and the combined measure). The responses for each scenario across each measure depending on condition are displayed in Figure S4.

For *Sam*, MPS-4 scores were significantly higher for the non-diagnostic condition (M = 2.70, SD = 0.82), than in the diagnostic condition (M = 2.42, SD = 0.87), $t(798.90)$ = -4.66, $p < .001$, d = 0.33; MM-1 ratings were higher in the non-diagnostic condition (M = 26.55, SD = 16.41), than in the diagnostic condition (M = 21.50, SD = 15.59), $t(787.84)$ = -4.45, $p < .001$, d = 0.32. For the combined measure ratings were also higher in the non-diagnostic condition (M = 0.18, SD = 0.88), than in the diagnostic condition (M =

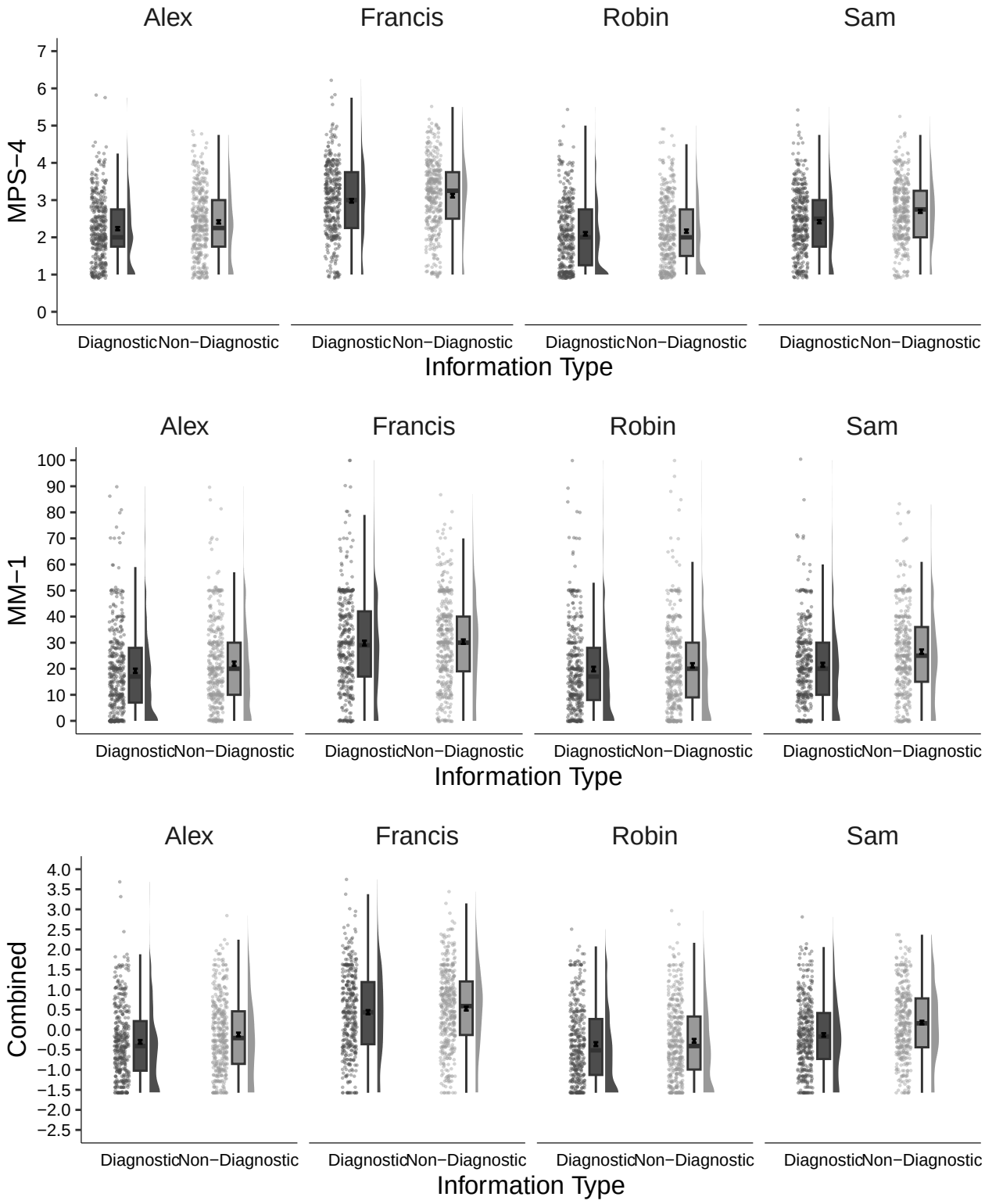

Figure S4. Study 1: Differences in moral perception for each description

-0.13, $SD = 0.88$), $t(795.41) = -4.98$, $p < .001$, $d = 0.35$.

For *Robin*, MPS-4 scores were not significantly different for the non-diagnostic condition ($M = 2.16$, $SD = 0.90$), than in the diagnostic condition ($M = 2.09$, $SD = 0.92$), $t(793.94) = -1.09$, $p = .275$, $d = 0.08$; MM-1 ratings were similar in the non-diagnostic condition ($M = 21.29$, $SD = 16.94$), and in the diagnostic condition ($M = 19.87$, $SD = 17.17$), $t(794.97) = -1.18$, $p = .239$, $d = 0.08$. For the combined measure ratings were also similar in the non-diagnostic condition ($M = -0.28$, $SD = 0.94$), and in the diagnostic condition ($M = -0.36$, $SD = 0.94$), $t(796.03) = -1.24$, $p = .217$, $d = 0.09$.

For *Alex*, MPS-4 scores were significantly higher for the non-diagnostic condition ($M = 2.41$, $SD = 0.88$), than in the diagnostic condition ($M = 2.23$, $SD = 0.86$), $t(796.97) = -2.92$, $p = .004$, $d = 0.21$; MM-1 ratings were higher in the non-diagnostic condition ($M = 21.93$, $SD = 16.47$), than in the diagnostic condition ($M = 19.20$, $SD = 16.73$), $t(798.89) = -2.33$, $p = .020$, $d = 0.16$. For the combined measure ratings were also higher in the non-diagnostic condition ($M = -0.12$, $SD = 0.92$), than in the diagnostic condition ($M = -0.30$, $SD = 0.92$), $t(798.40) = -2.82$, $p = .005$, $d = 0.20$.

For *Francis*, MPS-4 scores were significantly higher for the non-diagnostic condition ($M = 3.12$, $SD = 0.95$), than in the diagnostic condition ($M = 2.98$, $SD = 0.97$), $t(796.12) = -1.99$, $p = .047$, $d = 0.14$; MM-1 ratings were not significantly different in the non-diagnostic condition ($M = 30.38$, $SD = 17.17$), than in the diagnostic condition ($M = 29.84$, $SD = 18.56$), $t(788.61) = -0.43$, $p = .668$, $d = 0.03$. For the combined measure ratings were also similar in the non-diagnostic condition ($M = 0.53$, $SD = 0.98$), and in the diagnostic condition ($M = 0.44$, $SD = 1.02$), $t(794.36) = -1.29$, $p = .198$, $d = 0.09$.

Study 2 (good): Supplementary Analyses

Study 2: T-tests

A paired samples t-test indicated that there were no significant differences in MPS-4 responses depending on information type $t(1639)$, = -0.14, p = .893 d = 0.00, 95% CI [-0.04, 0.03]. MPS-4 responses were similar in the non-diagnostic condition (M = 6.05, SD = 0.99) compared to the diagnostic condition (M = 6.05, SD = 1.03).

A paired samples t-test indicated that there were no significant differences in MM1 responses depending on information type $t(1639)$, = 0.72, p = .475 d = 0.01, 95% CI [-0.36, 0.78]. MM-1 responses were similar in the non-diagnostic condition (M = 84.05, SD = 14.68) compared to the diagnostic condition (M = 84.26, SD = 14.76).

Study 2: Descriptives for Each Scenario

The means and standard deviations for MPS-4 for each scenario are as follows: *Sam*, $M_{\text{MPS-4}}$ = 6.12, $SD_{\text{MPS-4}}$ = 0.97, *Francis*, $M_{\text{MPS-4}}$ = 5.86, $SD_{\text{MPS-4}}$ = 1.07, *Alex*, $M_{\text{MPS-4}}$ = 6.13, $SD_{\text{MPS-4}}$ = 0.99, *Robin*, $M_{\text{MPS-4}}$ = 6.10, $SD_{\text{MPS-4}}$ = 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 characters (all ps $< .001$).

The means and standard deviations for MM-1 for each scenario are as follows: *Sam* (diagnostic/moral), $M_{\text{MM-1}}$ = 84.60, $SD_{\text{MM-1}}$ = 14.47; *Francis* (diagnostic/moral), $M_{\text{MM-1}}$ = 82.05, $SD_{\text{MM-1}}$ = 15.24; *Alex* (diagnostic/moral), $M_{\text{MM-1}}$ = 85.02, $SD_{\text{MM-1}}$ = 15.01; *Robin* (diagnostic/moral), $M_{\text{MM-1}}$ = 84.95, $SD_{\text{MM-1}}$ = 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 characters (all ps $< .001$).

Study 2: Combined Measure

Below we report the results for the combined measure of moral perception. We additionally report the effect of condition on responses to each description individually

The means and standard deviations for the combined measure for each scenario are as follows: *Sam*, $M = 0.07$, $SD = 0.97$, *Francis*, $M = -0.17$, $SD = 1.06$, *Alex*, $M = 0.09$, $SD = 1.02$, *Robin*, $M = 0.07$, $SD = 0.96$. There was significant variation depending on the description, $F(3,2335) = 48.01$, $p < .001$, partial $\eta^2 = 0.01$. *Francis* appeared to be rated as the less favorable than all other characters (all ps $< .001$), there were no differences between *Sam*, *Robin*, and *Alex* (all ps $> .05$).

We conducted a linear-mixed-effects model to test if condition influenced moral perception. Our outcome measure was the combined moral perception measure, 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, $\chi^2(8) = 142.42$, $p < .001$. Condition did not influence moral perception, $F(1, 2{,}452.92) = 0.88$, $p = .349$; and was not a significant predictor in the model when controlling for scenario, $b = -0.01$, $t(2{,}613.53) = -0.42$, $p = .673$, see Figure S5.

Study 2: Alternative Model

As in Study 1, the analyses reported in the main text included random effects for participant, as well as participant × condition random effects (allowing the effect of condition to vary across participants). When we attempted to additionally include random effects for scenario, the model failed to converge. Here we report alternative analyses where we replace the participant × condition random effects with random effects for scenario.

First we conducted a linear-mixed-effects model to test if condition influenced MPS-4 responses. Our outcome measure was MPS-4, our predictor variable was condition, and

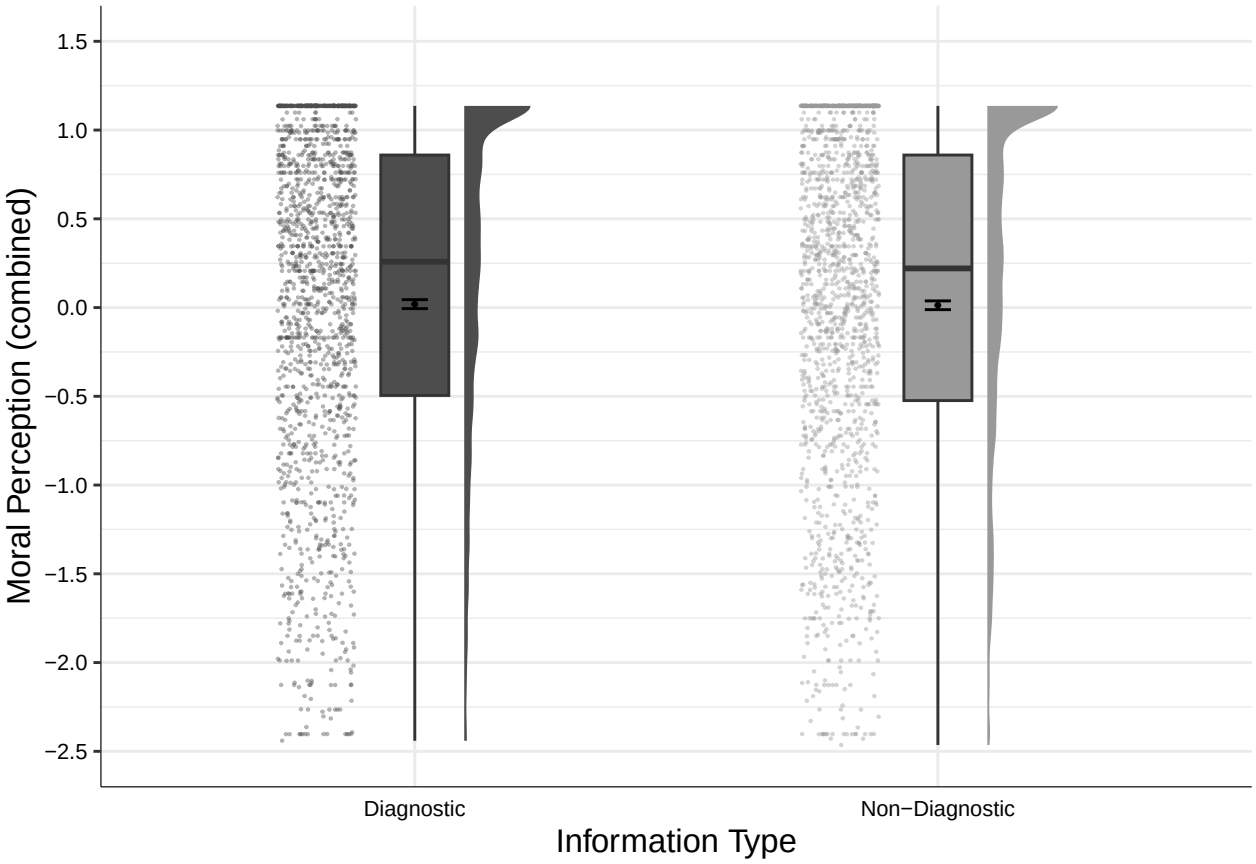

Figure S5. Study 2: Differences in combined measure depending on condition

scenario was also included in the model; we allowed intercepts to vary across participants, and across scenarios.

Overall, the model significantly predicted participants responses, and provided a better fit for the data than the baseline model, $\chi^2(9) = 160.00$, $p < .001$. Condition did not significantly influence responses to the MPS-4, $F(1, 838.26) = 0.24$, $p = .624$; and was not a significant predictor in the model when controlling for scenario, $b = 0.00$, $t(838.26) = 0.49$, $p = .624$ ($d = 0$).

Next we conducted the same linear-mixed-effects model, but for MM-1 responses, this time to test if condition influenced MM-1 responses. Our outcome measure was MM-1, our predictor variable was condition, and scenario was also included in the model; we allowed intercepts to vary across participants, and across scenarios.

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 significantly influence responses to the MM-1, $F(1, 2{,}453.06) = 1.23$, $p = .267$; and was not a significant predictor in the model when controlling for scenario, $b = 0.16$, $t(2{,}453.06) = 1.11$, $p = .267$, ($d = 0.02$).

Study 2: Differences between the Descriptions

Below we provide analyses of the effect of condition on responses to each scenario individually. The responses for each scenario across each measure depending on condition are displayed in Figure S6.

For *Sam*, MPS-4 scores were not significantly different in the non-diagnostic condition (M = 6.17, SD = 0.89), than in the diagnostic condition (M = 6.05, SD = 1.06), t(680.49) = -1.71, p = .088, d = 0.12; MM-1 ratings were similar in the non-diagnostic condition (M = 84.90, SD = 14.26), and in the diagnostic condition (M = 84.20, SD = 14.76), t(744.17) = -0.69, p = .490, d = 0.05. For the combined measure ratings were also similar in the non-diagnostic condition (M = 0.11, SD = 0.93), and in the diagnostic condition (M = 0.02, SD = 1.03), t(717.94) = -1.33, p = .183, d = 0.10.

