Stimulus-category confounds reduce, but do not explain, different roles of intent across moral domains: Valid statistical inference helps explain 'inconsistent' findings

Joseph Sweetman[1*] & Ryan M McManus[2]

[1]Department of Psychology, Faculty of Health and Life Sciences, University of Exeter, Exeter, Devon, United Kingdom

[2]Department of Psychology and Neuroscience, Boston College, Boston, Massachusetts, USA

* Corresponding author at

E-mail: j.p.sweetman@exeter.ac.uk

Acknowledgments: We would like to thank Liane Young for her helpful feedback and encouragement throughout this project. We would also like to thank Tom Kupfer and Alek Chakroff for helpful feedback on the expanded stimulus set. Finally, we thank Lily Tsoi and Joshua Rottman for helpful conversations at the beginning of this project.

Abstract

The reduced role of intent when judging impure, vs. harmful, acts provides evidence that our moral mind/brain may be composed of distinct, domain-specific systems. However, a study that manipulated and statistically adjusted for the target (i.e., self vs. other) of the act reported a significant, but notably smaller, intent × domain effect (Chakroff et al., 2013). Recent studies that either manipulated or controlled for the context, severity, weirdness, and target of the act reported nonsignificant effects (Kupfer et al., 2020; Parkinson & Byrne, 2017). We tested two competing hypotheses to explain these inconsistent findings: the intent × domain effect is a methodological artifact of the stimuli employed, or the effect is merely reduced in magnitude when controlling for potential stimulus-category confounds. To test these hypotheses, we conducted a close replication of Kupfer et al. We found a small, significant intent × domain effect, supporting the reduced-magnitude hypothesis. We find that the "inconsistent" findings can be explained by the reduced magnitude of the effect when employing carefully controlled stimuli, and by whether researchers test support for the null hypothesis, conduct sample size simulations, and perform model checking and appropriate robustness checks when modeling ordinal (i.e., Likert) items as interval data. Indeed, Kupfer et al.'s original null result becomes significant when an ordinal mixed model is fitted. Our findings are consistent with perspectives that view the moral mind/brain as constituted by multiple moral-cognitive modules. However, they are also consistent with a single moral-cognitive system with domain-dependent modulation. While carefully controlled experimental work and valid statistical inference help the replicability, robustness, and reproducibility of psychological science, delineating and decomposing the moral mind/brain will also require reforms to theory development.

Keywords: intent x moral domain, purity, harm, statistical inference, replication crisis, (ordinal) mixed effects models, theory crisis

Stimulus-category confounds reduce, but do not explain, different roles of intent across moral domains: Valid statistical inference helps explain 'inconsistent' findings

Accidentally head-butting a sibling while trying to hug them may display bad coordination, but it is not necessarily morally bad. In contrast, the exact same action, when intentionally performed, is almost certainly morally wrong. It is the mental state of the agent, in this case, their intent, or lack thereof, that distinguishes the moral status of these two acts. The importance of intent to moral cognition concerning harmful acts is clear; what is less clear is its role concerning "impure" acts like sex between siblings or ingesting taboo substances like your own (or another's) urine. Classic work on moral cognition has suggested that the importance of intent is much weaker for moral cognition concerning impure (vs. harmful) acts: the intent × domain effect. Such findings have been used to support the idea that moral cognition concerning harmful and impure acts is underpinned by multiple domain-specific modules. However, recent studies, purportedly employing more controlled stimuli, have reported a smaller intent × domain effect or a failure to find any evidence for the effect.

Here, we reexamine the different roles of intent across moral domains, testing two competing hypotheses that could explain these inconsistent experimental findings. First, it is possible that the intent × domain effect is a methodological artifact resulting from the stimuli employed to manipulate the moral domain of the act being judged: what we will term the *confounded-stimulus-category hypothesis*. Second, it is possible that the intent × domain effect is a "real" effect but is reduced in magnitude when statistically adjusting for, and/or employing stimuli that control for, the context, target (i.e., self vs. other) severity, and "weirdness" of the act: what we will term the *reduced-magnitude hypothesis*. We test these

competing hypotheses by conducting a pre-registered, close replication of one of the studies reporting a null intent × domain effect, and we meta-analyze the results from the original and replication studies to further test our hypotheses.