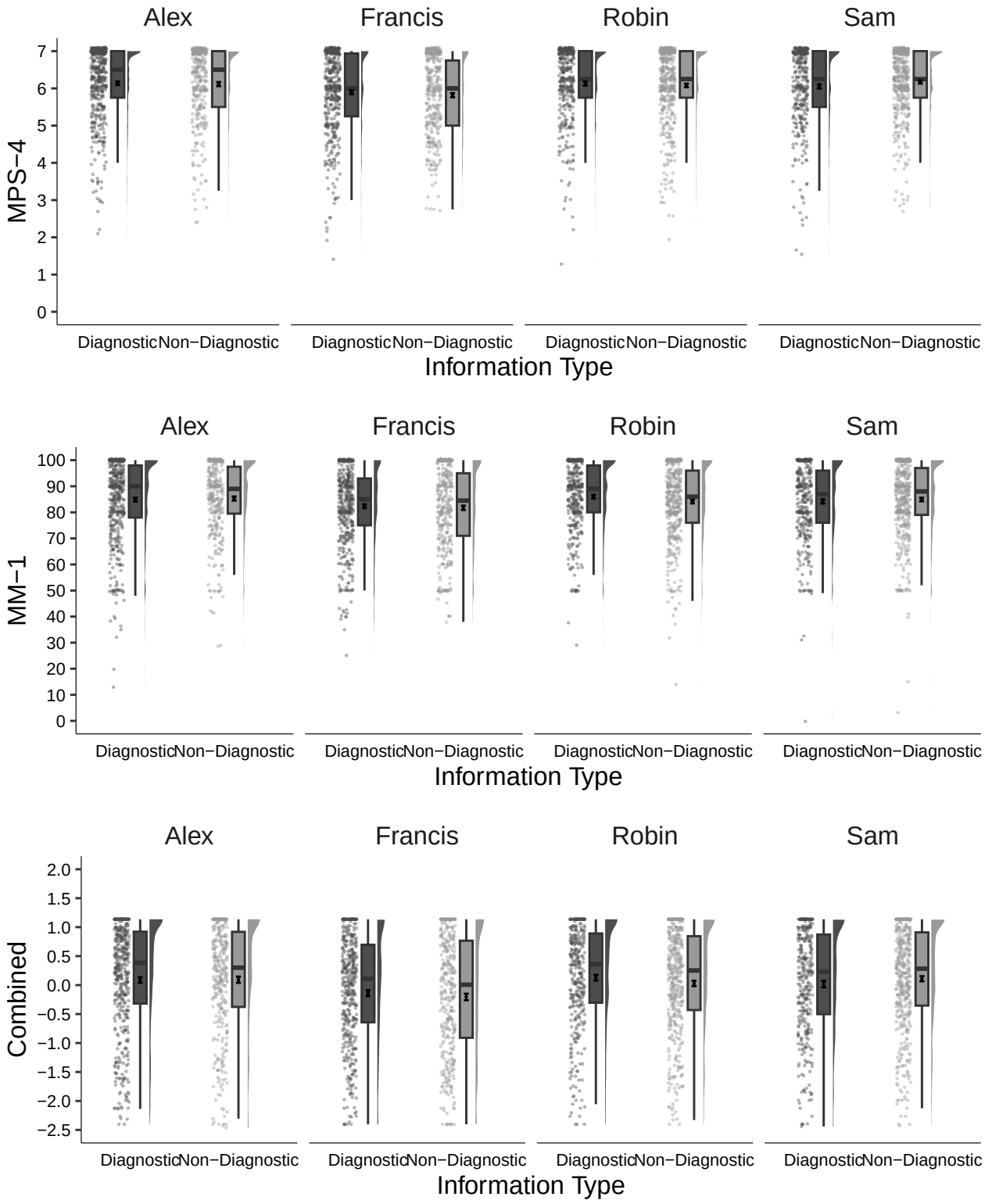

Figure S6. Study 2: Differences in moral perception for each description

For *Robin*, MPS-4 scores were not significantly different for the non-diagnostic

condition (M = 6.08, SD = 1.00), than in the diagnostic condition (M = 6.13, SD = 0.98), t(784.04) = 0.73, p = .463, d = 0.05; MM-1 ratings were similar in the non-diagnostic condition (M = 84.12, SD = 14.37), and in the diagnostic condition (M = 85.98, SD = 13.32), t(800.09) = 1.92, p = .055, d = 0.13. For the combined measure ratings were also similar in the non-diagnostic condition (M = 0.03, SD = 0.98), and in the diagnostic condition (M = 0.13, SD = 0.95), t(788.76) = 1.46, p = .145, d = 0.10.

For *Alex*, MPS-4 scores were not significantly different for the non-diagnostic condition (M = 6.11, SD = 1.00), than in the diagnostic condition (M = 6.14, SD = 0.99), t(737.60) = 0.32, p = .746, d = 0.02; MM-1 ratings were similar in the non-diagnostic condition (M = 85.28, SD = 14.31), and in the diagnostic condition (M = 84.83, SD = 15.51), t(776.47) = -0.43, p = .668, d = 0.03. For the combined measure ratings were also similar in the non-diagnostic condition (M = 0.09, SD = 0.98), and in the diagnostic condition (M = 0.09, SD = 1.04), t(767.89) = -0.06, p = .952, d = 0.00.

For *Francis*, MPS-4 scores were not significantly different for the non-diagnostic condition (M = 5.82, SD = 1.05), than in the diagnostic condition (M = 5.90, SD = 1.08), t(794.94) = 1.06, p = .290, d = 0.07; MM-1 ratings were not significantly different in the non-diagnostic condition (M = 81.74, SD = 15.67), than in the diagnostic condition (M = 82.31, SD = 14.90), t(771.23) = 0.54, p = .591, d = 0.04. For the combined measure ratings were also similar in the non-diagnostic condition (M = -0.20, SD = 1.08), and in the diagnostic condition (M = -0.14, SD = 1.04), t(777.51) = 0.88, p = .379, d = 0.06.

Study 3 (bad and good): Supplementary Analyses

Study 3: Descriptives for Each Scenario

The means and standard deviations for MPS-4 for each scenario are as follows: *Sam* (good), $M_{\text{MPS-4}} = 6.20$, $SD_{\text{MPS-4}} = 0.81$, *Francis* (bad), $M_{\text{MPS-4}} = 2.15$, $SD_{\text{MPS-4}} = 0.79$, *Alex* (bad), $M_{\text{MPS-4}} = 2.32$, $SD_{\text{MPS-4}} = 0.89$, *Robin* (good), $M_{\text{MPS-4}} = 6.32$, $SD_{\text{MPS-4}} = 0.76$. 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 $< .001$). There were no differences between *Robin* and *Sam* (*good*: $p = .003$) or between *Alex* and *Francis* (*bad*; $p < .001$).

The means and standard deviations for MM-1 for each scenario are as follows: *Sam* (good), $M_{\text{MM-1}} = 86.33$, $SD_{\text{MM-1}} = 13.68$; *Francis* (bad), $M_{\text{MM-1}} = 20.25$, $SD_{\text{MM-1}} = 16.93$; *Alex* (bad), $M_{\text{MM-1}} = 23.07$, $SD_{\text{MM-1}} = 17.51$; *Robin* (good), $M_{\text{MM-1}} = 88.44$, $SD_{\text{MM-1}} = 12.00$. 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 $< .001$). There were no differences between *Robin* and *Sam* (*good*: $p = .001$) or between *Alex* and *Francis* (*bad*; $p < .001$).

Study 3: Combined Measure

Below we report the results for the combined measure of moral perception from both DVs. We additionally report the effect of condition on responses to each description individually

The means and standard deviations for the combined measure for each scenario are as follows: *Sam*, $M = 0.93$, $SD = 0.39$, *Francis*, $M = -1.17$, $SD = 0.42$, *Alex*, $M = -1.08$,

SD = 0.46, *Robin*, *M* = 0.99, *SD* = 0.36. There was significant variation depending on the description, $F(2,1403) = 6{,}772.79$, $p < .001$, partial $\eta^2 = 0.87$. Both the *good* characters (*Robin* and *Sam*) were rated significantly more favorably than both the *bad* characters (*Alex* and *Francis*; all *p*s < .001). For the *good* characters, *Robin* was rated higher than *Sam* ($p < .001$), and for the *bad* characters *Francis* was rated more negatively than *Alex* ($p < .001$).

We conducted a linear-mixed-effects model to test if our predictors influenced responses on the combined moral perception measure. Our outcome measure was the combined moral perception measure, 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, $\chi^2(5) = 1{,}796.22$, $p < .001$. Condition significantly influenced responses to the combined moral perception measure, $F(1, 828) = 47.25$, $p < .001$ and was a significant predictor in the model when controlling for scenario, b = -0.07, $t(827.54)$ = -6.87, $p < .001$; valence significantly predicted responses, $F(1, 826) = 1{,}476.93$, $p < .001$; and there was also a significant condition × valence interaction, $F(1, 821) = 4.23$, $p = .040$, see Figure S7.

For the *bad* characters, we conducted a linear-mixed-effects model to test if condition influenced responses to the combined measure. Our outcome measure was the combined moral perception measure, 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, $\chi^2(3) = 74.54$, $p < .001$. Condition significantly influenced MPS-4 responses $F(1, 820.39) = 37.63$, $p < .001$, and was a significant predictor in the model b = -0.04, $t(820.39)$ = -6.13, $p < .001$.

For the *good* characters, we conducted a linear-mixed-effects model to test if

condition influenced responses to the combined measure. Our outcome measure was the combined moral perception measure, 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, $\chi^2(3) = 45.20$, $p < .001$. Condition significantly influenced MPS-4 responses $F(1, 826.21) = 15.67$, $p < .001$, and was a significant predictor in the model $b = 0.02$, $t(826.21) = 3.96$, $p < .001$.

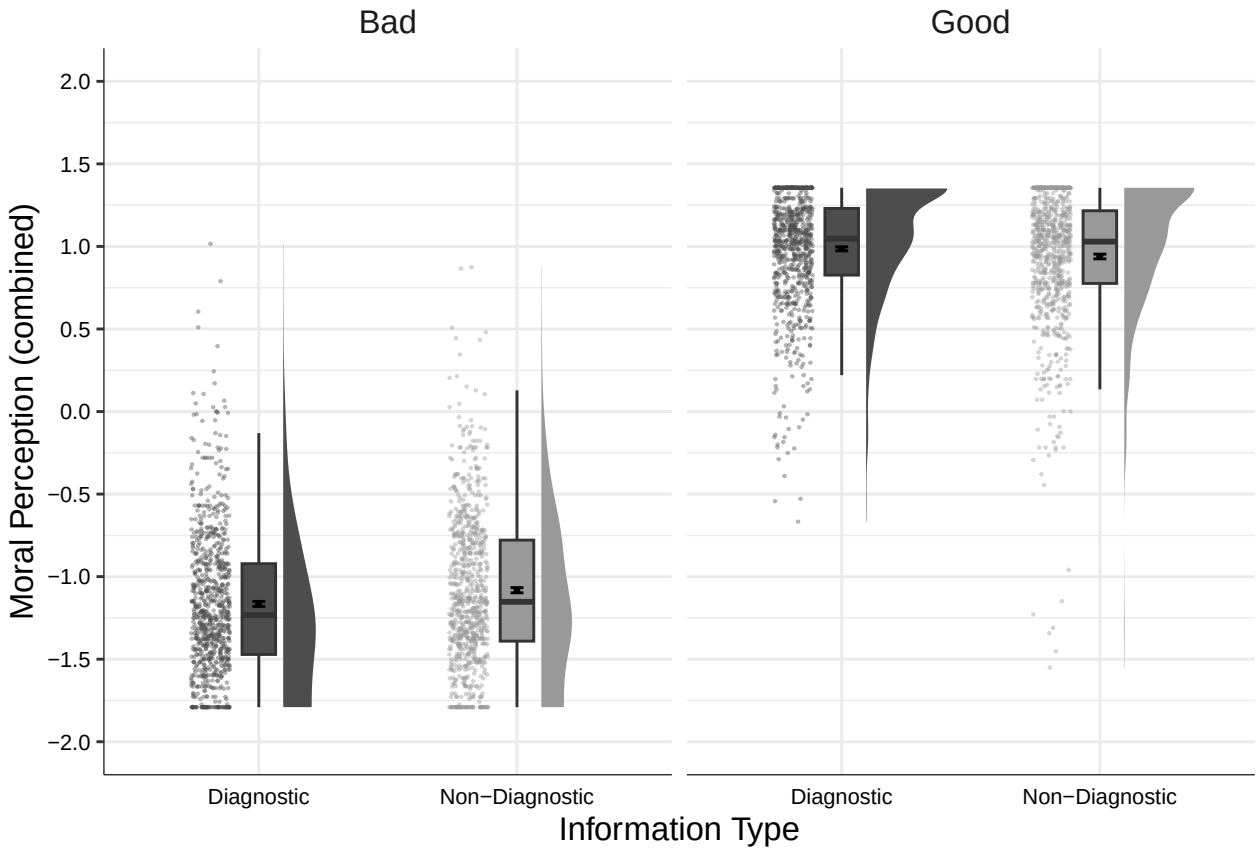

Figure S7. Study 3: Differences in the combined measure depending on condition

Study 3: Alternative Model

As with Studies 1 and 2 we report here a follow-up analysis with scenario additionally included as a random effect. We conducted a linear-mixed-effects model to test if our predictors influenced MPS-4R responses. Our outcome measure was MPS-4R, our predictor variables were condition and valence; we allowed intercepts and the effects of condition and valence to vary across participants, and we additionally included random effects for scenario. Overall, the model significantly predicted participants responses, and provided a better fit for the data than the baseline model, $\chi^2(5) = 2{,}774.91$, $p < .001$. Overall, there was a significant main effect for condition, $F(1, 1643) = 52.43$, $p < .001$; valence significantly predicted responses, $F(1, 2) = 201.92$, $p = .004$; and there was a significant condition $\times$ valence interaction, $F(1, 1641) = 6.82$, $p = .009$.

We conducted a linear-mixed-effects model to test if our predictors influenced MM-1R responses. The model was the same as the previous model, with a change to the outcome measure, our outcome measure for this model was MM-1R. As above, our predictor variables were condition and valence; we allowed intercepts and the effects of condition and valence to vary across participants, and we included random effects for scenario. Overall, the model significantly predicted participants responses, and provided a better fit for the data than the baseline model, $\chi^2(6) = 818.89$, $p < .001$. Overall there was a main effect for condition, $F(1, 827) = 32.03$, $p < .001$; valence significantly predicted responses, $F(1, 2) = 25.60$, $p = .028$; and there was no significant condition $\times$ valence interaction, $F(1, 824) = 1.82$, $p = .177$.

Interestingly, there was a consistent effect for both condition and valence, as well as a condition $\times$ valence interaction effect. To further examine this interaction effect, we report separate analyses for the good and bad descriptions below.

Study 3: Differences between the descriptions

Again, we conducted separate analyses to investigate of condition on responses to each scenario individually. The responses for each scenario across each measure depending on condition are displayed in Figure S8.

For *Sam*, MPS-4 scores were not significantly lower in the non-diagnostic condition ($M = 6.15$, $SD = 0.86$), than in the diagnostic condition ($M = 6.25$, $SD = 0.76$), $t(812.83) = 1.68$, $p = .094$, $d = 0.12$; Similarly, MM-1 ratings were similar in the non-diagnostic condition ($M = 85.49$, $SD = 14.10$), in the diagnostic condition ($M = 87.18$, $SD = 13.21$), $t(821.76) = 1.78$, $p = .075$, $d = 0.12$. For the combined measure ratings was also no significant difference between the non-diagnostic condition ($M = 0.90$, $SD = 0.42$), and the diagnostic condition ($M = 0.96$, $SD = 0.37$), $t(811.12) = 1.88$, $p = .060$, $d = 0.13$.

For *Robin*, MPS-4 scores were not significantly different for the non-diagnostic condition ($M = 6.28$, $SD = 0.80$), than in the diagnostic condition ($M = 6.36$, $SD = 0.71$), $t(809.44) = 1.60$, $p = .111$, $d = 0.11$; MM-1 ratings were similar in the non-diagnostic condition ($M = 87.84$, $SD = 13.49$), and in the diagnostic condition ($M = 89.02$, $SD = 10.30$), $t(765.30) = 1.42$, $p = .156$, $d = 0.10$. For the combined measure ratings were also similar in the non-diagnostic condition ($M = 0.97$, $SD = 0.39$), than in the diagnostic condition ($M = 1.01$, $SD = 0.32$), $t(784.03) = 1.63$, $p = .103$, $d = 0.11$.

For *Alex*, MPS-4 scores were significantly higher for the non-diagnostic condition ($M = 2.41$, $SD = 0.88$), than in the diagnostic condition ($M = 2.24$, $SD = 0.90$), $t(830.38) = -2.69$, $p = .007$, $d = 0.19$; MM-1 ratings were similar in the non-diagnostic condition ($M = 23.53$, $SD = 16.61$), and in the diagnostic condition ($M = 22.62$, $SD = 18.34$), $t(828.19) = -0.75$, $p = .454$, $d = 0.05$. For the combined measure ratings were also similar in the non-diagnostic condition ($M = -1.05$, $SD = 0.45$), and in the diagnostic condition ($M = -1.11$, $SD = 0.47$), $t(830.90) = -1.77$, $p = .077$, $d = 0.12$.

For *Francis*, MPS-4 scores were significantly higher for the non-diagnostic condition

(M = 2.26, SD = 0.85), than in the diagnostic condition (M = 2.05, SD = 0.70), t(802.80) = -3.96, $p < .001$, d = 0.27; MM-1 ratings were significantly higher in the non-diagnostic condition (M = 22.01, SD = 17.84), than in the diagnostic condition (M = 18.45, SD = 15.76), t(817.94) = -3.05, p = .002, d = 0.21. For the combined measure ratings were also significantly higher in the non-diagnostic condition (M = -1.11, SD = 0.46), than in the diagnostic condition (M = -1.23, SD = 0.38), t(808.55) = -3.85, $p < .001$, d = 0.27.

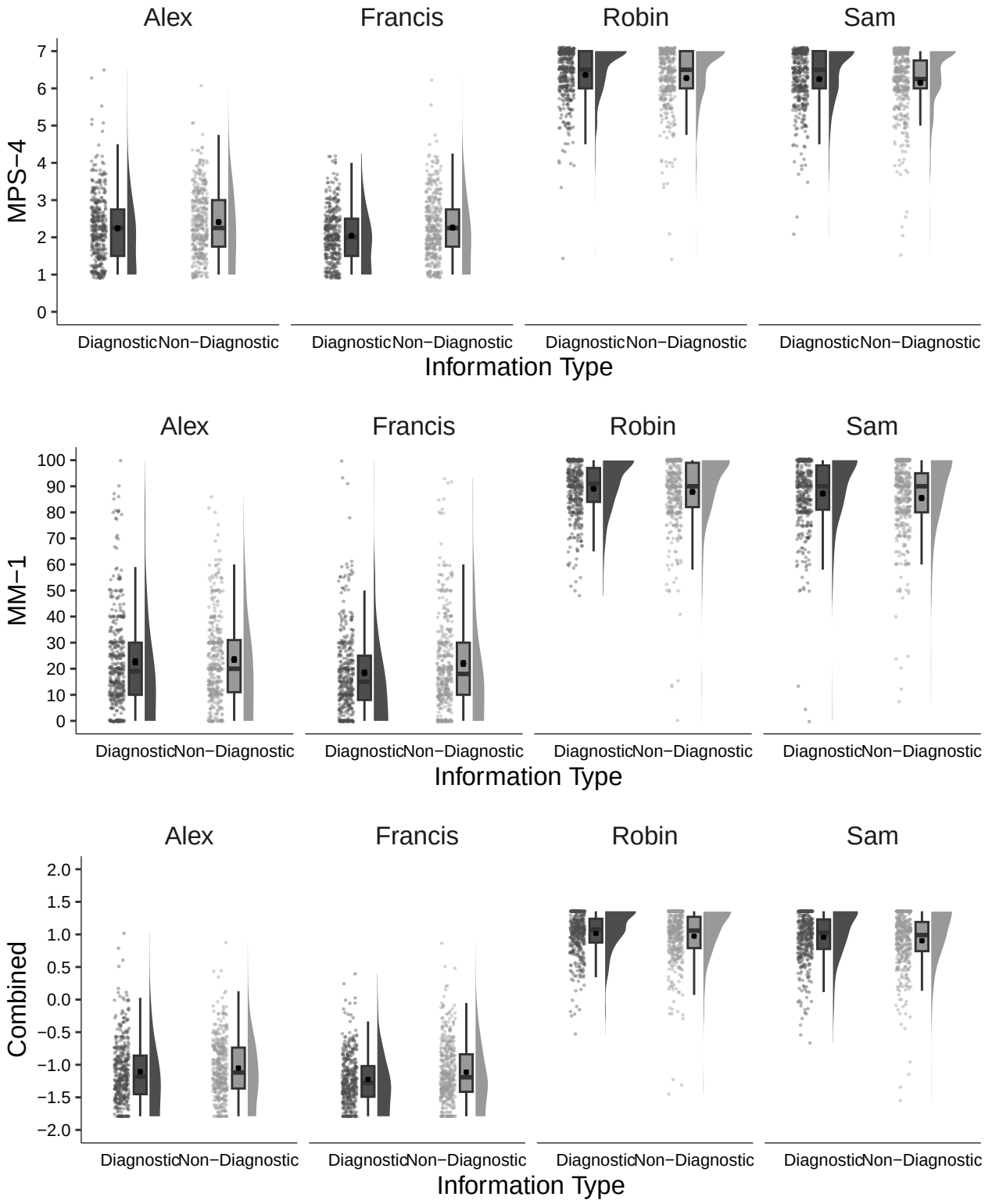

Figure S8. Study 3: Differences in moral perception for each description

Pilot Study 1

The aim of this pilot study was to develop and test materials that could be used to study the dilution effect for moral characters. We developed diagnostic and non-diagnostic character descriptions. We hypothesized that moral evaluations of the diagnostic descriptions would be more severe (more immoral) than for the non-diagnostic descriptions.

Pilot Study 1: Method

Pilot 1: Participants and design. The pilot study was a within-subjects design. The independent variable was description type with two levels, *diagnostic* and *non-diagnostic*. We used two dependent variables. The first dependent variable was the four item moral perception scale (MPS-4), participants rated the characters on four dimensions using 7-point bipolar scales. The dimensions and scale endpoints were: Bad-Good, Immoral-Moral, Violent-Peaceful, Merciless-Empathetic, this showed excellent reliability, $\alpha = 0.93$. The second dependent variable was a single item moral perception measure (MM-1) which consisted of a 100-point slider ranging from 0 = *Very Immoral* to 100 = *Very Moral*. Both dependent variables were taken from Walker et al. (2021).

A total sample of 235 (89 female, 142 male, 1 non-binary, 1 prefer not to say; $M_{\text{age}} = 36.45$, min = 20, max = 72, $SD = 10.23$) started the survey. Participants were recruited from MTurk.

We removed participants who failed both manipulation checks ($n = 23$), leaving a total sample of 212 participants (80 female, 128 male, 1 non-binary, 1 prefer not to say; $M_{\text{age}} = 36.63$, min = 20, max = 72, $SD = 10.34$).

Pilot 1: Procedure and materials. Data were collected using an online questionnaire presented with Qualtrics (www.qualtrics.com). Participants were presented with descriptions of six characters.

Moral character descriptions were developed by combining descriptions relating to three different moral foundations. These descriptions were adapted from the items of the extended character morality questionnaire (Grizzard et al., 2020), and read as follows:

(i) *Imagine a person named Sam. Throughout their life they have been known to be cruel, act unfairly, and to betray their own group*;
(ii) *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*;
(iii) *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*
(iv) *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.*

We developed neutral descriptions that included information relating to physical appearance/attributes, hobbies/activities, and family information that read as follows:

(i) *Imagine a person named Jackie. They have red hair, play tennis four times a month, and have one older sibling and one younger sibling*;
(ii) *Imagine a person named Charlie. They are left-handed, drink tea in the morning, and have two older siblings and one younger sibling.*

Character descriptions did not specify the gender of the charcters, and all characters had names that could be either male or female (Sam, Robin, Francis, Alex, Jackie, Charlie). All participants read six descriptions, four moral descriptions and two neutral. Pilot Study 1 was pre-registered at https://aspredicted.org/3VK_8FD.

Pilot 1: Results

Pilot 1: Main Measures. The means and standard deviations for MPS-4 for each scenario are as follows: *Sam* (diagnostic), $M_{\text{MPS-4}} = 4.35$, $SD_{\text{MPS-4}} = 1.90$, *Francis* (diagnostic), $M_{\text{MPS-4}} = 4.46$, $SD_{\text{MPS-4}} = 1.73$, *Alex* (diagnostic), $M_{\text{MPS-4}} = 4.44$, $SD_{\text{MPS-4}} = 1.79$, *Robin* (diagnostic), $M_{\text{MPS-4}} = 4.35$, $SD_{\text{MPS-4}} = 1.96$, *Jackie* (non-diagnostic), $M_{\text{MPS-4}} = 5.40$, $SD_{\text{MPS-4}} = 1.01$, *Charlie* (non-diagnostic), $M_{\text{MPS-4}} = 5.38$, $SD_{\text{MPS-4}} = 1.01$. For the diagnostic descriptions, there was no significant variation depending on the description, $F(3,600) = 1.58$, $p = .194$, partial $\eta^2 = 0.00$. For the non-diagnostic descriptions there was no significant difference in ratings depending on description, $t(211) = -0.67$, $p = .506$, $d = 0.05$.

The means and standard deviations for MM-1 for each scenario are as follows: *Sam* (diagnostic), $M_{\text{MM-1}} = 55.67$, $SD_{\text{MM-1}} = 30.47$; *Francis* (diagnostic), $M_{\text{MM-1}} = 58.22$, $SD_{\text{MM-1}} = 28.61$; *Alex* (diagnostic), $M_{\text{MM-1}} = 56.80$, $SD_{\text{MM-1}} = 29.45$; *Robin* (diagnostic), $M_{\text{MM-1}} = 55.49$, $SD_{\text{MM-1}} = 31.38$; *Jackie* (non-diagnostic), $M_{\text{MM-1}} = 73.00$, $SD_{\text{MM-1}} = 14.72$; *Charlie* (non-diagnostic), $M_{\text{MM-1}} = 72.94$, $SD_{\text{MM-1}} = 14.79$. For the diagnostic descriptions, we observed significant variation depending on the description, $F(3,608) = 3.01$, $p = .032$, partial $\eta^2 = 0.001$. When correcting for multiple comparisons, pairwise comparisons did not reveal significant differences between descriptions. We note that without correction, *Francis* appeared to be rated as more moral than both *Robin* ($p = .012$), and *Sam* ($p = .009$). For the non-diagnostic descriptions there was no significant difference in ratings depending on description, $t(211) = -0.09$, $p = .929$, $d = 0.01$.

We conducted a linear-mixed-effects model to test if condition influenced MPS-4 responses. Our outcome measure was MPS-4, our predictor variable was condition; we allowed intercepts and the effect of condition to vary across participants. Overall, the model significantly predicted participants responses, and provided a better fit for the data than the baseline model, $\chi^2(2) = 860.16$, $p < .001$. Condition was a significant predictor in

the model b = -0.49, $t(211.05)$ = -8.54, $p < .001$, with the non-diagnostic descriptions being rated as more moral than the diagnostic descriptions of immoral characters Figure S9.

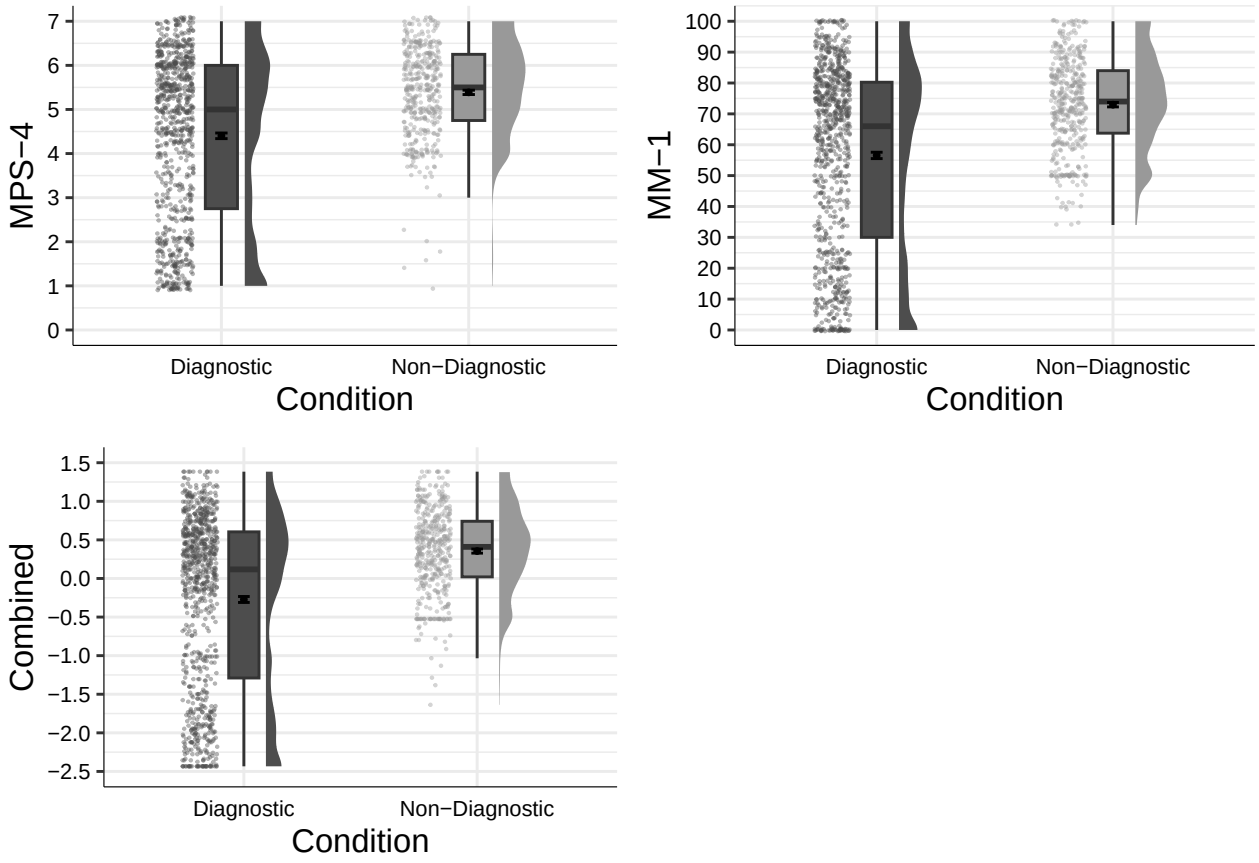

Figure S9. Pilot Study 1: Differences in moral perception depending on condition

We conducted a linear-mixed-effects model to test if condition influenced MM-1 responses. Our outcome measure was MM-1, our predictor variable was condition; we allowed intercepts and the effect of condition to vary across participants. Overall, the model significantly predicted participants responses, and provided a better fit for the data than the baseline model, $\chi^2(2) = 924.82$, $p < .001$. Condition was a significant predictor in the model b = -8.22, $t(210.98)$ = -8.60, $p < .001$, with the non-diagnostic descriptions being rated as more moral than the diagnostic descriptions, see Figure S9.