Multiple Moral-Cognitive Modules and the Role of Intent Across Moral Domains

One view on the moral mind/brain is that it is underpinned by multiple domain-specific cognitive modules. Moral foundations theory (MFT) is one leading example of this view, positing a set of five distinct mental modules, or "foundations," each processing domain-specific moral information (Graham et al., 2012; Haidt & Joseph, 2004). Each of these moral modules is said to be "triggered" by domain-specific actions (i.e., violations and virtues). For example, the harm module is said to be triggered by suffering, distress, or neediness (e.g., assault would be a prototypical harm violation). MFT posits that the purity module is triggered by bodily fluids, taboo diets, or sexual practices (e.g., incest and drinking urine would be prototypical purity violations; for full details, see Graham et al., 2012; Haidt & Joseph, 2004).

Drawing inspiration from evolutionary psychology and anthropology, MFT holds that each of these moral modules is an adaptation to recurrent problems of social coordination. We set aside the question of how challenging it is to say much of scientific value about the evolution of high-level human cognition (see Lewontin, 1998). Instead, we point out that, regardless of its phylogeny or adaptive value/function, understanding whether human moral cognition is constituted by multiple, domain-specific modules is a fundamental question in the study of the moral mind/brain. Indeed, correctly delineating and decomposing cognitive phenomena is an important step in developing theoretical explanations in psychology and cognitive (neuro)science (Bechtel, 2012).

Some of the strongest evidence for multiple moral-cognitive modules stems from a set of four experiments by Young and Saxe (2011) showing that the exculpatory effect of innocent intentions was significantly reduced for impure, compared to harmful, acts. For example, accidentally putting poison in a coworker's coffee was judged as much less morally wrong than doing the same action intentionally. However, accidentally washing your mouth out with urine was judged as only slightly less morally wrong than doing so intentionally. In other words, the simple main effect of mental state (intentional vs. accidental) was stronger for harmful than for impure acts. Young & Saxe showed evidence for the intent × domain effect across Experiments 1A (N = 262), 1B (N = 80), 2 (N = 320), and 3 (N = 160), employing a total of two harmful (e.g., poisoning a co-worker, causing an allergic reaction) and four impure (e.g., incest, drinking your own urine) acts.

Subsequently, converging evidence for the intent × domain effect has been obtained using both behavioral and fMRI methods (Chakroff et al., 2016). Chakroff et al. provided converging behavioral evidence for the intent × domain effect, employing 24 harmful and 24 impure acts (N = 23). In addition, imaging data from this experiment showed that the right temporoparietal junction (RTPJ), an area recruited for mental state reasoning or theory of mind, was preferentially recruited to process harmful vs. impure acts. Multivoxel pattern analysis revealed that spatial patterns of activity in RTPJ distinguished the mental state of the agent for harmful, but not impure, acts (Chakroff et al., 2016) . These findings have largely been replicated in recent work employing 28 harmful and 28 impure acts (Dungan & Young, 2019; N = 29). The authors also found that the intent × domain effect remained even when participants were instructed to focus on "why" (vs. how) harm and purity actions were being carried out (Dungan & Young, 2019). More converging evidence for the intent × domain effect comes from cross-cultural work across several traditional small-scale societies (e.g., hunter-gatherer and pastoralist) and Western societies (Barrett et al., 2016). Furthermore, the

intent × domain effect has been successfully replicated in an independent, pre-registered, close replication of one of the original experiments (Young & Saxe, 2011; Experiment 1B) documenting the effect (Sweetman & Newman, 2020) . Taken together, this converging evidence seems to offer some support for the intent × domain effect and is consistent with the idea that moral cognition concerning harmful and impure acts may be underpinned by multiple domain-specific cognitive modules.

Confounded-Stimulus Category or Reduced Magnitude?