Pilot 1: Combined Measure. We developed a combined moral perception measure by calculating the mean of the combined mean-centered scores for MPS-4 and

MM-1, and mean-centering this result. Below we report the analyses for this combined measure.

The standardized means and standard deviations for the combined measure for each scenario are as follows: *Sam* (diagnostic), M = -0.30, SD = 1.16; *Francis* (diagnostic), M = -0.22, SD = 1.06; *Alex* (diagnostic), M = -0.25, SD = 1.10; *Robin* (diagnostic), M = -0.31, SD = 1.19; *Jackie* (non-diagnostic), M = 0.36, SD = 0.55; *Charlie* (non-diagnostic), M = 0.35, SD = 0.55. For the moral descriptions, we observed significant variation depending on the description, $F(3,602) = 2.67$, $p = .050$, partial $\eta^2 = 0.001$. When correcting for multiple comparisons, pairwise comparisons did not reveal significant differences between descriptions. We note that without correction, *Francis* appeared to be rated as more moral than both *Robin* ($p = .022$), and *Sam* ($p = .021$). For the neutral descriptions there was no significant difference in ratings depending on description, $t(211)$ = -0.46, $p = .645$, $d = 0.03$.

We conducted a linear-mixed-effects model to test if condition influenced responses on this combined measure. Overall, the model significantly predicted participants responses, and provided a better fit for the data than the baseline model $\chi^2(2) = 1{,}035.36$, $p < .001$, and condition was a significant predictor in the model b = -0.31, $t(210.99)$ = -8.74, $p < .001$. Participants rated the neutral/non-diagnostic descriptions as more moral than the immoral/diagnostic descriptions (see Figure S9).

Pilot Study 2

Pilot Study 1 developed materials for studying the dilution effect with morally *bad* characters. In Pilot Study 2, we develop materials for studying the dilution effect with morally *good* characters. As with Pilot Study 1, we developed diagnostic and non-diagnostic descriptions. We hypothesized that evaluations of the diagnostic descriptions would be more extreme (more moral) than for the non-diagnostic descriptions

Pilot Study 2: Method

Pilot 2: Participants and design. The pilot study was a within-subjects design. The independent variable was description type with two levels, *diagnostic* and *non-diagnostic*. We used the same two dependent variables as in previous studies, the four item moral perception scale (MPS-4, $\alpha = 0.84$), and the single item moral perception measure (MM-1).

A total sample of 245 (70 female, 175 male, 0 non-binary, 0 prefer not to say; $M_{\text{age}} = 36.69$, min = 18, max = 71, $SD = 9.57$) started the survey. Participants were recruited from MTurk.

We removed participants who failed both manipulation checks ($n = 30$), leaving a total sample of 215 participants (63 female, 152 male, 0 non-binary, 0 prefer not to say; $M_{\text{age}} = 36.59$, min = 18, max = 71, $SD = 9.59$).

Pilot 2: Procedure and materials. Data were collected using an online questionnaire presented with Qualtrics (www.qualtrics.com). Participants were presented with descriptions of six characters.

Moral character descriptions were developed by combining descriptions relating to three different moral foundations, focusing on upholding the moral foundations (rather than transgressions as in previous studies). We developed 4 descriptions of moral characters that read as follows:

(i) *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*;

(ii) *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*;

(iii) *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*;

(iv) *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.*

We developed 2 descriptions of morally neutral characters that included information relating to physical appearance/attributes, hobbies/activities, and a color preference:

(i) *Imagine a person named Jackie. They have dark hair, go for a jog twice a week, and their favourite colour is blue*;

(ii) *Imagine a person named Charlie. They have blue eyes, drink coffee in the morning, and their favourite colour is green.*

We used the same gender ambiguous names, and we did not specify the gender of the characters. Pilot Study 2 was pre-registered at https://aspredicted.org/W52_VPX.

Pilot 2: Results

Pilot 2: Main Measures. The means and standard deviations for MPS-4 for each scenario are as follows: *Sam* (diagnostic), $M_{\text{MPS-4}} = 6.01$, $SD_{\text{MPS-4}} = 0.91$, *Francis* (diagnostic), $M_{\text{MPS-4}} = 5.89$, $SD_{\text{MPS-4}} = 0.95$, *Alex* (diagnostic), $M_{\text{MPS-4}} = 5.94$, $SD_{\text{MPS-4}} =$

0.94, *Robin* (diagnostic), $M_{\text{MPS-4}} = 5.93$, $SD_{\text{MPS-4}} = 0.92$, *Jackie* (non-diagnostic), $M_{\text{MPS-4}} = 5.60$, $SD_{\text{MPS-4}} = 0.99$, *Charlie* (non-diagnostic), $M_{\text{MPS-4}} = 5.53$, $SD_{\text{MPS-4}} = 1.08$. For the diagnostic descriptions, there was significant variation depending on the description, $F(3,613) = 2.91$, $p = .036$, partial $\eta^2 = 0.00$, *Sam* was viewed significantly more favorably than *Francis* ($p = .040$). For the non-diagnostic descriptions there was no significant difference in ratings depending on description, $t(214) = -1.79$, $p = .075$, $d = 0.12$.

The means and standard deviations for MM-1 for each scenario are as follows: *Sam* (diagnostic), $M_{\text{MM-1}} = 79.85$, $SD_{\text{MM-1}} = 15.44$; *Francis* (diagnostic), $M_{\text{MM-1}} = 78.30$, $SD_{\text{MM-1}} = 15.84$; *Alex* (diagnostic), $M_{\text{MM-1}} = 79.78$, $SD_{\text{MM-1}} = 15.71$; *Robin* (diagnostic), $M_{\text{MM-1}} = 79.46$, $SD_{\text{MM-1}} = 15.41$; *Jackie* (non-diagnostic), $M_{\text{MM-1}} = 73.44$, $SD_{\text{MM-1}} = 15.83$; *Charlie* (non-diagnostic), $M_{\text{MM-1}} = 73.07$, $SD_{\text{MM-1}} = 16.22$. For the diagnostic descriptions, we observed no significant variation depending on the description, $F(3,594) = 1.45$, $p = .231$, partial $\eta^2 = 0.002$. For the non-diagnostic descriptions there was no significant difference in ratings depending on description, $t(214) = -0.60$, $p = .552$, $d = 0.04$.

We conducted a linear-mixed-effects model to test if condition influenced MPS-4 responses. Our outcome measure was MPS-4, our predictor variable was condition; we allowed intercepts and the effect of condition to vary across participants. Overall, the model significantly predicted participants responses, and provided a better fit for the data than the baseline model, $\chi^2(2) = 475.42$, $p < .001$. Condition was a significant predictor in the model $b = 0.19$, $t(214.35) = 6.53$, $p < .001$, with the diagnostic descriptions being rated as more moral than the non-diagnostic descriptions of immoral characters Figure S10.

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

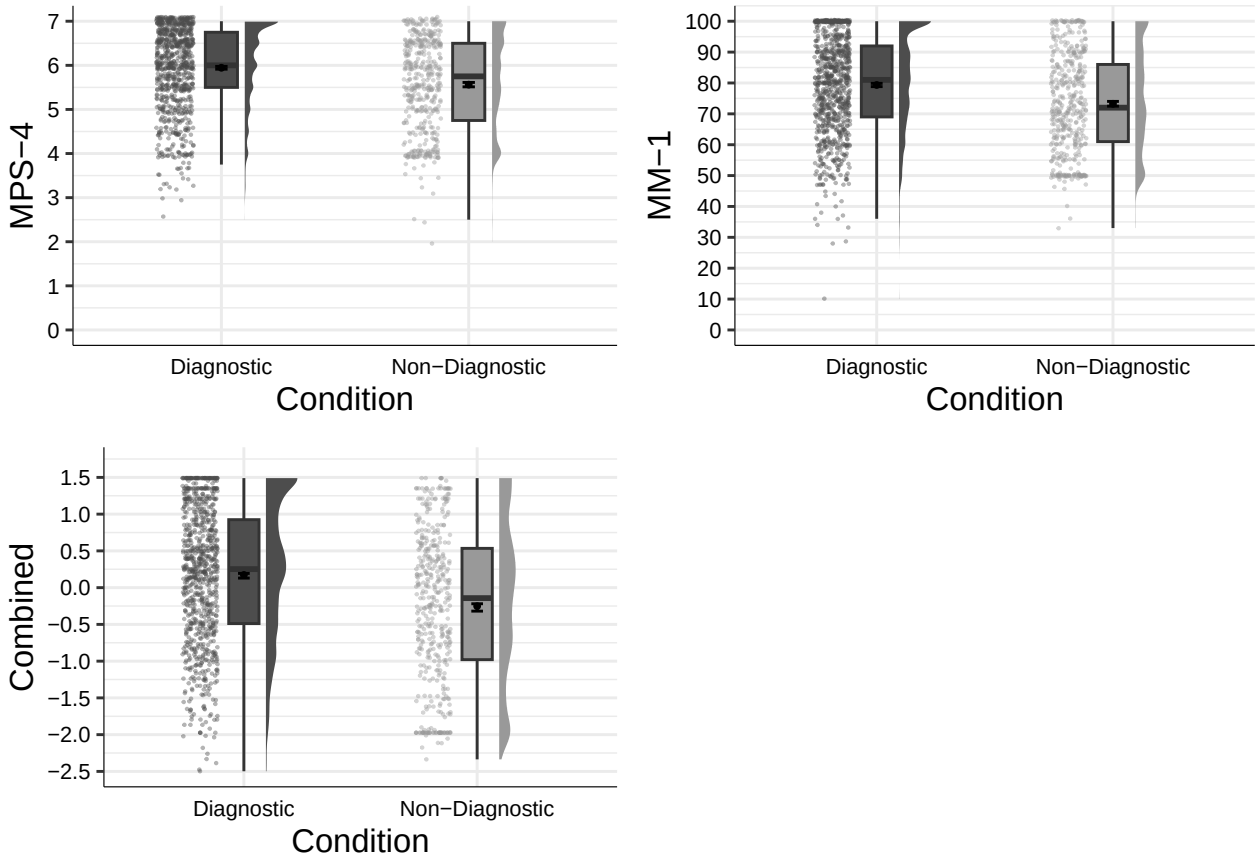

Figure S10. Pilot Study 2: Differences in moral perception depending on condition

than the baseline model, $\chi^2(2) = 324.13$, $p < .001$. Condition was a significant predictor in the model $b = 3.04$, $t(214.90) = 6.02$, $p < .001$, with the diagnostic descriptions being rated as more moral than the non-diagnostic descriptions, see Figure S10.

Pilot 2: Combined Measure. As in previous studies, we developed a combined moral perception measure by calculating the mean of the combined mean-centered scores for MPS-4 and MM-1, and mean-centering this result. Below we report the analyses for this combined measure.

The standardized means and standard deviations for the combined measure for each scenario are as follows: *Sam* (moral), $M = 0.21$, $SD = 0.91$; *Francis* (moral), $M = 0.10$, $SD = 0.96$; *Alex* (moral), $M = 0.18$, $SD = 0.94$; *Robin* (moral), $M = 0.16$, $SD = 0.93$; *Jackie* (neutral), $M = -0.24$, $SD = 1.01$; *Charlie* (neutral), $M = -0.30$, $SD = 1.07$. For the moral

descriptions, we observed significant variation depending on the description, $F(3,588) = 2.90$, $p = .039$, partial $\eta^2 = 0.002$. *Sam* was viewed significantly more favorably than *Francis* ($p = .045$). For the neutral descriptions there was no significant difference in ratings depending on description, $t(426.74) = -0.51$, $p = .609$, $d = 0.10$.

We conducted a linear-mixed-effects model to test if condition influenced responses to the combined measure. Overall, the model significantly predicted participants responses, and provided a better fit for the data than the baseline model $\chi^2(2) = 564.98$, $p < .001$, and condition was a significant predictor in the model $b = 0.22$, $t(214.32) = 6.60$, $p < .001$ (see Figure S10).

Study S1 - Good Characters

Study S1 is a replication of Study 2, but with an MTurk Sample.

Study S1: Method

Study S1: Participants and design. The design, materials, and procedure for Study S1 were the same as for Study 2, the only change from Study 2 was that all participants in Study S2 were recruited from MTurk. Study S1 was a within-subjects design. The independent variable was condition with two levels, diagnostic and non-diagnostic. We used the same two dependent variables as in previous studies, the four item moral perception scale (MPS-4, $\alpha = 0.81$), and the single item moral perception measure MM-1.

A total sample of 1118 (445 female, 642 male, 2 non-binary, 3 other; 1 prefer not to say, $M_{\text{age}} = 37.44$, min = 19, max = 84, $SD = 11.08$) started the survey. Participants were recruited from MTurk and paid $0.40 for their participation.

Participants who failed both manipulation checks were removed ($n = 262$), leaving a total sample of 856 participants (347 female, 507 male, 0 other, 0 prefer not to say; $M_{\text{age}} = 37.12$, min = 19, max = 84, $SD = 11.04$).

Study S1: Procedure and materials. All materials and procedures were the same as in Study 2.

Study S1: Results

Study S1: Main Measures. The means and standard deviations for MPS-4 for each scenario are as follows: *Sam*, $M_{\text{MPS-4}} = 5.95$, $SD_{\text{MPS-4}} = 0.93$, *Francis*, $M_{\text{MPS-4}} = 5.89$, $SD_{\text{MPS-4}} = 0.91$, *Alex*, $M_{\text{MPS-4}} = 5.94$, $SD_{\text{MPS-4}} = 0.96$, *Robin*, $M_{\text{MPS-4}} = 5.95$, $SD_{\text{MPS-4}} = 0.94$. There was significant variation depending on the description, $F(3,2527) = 3.30$, $p =$

.020, partial $\eta^2 = 0.001$. Pairwise comparisons did not reveal any significant differences between individual descriptions (all ps $> .05$).

The means and standard deviations for MM-1 for each scenario are as follows: *Sam* (diagnostic/moral), $M_{\text{MM-1}} = 81.34$, $SD_{\text{MM-1}} = 14.14$; *Francis* (diagnostic/moral), $M_{\text{MM-1}} = 80.65$, $SD_{\text{MM-1}} = 14.16$; *Alex* (diagnostic/moral), $M_{\text{MM-1}} = 81.15$, $SD_{\text{MM-1}} = 14.42$; *Robin* (diagnostic/moral), $M_{\text{MM-1}} = 81.63$, $SD_{\text{MM-1}} = 14.15$. There was significant variation depending on the description, $F(3,2518) = 2.89$, $p = .035$, partial $\eta^2 = 0.001$. Pairwise comparisons did not reveal any significant differences between individual descriptions (all ps $> .05$).

We conducted a linear-mixed-effects model to test if condition influenced MPS-4 responses. Our outcome measure was MPS-4, our predictor variable was condition; we allowed intercepts and the effect of condition to vary across participants, and scenario was also included in the model. Overall, the model significantly predicted participants responses, and provided a better fit for the data than the baseline model, $\chi^2(8) = 17.86$, $p = .022$. Condition did not influence responses to the MPS-4, $F(1, 866.60) = 2.80$, $p = .095$; and was not a significant predictor in the model when controlling for scenario, $b = 0.01$, $t(867) = 1.67$, $p = .095$, ($d = 0.04$), see Figure S11.

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, $\chi^2(8) = 40.10$, $p < .001$. Condition significantly influenced MM-1 responses $F(1, 864) = 4.79$, $p = .029$, and was a significant predictor in the model $b = 0.29$, $t(864) = 2.19$, $p = .029$, ($d = 0.05$), see Figure S11.

Study S1: Combined Measure. The means and standard deviations for the combined measure for each scenario are as follows: *Sam*, $M = 0.03$, $SD = 1.02$, *Francis*, M

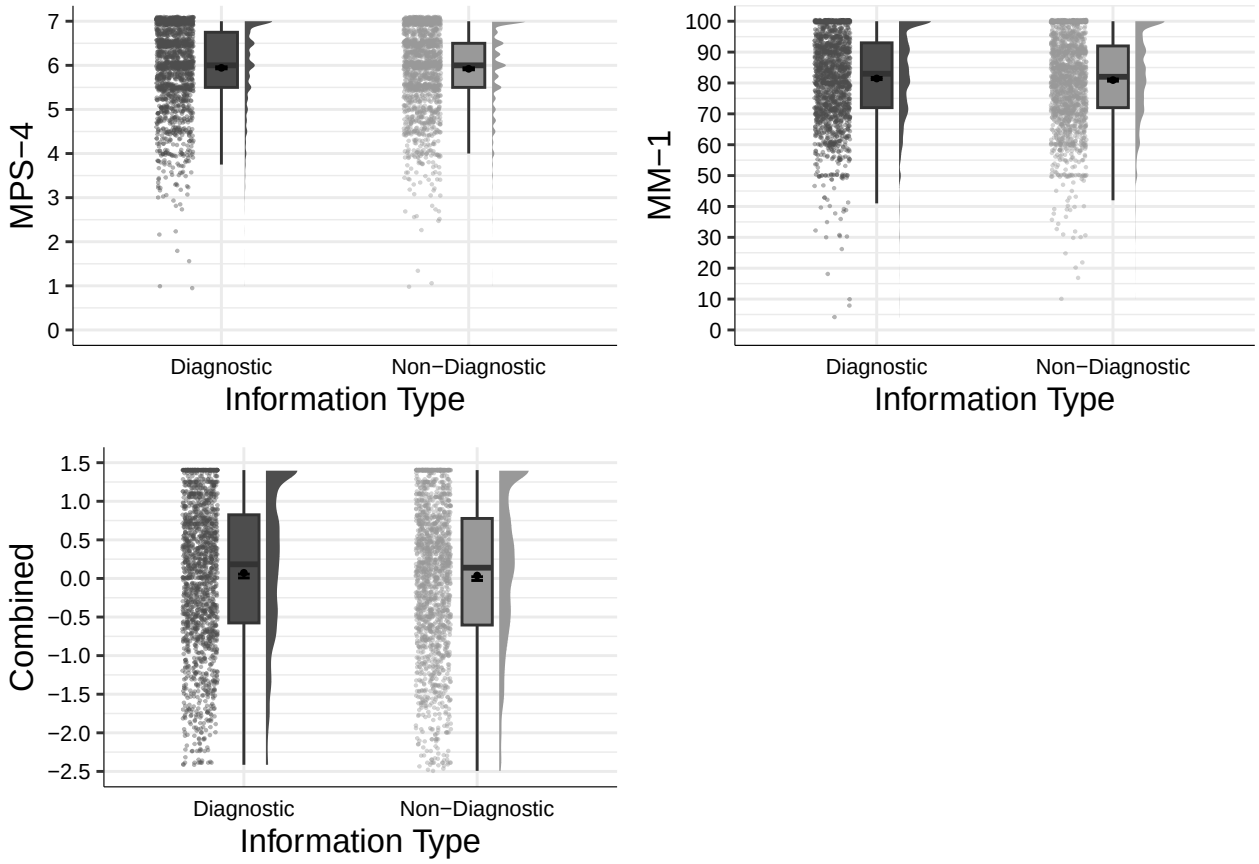

Figure S11. Study S1: Responses to moral perception measures depending on condition

= -0.03, *SD* = 0.98, *Alex*, *M* = 0.02, *SD* = 1.04, *Robin*, *M* = 0.04, *SD* = 1.01. There was significant variation depending on the description, $F(3,2493) = 4.32$, $p = .005$, partial $\eta^2 = 0.00$. Follow-up pairwise comparisons did not reveal any significant differences between the different characters (all *p*s > .05).

We conducted a linear-mixed-effects model to test if condition influenced moral perception. Our outcome measure was the combined moral perception measure, 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, $\chi^2(8) = 42.42$, $p < .001$. Condition did not influence moral perception, $F(1, 865.01) = 5.31$, $p = .021$; and was not a significant predictor in the model when

controlling for scenario, b = -0.01, t(2,541.03) = -0.82, p = .410, see Figure S5.

Study S1: Differences between the Descriptions. Below we provide analyses of the effect of condition on responses to each scenario individually. The responses for each scenario across each measure depending on condition are displayed in Figure S12.

For *Sam*, MPS-4 scores were not significantly different in the non-diagnostic condition (M = 5.89, SD = 0.91), than in the diagnostic condition (M = 6.02, SD = 0.95), t(810.53) = 1.97, p = .049, d = 0.14; MM-1 ratings were similar in the non-diagnostic condition (M = 79.75, SD = 14.62), than in the diagnostic condition (M = 83.25, SD = 13.30), t(845.88) = 3.66, $p < .001$, d = 0.25. For the combined measure ratings were also similar in the non-diagnostic condition (M = -0.06, SD = 1.03), than in the diagnostic condition (M = 0.15, SD = 1.01), t(829.20) = 3.07, p = .002, d = 0.21.

For *Robin*, MPS-4 scores were not significantly different for the non-diagnostic condition (M = 5.95, SD = 0.93), than in the diagnostic condition (M = 5.94, SD = 0.95), t(811.83) = -0.20, p = .841, d = 0.01; MM-1 ratings were similar in the non-diagnostic condition (M = 81.62, SD = 14.28), and in the diagnostic condition (M = 81.64, SD = 14.02), t(824.54) = 0.02, p = .982, d = 0.00. For the combined measure ratings were also similar in the non-diagnostic condition (M = 0.04, SD = 1.03), than in the diagnostic condition (M = 0.04, SD = 0.99), t(828.47) = -0.10, p = .919, d = 0.01.

For *Alex*, MPS-4 scores were not significantly different for the non-diagnostic condition (M = 5.97, SD = 0.91), than in the diagnostic condition (M = 5.91, SD = 0.99), t(845.29) = -0.91, p = .362, d = 0.06; MM-1 ratings were similar in the non-diagnostic condition (M = 81.93, SD = 13.38), than in the diagnostic condition (M = 80.51, SD = 15.21), t(850.53) = -1.46, p = .145, d = 0.10. For the combined measure ratings were also similar in the non-diagnostic condition (M = 0.07, SD = 0.98), than in the diagnostic condition (M = -0.02, SD = 1.09), t(847.27) = -1.30, p = .192, d = 0.09.

For *Francis*, MPS-4 scores were not significantly different for the non-diagnostic

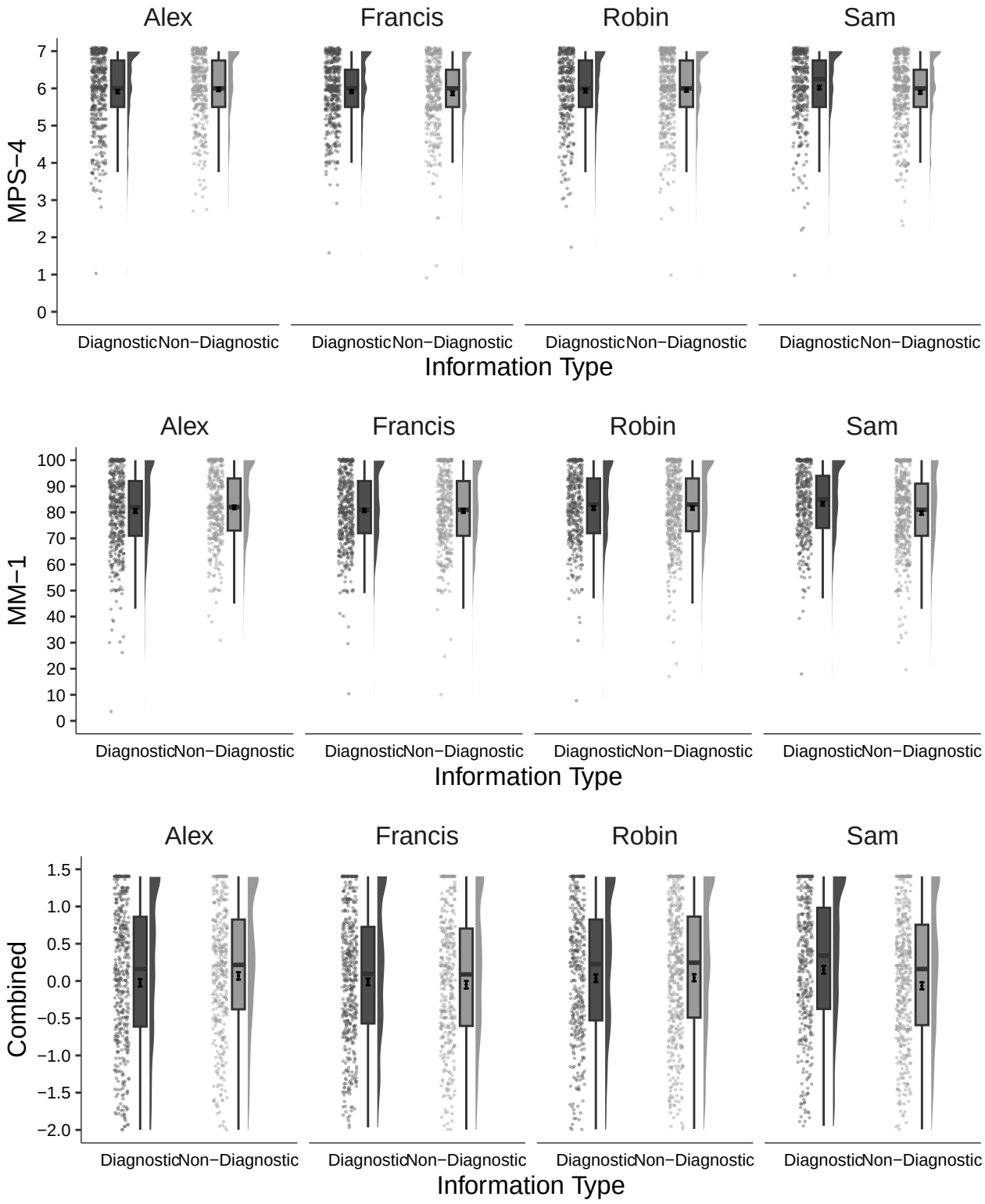

Figure S12. Study 2: Differences in moral perception for each description

condition (M = 5.87, SD = 0.95), than in the diagnostic condition (M = 5.91, SD = 0.87), $t(787.36) = 0.77$, $p = .443$, $d = 0.05$; MM-1 ratings were not significantly different in the non-diagnostic condition (M = 80.54, SD = 14.38), than in the diagnostic condition (M = 80.75, SD = 13.99), $t(809.63) = 0.21$, $p = .832$, $d = 0.01$. For the combined measure ratings were also similar in the non-diagnostic condition (M = -0.05, SD = 0.99), and in the diagnostic condition (M = -0.01, SD = 0.98), $t(814.30) = 0.55$, $p = .581$, $d = 0.04$.

Study S2 - Good and Bad Characters

Study S2 is the same as Study 3, but with an MTurk sample. Study 3 was pre-registered at https://aspredicted.org/QDF_XT1.

Study S2: Method

Study S2: Participants and design. Study S2 was a 2 × 2 within-subjects factorial design. The first independent variable was condition with two levels, diagnostic and non-diagnostic. The second independent variable was valence of character description, with two levels morally good and morally bad. We used the same two dependent variables as in previous studies, the four item moral perception scale (MPS-4, $\alpha = 0.94$), and the single item moral perception measure MM-1.

A total sample of 1095 (386 female, 700 male, 2 non-binary, 0 other; 2 prefer not to say, $M_{\text{age}} = 36.42$, min = 19, max = 77, $SD = 10.65$) started the survey. Participants were recruited from MTurk and paid $0.40 for their participation.

Participants who failed both manipulation checks were removed ($n = 221$), leaving a total sample of 874 participants (320 female, 550 male, 0 other, 0 prefer not to say; $M_{\text{age}} = 36.37$, min = 19, max = 77, $SD = 10.72$).

Study S2: Procedure and materials. Again, data were collected using an online questionnaire presented with Qualtrics (www.qualtrics.com). Participants were presented with four descriptions of characters as in Study 3. To ensure consistency across character judgments, we selected descriptions that related to the same moral foundations (care, fairness, and loyalty). We used the same four character names as in previous studies. The *good* characters were *Sam* and *Robin*, and the *bad* characters were *Francis* and *Alex*, e.g., *Imagine a person named Robin. Throughout their life they have been known to show compassion and empathy for others, act with a sense of fairness and justice, and, never to break their word.* or, *Imagine a person named Alex. Throughout their life they have been*

known to be cruel, act unfairly, and to betray their own group. Full descriptions for each character are in the supplementary materials. One description for each the *good* and *bad* characters was randomly assigned to include non-diagnostic information for each participant thus all participants were exposed to all conditions (see https://osf.io/mdnpv/?view_only=77883e3fbc3d45f1a35fe92d5318cb67for details of the randomization blocks). Study S2 was pre-registered at https://aspredicted.org/QDF_XT1

Study S2: Results

The means and standard deviations for MPS-4 for each scenario are as follows: *Sam* (good), $M_{\text{MPS-4}} = 5.90$, $SD_{\text{MPS-4}} = 1.03$, *Francis* (bad), $M_{\text{MPS-4}} = 4.07$, $SD_{\text{MPS-4}} = 2.07$, *Alex* (bad), $M_{\text{MPS-4}} = 4.03$, $SD_{\text{MPS-4}} = 2.03$, *Robin* (good), $M_{\text{MPS-4}} = 5.85$, $SD_{\text{MPS-4}} = 1.05$. There was significant variation depending on the description, $F(1,1080) = 442.71$, $p < .001$, partial $\eta^2 = 0.24$. Both the *good* characters (*Robin* and *Sam*) were rated significantly more favorably than both the *bad* characters (*Alex* and *Francis*; all *p*s $< .001$). There were no differences between *Robin* and *Sam* (*good*: $p = .366$) or between *Alex* and *Francis* (*bad*; $p = .648$).

The means and standard deviations for MM-1 for each scenario are as follows: *Sam* (good), $M_{\text{MM-1}} = 81.01$, $SD_{\text{MM-1}} = 15.23$; *Francis* (bad), $M_{\text{MM-1}} = 51.49$, $SD_{\text{MM-1}} = 33.18$; *Alex* (bad), $M_{\text{MM-1}} = 50.89$, $SD_{\text{MM-1}} = 32.14$; *Robin* (good), $M_{\text{MM-1}} = 80.81$, $SD_{\text{MM-1}} = 15.16$. There was significant variation depending on the description, $F(1,1080) = 458.92$, $p < .001$, partial $\eta^2 = 0.254$. Again, the *good* characters (*Robin* and *Sam*) were rated significantly more favorably than the *bad* characters (*Alex* and *Francis*; all *p*s $< .001$). There were no differences between *Robin* and *Sam* (*good*: $p = .776$) or between *Alex* and *Francis* (*bad*; $p = .683$).

We conducted a linear-mixed-effects model to test if our predictors influenced MPS-4 responses. Our outcome measure was MPS-4, 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, $\chi^2(5) = 4{,}554.31$, $p < .001$. Overall, there was a significant main effect for condition, $F(1, 873) = 8.61$, $p = .003$; valence significantly predicted responses, $F(1, 873) = 1{,}859.34$, $p < .001$; and there was no significant condition × valence interaction, $F(1, 873) = 0.01$, $p = .935$.

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 MM-1. 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, $\chi^2(5) = 3{,}496.86$, $p < .001$. Overall there was a main effect for condition, $F(1, 873) = 16.61$, $p < .001$; valence significantly predicted responses, $F(1, 873) = 986.37$, $p < .001$; and there was no significant condition × valence interaction, $F(1, 873) = 0.04$, $p = .849$.

We conducted a linear-mixed-effects model to test if our predictors influenced responses on the combined moral perception measure. Our outcome measure was the combined moral perception measure, 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, $\chi^2(5) = 4{,}467.15$, $p < .001$. Condition significantly influenced responses to the combined moral perception measure, $F(1, 873) = 16.65$, $p < .001$ and was a significant predictor in the model when controlling for scenario, $b = -0.02$, $t(873.00) = -4.08$, $p < .001$; valence significantly predicted responses, $F(1, 873) = 1{,}598.27$, $p < .001$; and there was also a significant condition × valence interaction, $F(1, 873) = 0.03$, $p = .867$, see Figure S7.

For both MP-4 and MM-1 (and the combined measure) we found a main effect for condition and valence, and there was no condition × valence interaction. We conducted follow-up analyses to test the if the main effect for condition holds for both good and bad descriptions separately.

Differences in the *Bad* Descriptions

For the *bad* characters, we conducted a linear-mixed-effects model to test if condition influenced MPS-4 responses. Our outcome measure was MPS-4, our predictor variable was condition; we allowed intercepts and the effect of condition to vary across participants. Overall, the model did not significantly predict participants responses, or provide a better fit for the data than the baseline model, $\chi^2(3) = 5.40$, $p = .145$. Condition did not significantly influence MPS-4 responses $F(1, 872.00) = 3.54$, $p = .060$, and was not a significant predictor in the model b = -0.03 (β = -0.02), $t(872.00)$ = -1.88, $p = .060$, (d = -0.06), see Figure S13.

We also conducted a linear-mixed-effects model to test if condition influenced MM-1 responses. Our outcome measure was MM-1, our predictor variable was condition; we allowed intercepts and the effect of condition to vary across participants. Overall, the model significantly predicted participants responses, and provided a better fit for the data than the baseline model, $\chi^2(3) = 8.67$, $p = .034$. Condition significantly influenced MM-1 responses $F(1, 872.00) = 7.01$, $p = .008$, and was a significant predictor in the model b = -0.69 (β = -0.02), $t(872.00)$ = -2.65, $p = .008$, (d = -0.09), see Figure S13.

Differences in the *Good* Descriptions

For the *good* characters, we conducted a linear-mixed-effects model to test if condition influenced MPS-4 responses. Our outcome measure was MPS-4, our predictor variable was condition; we allowed intercepts and the effect of condition to vary across participants.

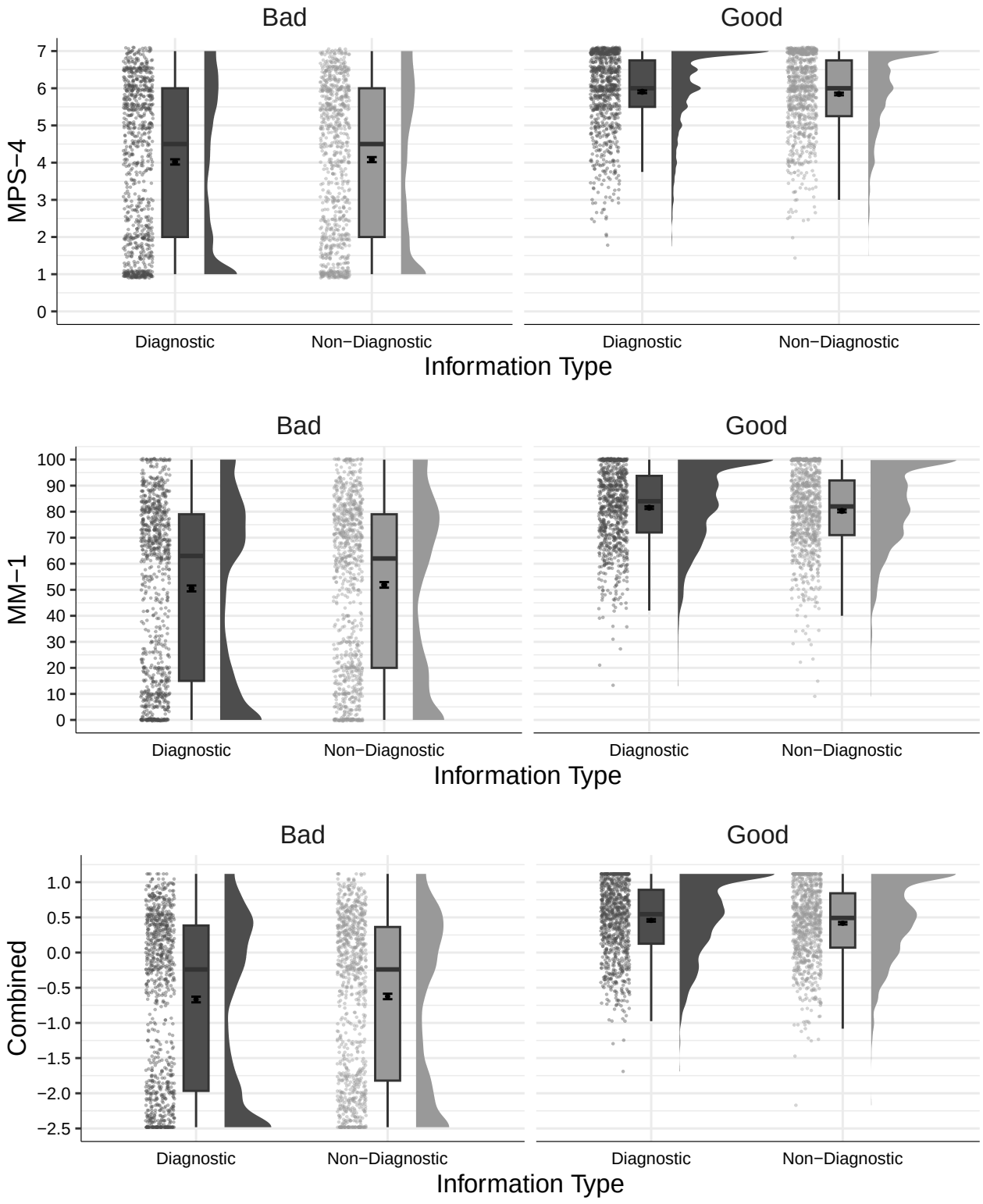

Figure S13. Study S2: Differences in moral perception depending on condition

Overall, the model significantly predicted participants responses, and provided a better fit for the data than the baseline model, $\chi^2(3) = 13.66$, $p = .003$. Condition significantly influenced MPS-4 responses $F(1, 872.00) = 6.82$, $p = .009$, and was a significant predictor in the model $b = 0.03$ ($\beta = 0.03$), $t(872.00) = 2.61$, $p = .009$, ($d = 0.09$), see Figure S13.

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, $\chi^2(1) = 11.97$, $p < .001$. Condition significantly influenced MM-1 responses $F(1, 873) = 12.04$, $p < .001$, and was a significant predictor in the model $b = 0.63$ ($\beta = 0.04$), $t(873) = 3.47$, $p < .001$, ($d = 0.12$), see Figure S13.

Study S2: Differences between the descriptions

Again, we conducted separate analyses to investigate of condition on responses to each scenario individually. The responses for each scenario across each measure depending on condition are displayed in Figure S14.

For *Sam* (*good*), MPS-4 scores were significantly lower in the non-diagnostic condition ($M = 5.81$, $SD = 1.09$), than in the diagnostic condition ($M = 5.98$, $SD = 0.97$), $t(859.15) = 2.46$, $p = .014$, $d = 0.17$; Similarly, MM-1 ratings were significantly lower in the non-diagnostic condition ($M = 79.64$, $SD = 15.68$), than in the diagnostic condition ($M = 82.37$, $SD = 14.67$), $t(867.08) = 2.66$, $p = .008$, $d = 0.18$. For the combined measure ratings were also lower in the non-diagnostic condition ($M = 0.39$, $SD = 0.54$), than in the diagnostic condition ($M = 0.50$, $SD = 0.50$), $t(863.14) = 2.85$, $p = .004$, $d = 0.19$.

For *Robin* (*good*), MPS-4 scores were not significantly different for the non-diagnostic condition ($M = 5.88$, $SD = 0.96$), than in the diagnostic condition ($M = 5.83$, $SD = 1.14$), $t(844.53) = -0.77$, $p = .440$, $d = 0.05$; MM-1 ratings were similar in the non-diagnostic

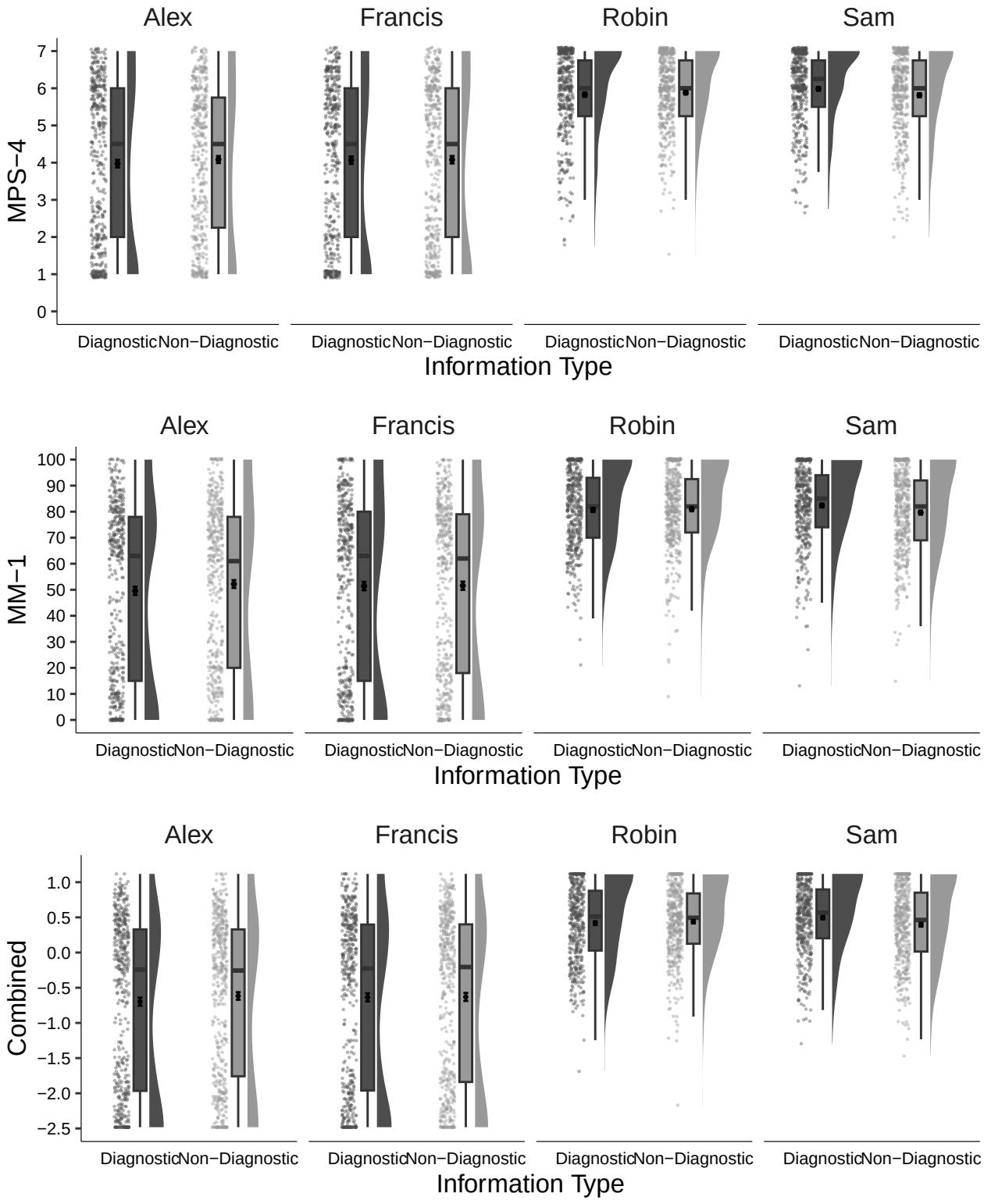

Figure S14. Study S2: Differences in moral perception for each description

condition (M = 80.92, SD = 15.27), and in the diagnostic condition (M = 80.70, SD = 15.07), t(871.98) = -0.22, p = .828, d = 0.01. For the combined measure ratings were also similar in the non-diagnostic condition (M = 0.44, SD = 0.51), than in the diagnostic condition (M = 0.42, SD = 0.54), t(867.63) = -0.57, p = .569, d = 0.04.

For *Alex* (*bad*), MPS-4 scores were not significantly different for the non-diagnostic condition (M = 4.08, SD = 1.96), than in the diagnostic condition (M = 3.97, SD = 2.11), t(865.81) = -0.80, p = .421, d = 0.05; MM-1 ratings were similar in the non-diagnostic condition (M = 52.19, SD = 31.29), and in the diagnostic condition (M = 49.58, SD = 32.95), t(868.76) = -1.20, p = .230, d = 0.08. For the combined measure ratings were also similar in the non-diagnostic condition (M = -0.62, SD = 1.11), and in the diagnostic condition (M = -0.70, SD = 1.19), t(867.67) = -1.04, p = .301, d = 0.07.

For *Francis* (*bad*), MPS-4 scores were not significantly different for the non-diagnostic condition (M = 4.08, SD = 2.07), than in the diagnostic condition (M = 4.07, SD = 2.07), t(871.94) = -0.09, p = .928, d = 0.01; MM-1 ratings were not significantly different in the non-diagnostic condition (M = 51.56, SD = 32.68), than in the diagnostic condition (M = 51.42, SD = 33.70), t(871.59) = -0.06, p = .952, d = 0.00. For the combined measure ratings were also similar in the non-diagnostic condition (M = -0.63, SD = 1.18), and in the diagnostic condition (M = -0.64, SD = 1.20), t(871.88) = -0.08, p = .939, d = 0.01.

Study S3 - Good and Bad Characters

The aim of Study S3 was to test for the moral dilution effect in both good and bad characters, while attempting to eliminate the confounding influence of the presence of other descriptions by adopting a between-subjects design.

Study S3: Method

Study S3: Participants and design. Study S3 was a 2 × 2 between-subjects factorial design. As in Study 3, 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 (MPS-4, $\alpha = 0.97$, and MM-1).

A total sample of 2389 (1137 female, 1236 male, 5 non-binary, 3 other; 8 prefer not to say, $M_{\text{age}} = 38.78$, min = 2, max = 1995, $SD = 42.71$) started the survey. Participants were recruited from MTurk and paid $0.10 for their participation.

Participants who failed both manipulation checks were removed ($n = 445$), leaving a total sample of 1944 participants (960 female, 970 male, 2 other, 2 prefer not to say; $M_{\text{age}} = 37.88$, min = 2, max = 454, $SD = 15.49$).

Study S3: Procedure and materials. The materials for Study S3 were the same as those used in Study 3. Participants were randomly presented with a single character description: *Sam*, *Robin* (*good* characters), *Francis* and *Alex* (*bad* characters), and were randomly assigned to the diagnostic condition (containing diagnostic information only), or the non-diagnostic condition (where the character description additionally included non-diagnostic information). Study S3 was not pre-registered however our predictions were the same as those for Study 3.

Study S3: Results

The means and standard deviations for MPS-4 for each scenario are as follows: *Sam* (good), $M_{\text{MPS-4}} = 6.15$, $SD_{\text{MPS-4}} = 0.87$, *Francis* (bad), $M_{\text{MPS-4}} = 3.65$, $SD_{\text{MPS-4}} = 2.16$, *Alex* (bad), $M_{\text{MPS-4}} = 3.65$, $SD_{\text{MPS-4}} = 2.09$, *Robin* (good), $M_{\text{MPS-4}} = 6.21$, $SD_{\text{MPS-4}} = 0.85$. There was significant variation depending on the description, $F(3,1940) = 396.86$, $p < .001$, partial $\eta^2 = 0.38$. Both the *good* characters (*Robin* and *Sam*) were rated significantly more favorably than both the *bad* characters (*Alex* and *Francis*; all *p*s $< .001$). There were no differences between *Robin* and *Sam* (*good*: $p = .932$) or between *Alex* and *Francis* (*bad*; $p > .999$).

The means and standard deviations for MM-1 for each scenario are as follows: *Sam* (good), $M_{\text{MM-1}} = 84.70$, $SD_{\text{MM-1}} = 15.32$; *Francis* (bad), $M_{\text{MM-1}} = 43.37$, $SD_{\text{MM-1}} = 34.96$; *Alex* (bad), $M_{\text{MM-1}} = 44.68$, $SD_{\text{MM-1}} = 34.57$; *Robin* (good), $M_{\text{MM-1}} = 85.33$, $SD_{\text{MM-1}} = 14.47$. There was significant variation depending on the description, $F(3,1940) = 383.99$, $p < .001$, partial $\eta^2 = 0.37$. Both the *good* characters (*Robin* and *Sam*) were rated significantly more favorably than both the *bad* characters (*Alex* and *Francis*; all *p*s $< .001$). There were no differences between *Robin* and *Sam* (*good*: $p = .982$) or between *Alex* and *Francis* (*bad*; ($p = .872$)).

We conducted a 2 × 2 between subjects ANOVA to test for an interaction between valence and condition in predicting MPS-4. Condition significantly influenced responses to the MPS-4, $F(1, 1940) = 5.16$, $p = .023$; valence significantly predicted responses, $F(1, 1940) = 1{,}495.09$, $p < .001$; and there was no significant condition × valence interaction, $F(1, 1940) = 0.03$, $p = .858$.

We conducted a 2 × 2 between subjects ANOVA to test for an interaction between valence and condition in predicting responses to MM-1. Condition significantly influenced responses to MM-1, $F(1, 1940) = 9.46$, $p = .002$; valence significantly predicted responses, $F(1, 1940) = 580.03$, $p < .001$; and there was no significant condition × valence

interaction, $F(1, 1940) = 0.32$, $p = .573$.

We conducted a 2 × 2 between subjects ANOVA to test for an interaction between valence and condition in predicting responses to the combined measure. Condition significantly influenced responses to the combined measure, $F(1, 1940) = 9.46$, $p = .002$; valence significantly predicted responses, $F(1, 1940) = 580.03$, $p < .001$; and there was no significant condition × valence interaction, $F(1, 1940) = 0.32$, $p = .573$.

As in previous studied we conducted separate analyses for the good and bad descriptions.

Differences in the *Bad* Descriptions

For the *bad* characters, there was no significant difference in responses to MPS-4 between the diagnostic condition ($M = 3.57$, $SD = 2.21$) and the non-diagnostic condition ($M = 3.72$, $SD = 2.03$) depending on condition, $t(941.99) = -1.09$, $p = .277$, $d = -0.07$.

For the *bad* characters, there was no significant difference in responses to MM-1 between the diagnostic condition ($M = 34.62$, $SD = 34.25$) and the non-diagnostic condition ($M = 37.04$, $SD = 32.16$) depending on condition, $t(680.45) = -0.96$, $p = .339$, $d = 0.07$.

Differences in the *Good* Descriptions

For the *good* characters, there was no significant difference in responses to MPS-4 between the diagnostic condition ($M = 6.27$, $SD = 0.84$) and the non-diagnostic condition ($M = 6.09$, $SD = 0.88$) depending on condition, $t(981.90) = 3.21$, $p = .001$, $d = 0.20$.

For the *good* characters, there was no significant difference in responses to MPS-4 between the diagnostic condition ($M = 87.23$, $SD = 13.68$) and the non-diagnostic condition ($M = 82.89$, $SD = 15.70$) depending on condition, $t(972.88) = 4.63$, $p < .001$, $d = 0.29$.

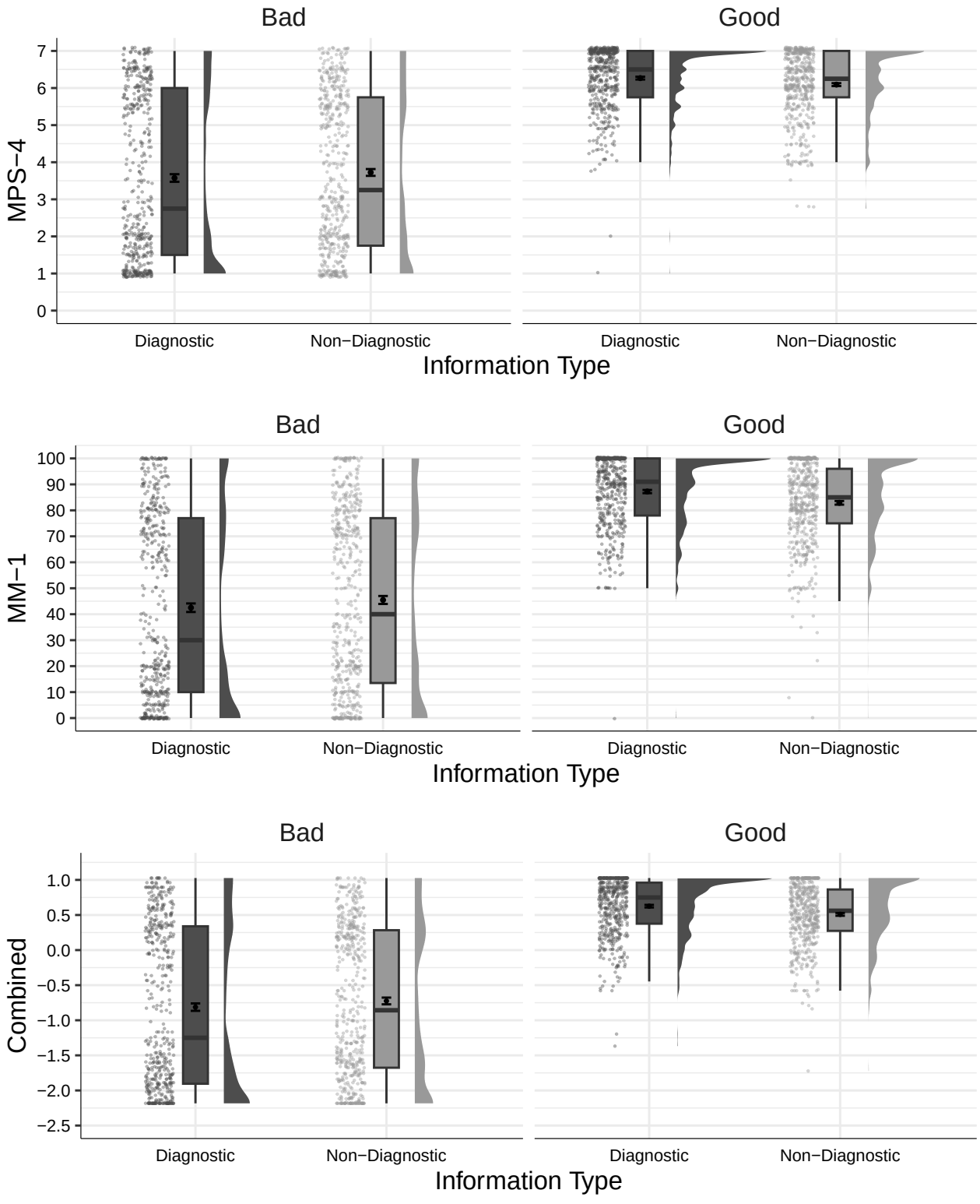

Figure S15. Study S3: Differences in moral perception depending on condition

Study S3: Differences between the descriptions

Again, we conducted separate analyses to investigate of condition on responses to each scenario individually. The responses for each scenario across each measure depending on condition are displayed in Figure S16.

For *Sam*, MPS-4 scores were significantly lower in the non-diagnostic condition (M = 6.06, SD = 0.89), than in the diagnostic condition (M = 6.24, SD = 0.85), $t(488.52)$ = 2.34, p = .020, d = 0.21; Similarly, MM-1 ratings were significantly lower in the non-diagnostic condition (M = 82.23, SD = 16.63), than in the diagnostic condition (M = 87.21, SD = 13.43), $t(471.87)$ = 3.65, $p < .001$, d = 0.33. For the combined measure ratings were also lower in the non-diagnostic condition (M = 0.49, SD = 0.45), than in the diagnostic condition (M = 0.62, SD = 0.41), $t(486.39)$ = 3.35, $p < .001$, d = 0.30.

For *Robin*, MPS-4 scores were not significantly different for the non-diagnostic condition (M = 6.13, SD = 0.87), than in the diagnostic condition (M = 6.30, SD = 0.83), $t(490.93)$ = 2.21, p = .027, d = 0.20; MM-1 ratings were similar in the non-diagnostic condition (M = 83.53, SD = 14.75), and in the diagnostic condition (M = 87.25, SD = 13.96), $t(490.98)$ = 2.88, p = .004, d = 0.26. For the combined measure ratings were also similar in the non-diagnostic condition (M = 0.53, SD = 0.42), than in the diagnostic condition (M = 0.63, SD = 0.40), $t(490.92)$ = 2.82, p = .005, d = 0.25.

For *Alex*, MPS-4 scores were not significantly different for the non-diagnostic condition (M = 3.71, SD = 2.02), than in the diagnostic condition (M = 3.59, SD = 2.16), $t(465.98)$ = -0.62, p = .534, d = 0.06; MM-1 ratings were similar in the non-diagnostic condition (M = 45.80, SD = 33.76), than in the diagnostic condition (M = 43.48, SD = 35.46), $t(468.21)$ = -0.73, p = .465, d = 0.07. For the combined measure ratings were also similar in the non-diagnostic condition (M = -0.72, SD = 1.05), than in the diagnostic condition (M = -0.79, SD = 1.13), $t(465.61)$ = -0.69, p = .489, d = 0.06.

For *Francis*, MPS-4 scores were not significantly different for the non-diagnostic

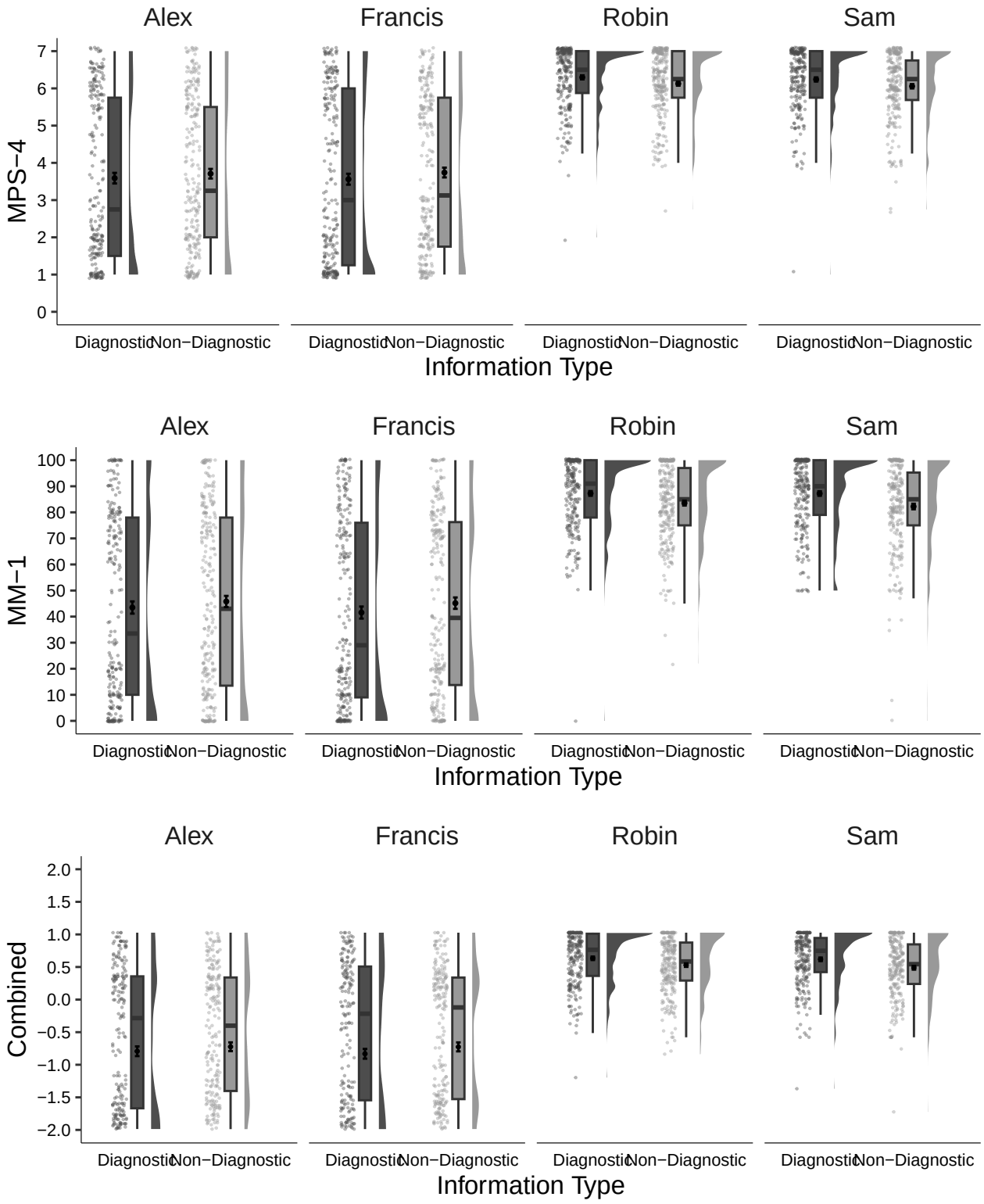

Figure S16. Study S3: Differences in moral perception for each description

condition (M = 3.74, SD = 2.05), than in the diagnostic condition (M = 3.56, SD = 2.27), t(474.15) = -0.91, p = .364, d = 0.08; MM-1 ratings were not significantly different in the non-diagnostic condition (M = 45.16, SD = 34.18), than in the diagnostic condition (M = 41.55, SD = 35.73), t(478.97) = -1.13, p = .258, d = 0.10. For the combined measure ratings were also similar in the non-diagnostic condition (M = -0.73, SD = 1.07), and in the diagnostic condition (M = -0.83, SD = 1.16), t(476.42) = -1.04, p = .297, d = 0.10.

Meta-analyses and Forest Plots

Overall, there was a significant dilution effect across Studies 1-3, for both good and bad characters, and for both measures, $d_{pooled} = 0.12$, $SE = 0.05$, $z = 2.38$, $p = .017$, 95% CI [0.02, 0.22].

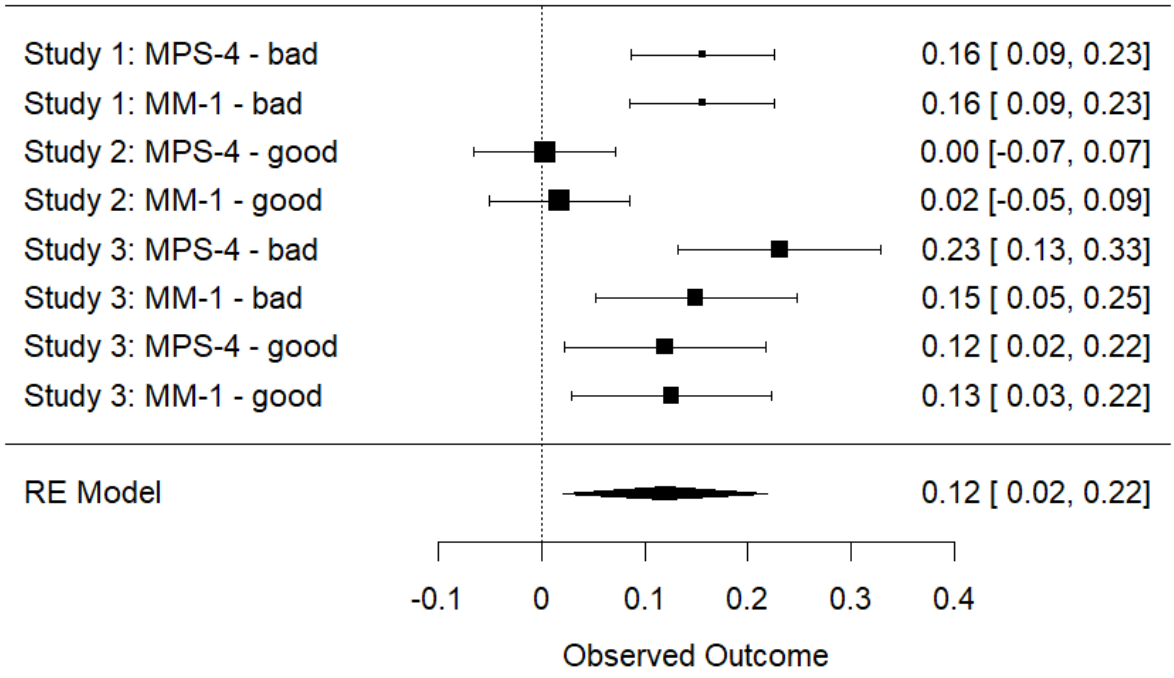

Figure S17. Forest plot showing effects for Studies 1-3 and pooled effect.

Overall, there was a significant dilution effect across Studies 1 and 3, for bad characters, and for both measures, d_{pooled} = -0.17, SE = 0.02, z = -7.97, $p < .001$, 95% CI [-0.22, -0.13].

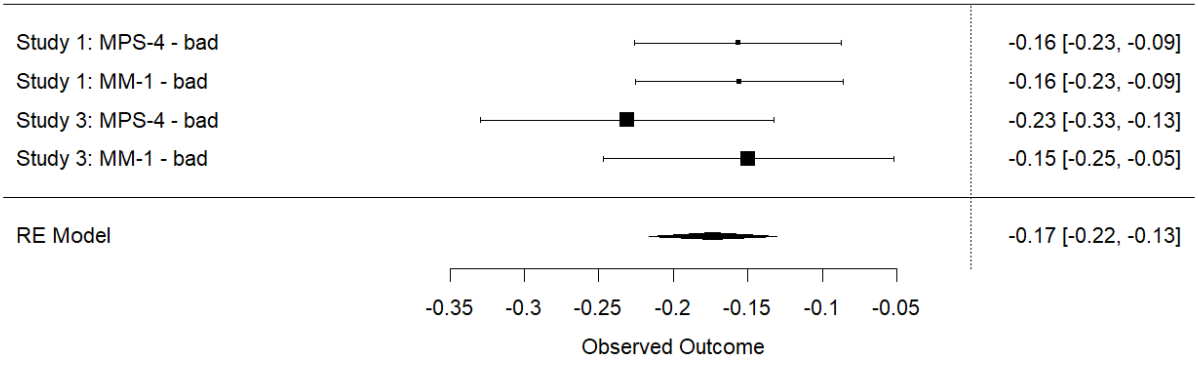

Figure S18. Forest plot showing effects for bad characters for Studies 1 and 3 and pooled effect.

Overall, there was no significant dilution effect across Studies 2 and 3, for good characters, for either measures, $d_{pooled} = 0.06$, $SE = 0.06$, $z = 1.12$, $p = .263$, 95% CI [-0.05, 0.18].

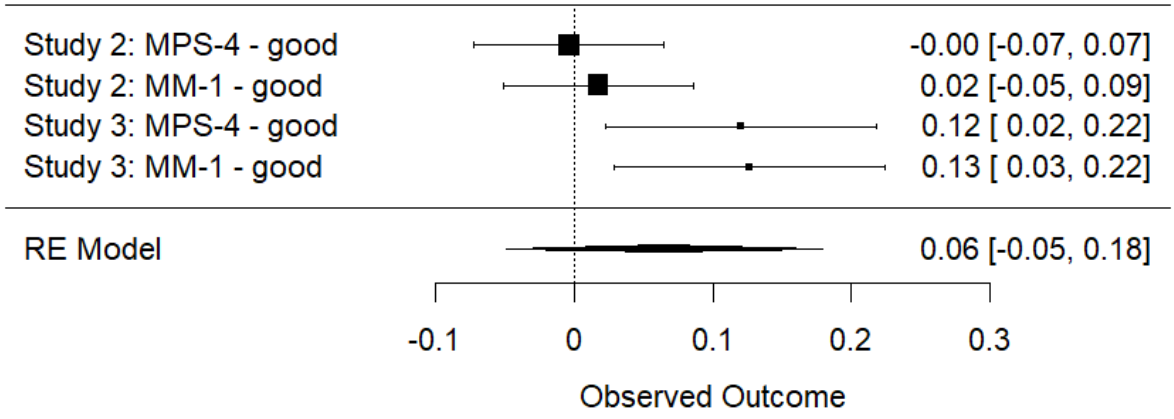

Figure S19. Forest plot showing effects for good characters Studies 2 and 3 and pooled effect.

Overall, there was a significant dilution effect across all studies, for bad and good characters, and for both measures, $d_{pooled} = 0.11$, $SE = 0.03$, $z = 3.80$, $p < .001$, 95% CI [0.05, 0.16].

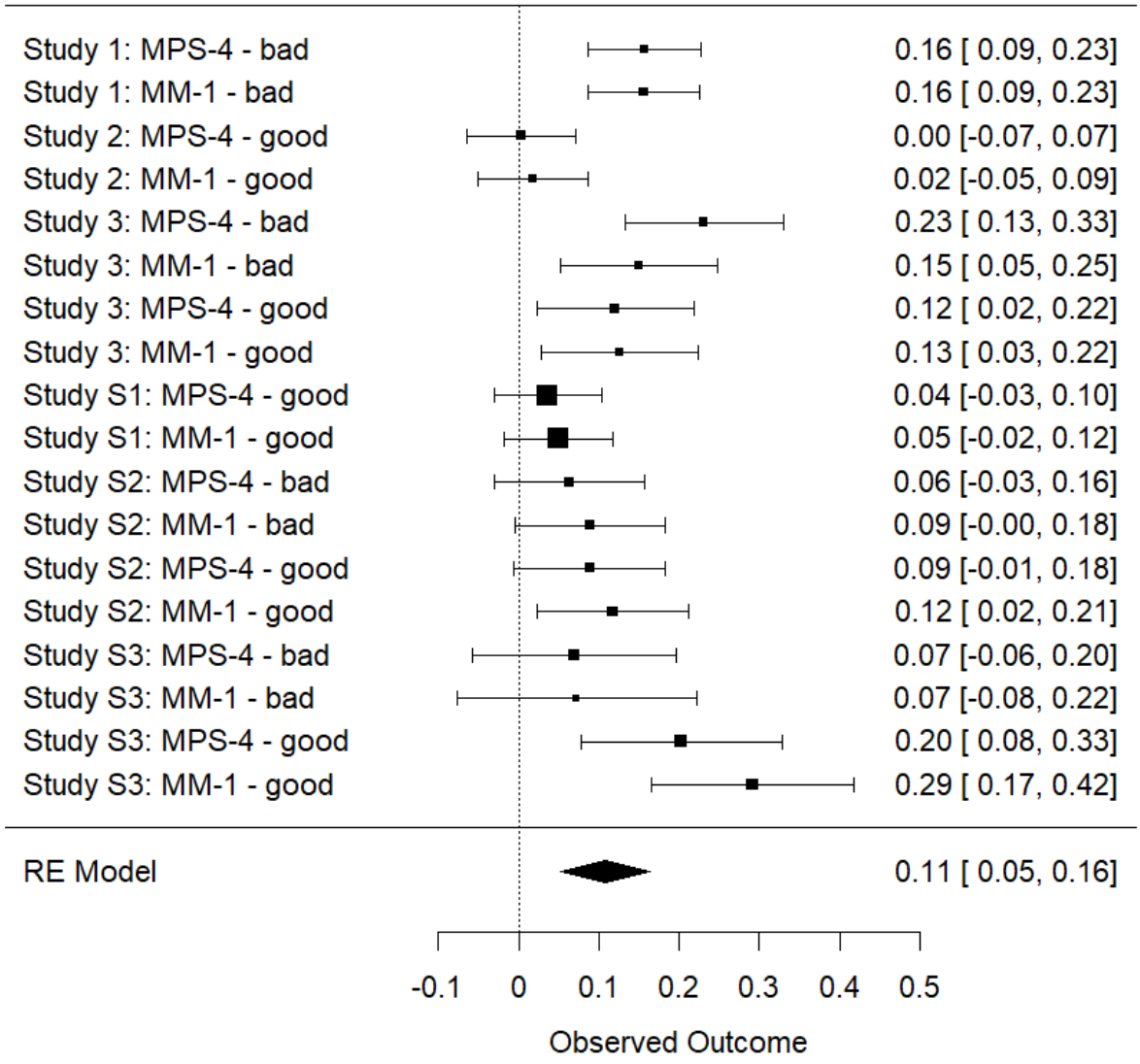

Figure S20. Forest plot showing effects for all studies and pooled effect.

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

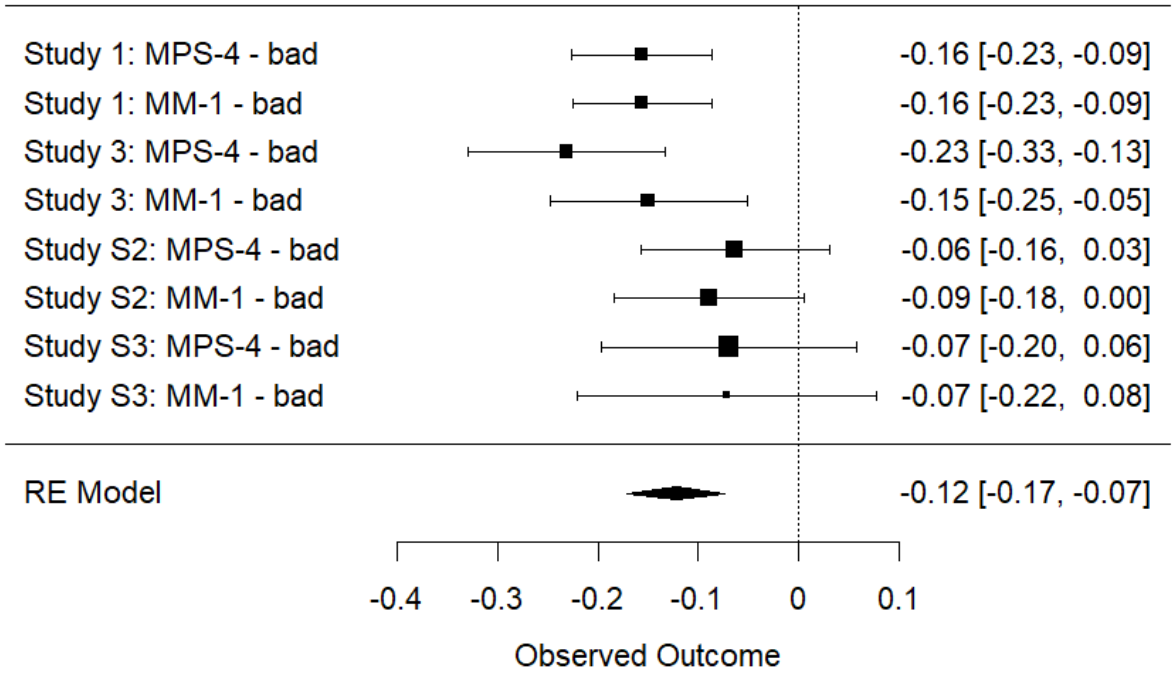

Figure S21. Forest plot showing effects for bad characters for all studies and pooled effect.

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

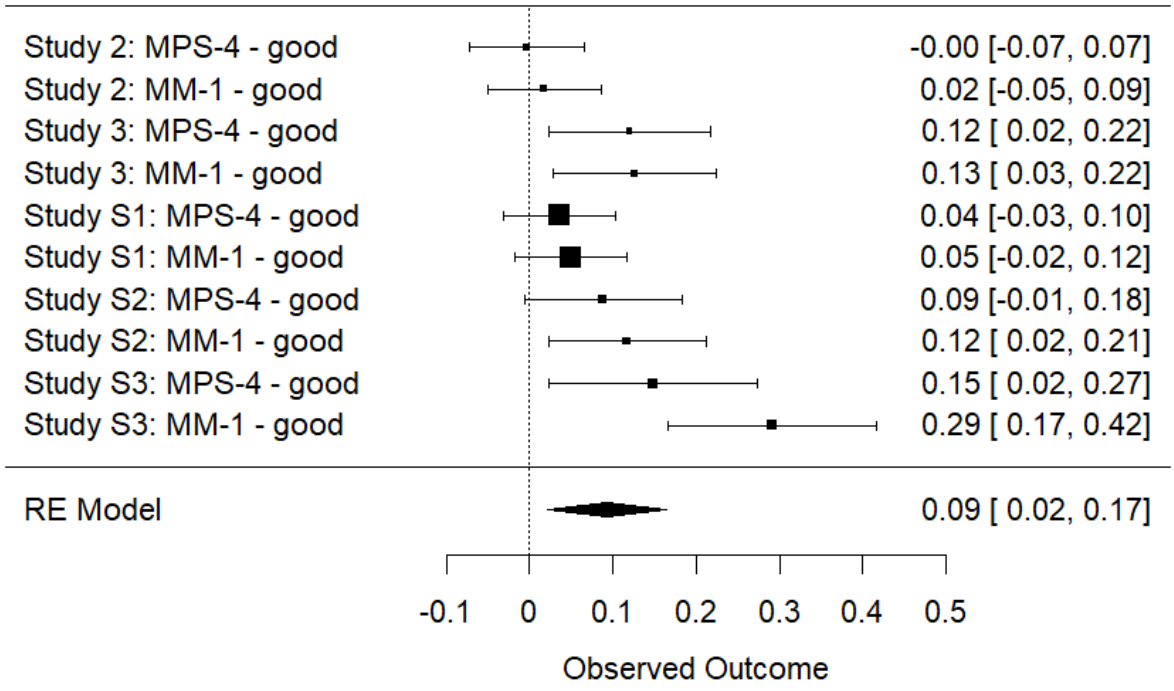

Figure S22. Forest plot showing effects for good characters for all studies and pooled effect.

964 **References**

965 Grizzard, M., Fitzgerald, K., Francemone, C. J., Ahn, C., Huang, J., Walton, J., ... Eden,
966 A. (2020). Validating the extended character morality questionnaire. *Media Psychology*,
967 *23*(1), 107–130. https://doi.org/10.1080/15213269.2019.1572523
968 Walker, A. C., Turpin, M. H., Fugelsang, J. A., & Białek, M. (2021). Better the two devils
969 you know, than the one you don't: Predictability influences moral judgments of
970 immoral actors. *Journal of Experimental Social Psychology*, *97*, 104220.
971 https://doi.org/10.1016/j.jesp.2021.104220