Critics of the multiple moral modules view claim that the above evidence for multiple, domain-specific moral modules reflects confounds (e.g., the weirdness, severity, target, and context of the act) in the stimuli used to manipulate an act's moral domain (Gray & Keeney, 2015). Gray & Keeney found that typical impure acts (e.g., incest, drinking urine, etc.) are both "weirder" (i.e., more atypical) and less severe than the harmful acts (e.g., assault, poisoning) employed in this literature. For example, the impure act of washing your mouth out with urine seems, intuitively, weirder and less severe than the harmful act of putting poison in a coworker's coffee. Gray & Keeney claim that it is these stimulus-category confounds, not the type of moral domain, that are responsible for reported domain effects in moral judgments. Controlling for weirdness and severity, the authors showed that the weirdness of the act was more predictive of character evaluations than the act's moral domain. This supports an alternative explanation, to multiple moral-cognitive modules, of earlier work suggesting that impure (vs. harmful) acts led to more lenient moral judgments of action but harsher judgments of moral character (Uhlmann & Zhu, 2014): the judgment type × domain effect.

The impure and harmful acts employed in research testing the intent × domain effect also tend to vary in terms of the target of the act. Specifically, impure acts (e.g., drinking

urine) tend to be self-directed, whereas harmful acts (poisoning a coworker) tend to be other-directed. Chakroff et al. (2013) explored the relation between intent (i.e., intentional vs. accidental), domain (i.e., harm vs. purity), target (i.e., self vs. other), and moral judgment type (i.e., action vs. character). Employing a between-subjects design with four harmful and four impure acts (Experiment 1; N = 331), the authors still found a significant intent × domain effect, adjusting for the manipulation of the act's target. This finding suggests that the potential target confound in the stimulus category may not, on its own, fully explain the intent × domain effect. It should be noted that the effect size reported from the ANOVA model, $F(1,315) = 13.17$, $p < .001$, $\eta_p^2 = .04$, is smaller than those reported in the original experiments in Young & Saxe (2011) that did not manipulate, and adjust for, the target of the act (Young & Saxe: Experiment 1A: $\eta_p^2 = .15$ and .23; Experiment 1B: $\eta_p^2 = .07$; Experiment 2: $\eta_p^2 = .17$; Experiment 3: $\eta_p^2 = .05$ and $\eta_p^2 = .05$). We could take this as prima facie evidence consistent with the hypothesis that the intent × domain effect exists, but its magnitude is when the target of the act is statistically adjusted for (i.e., the reduced-magnitude hypothesis). However, it should be noted that the context of the act was not carefully controlled; for example, closing a door on someone's fingers vs. serving someone dog meat.

Recently, Parkinson & Byrne (2017) reported finding a null intent × domain effect, $F(1, 21) = 2.29$, $p = .15$, when exploring the relation between intent (intentional vs. accidental), domain (harm vs. purity), target (self vs. other), and moral judgment type (wrongness vs. responsibility). The authors employed a within-subjects design with four harmful and four impure acts (Experiment 2; N = 22) that were more closely matched for context than previous harm and purity stimuli. For example, an agent either intentionally or accidentally pours themselves, or another, a glass of poisonous horse medicine (harmful act) or horse urine (impure act). We could take their null finding as prima facie evidence that the

intent × domain effect is a methodological artifact due to the stimuli employed to manipulate the moral domain of the act (i.e., the confounded-stimulus-category hypothesis). However, it should be noted that there is an important difference between simply failing to reject a null hypothesis (i.e., a nonsignificant, $p > .05$, finding) and accepting or finding support for the null hypothesis (Gallistel, 2009; Kruschke, 2018; Lakens, 2017).

Recently, Kupfer et al. (2020) also reported finding a null intent × domain effect, $F(1, 13) = 3.63$, $p = .079$. In a pre-registered, within-subjects experiment, the authors employed 14 acts (Experiment 1; N = 237), varying the agent's intent (intentional vs. accidental) and the act's moral domain (harmful vs. impure). While Parkinson & Byrne (2017) and Chakroff et al. (2013) controlled for the target (i.e., self vs. other) of the act by experimentally manipulating and statistically adjusting for it, Kupfer et al. instead employed harm and purity stimuli matched for target. Specifically, all impure and harmful acts affected another person or group of people (i.e., were other-targeted). In addition, the authors employed harmful and impure acts that were matched well in terms of context, severity, and weirdness (although the latter was not empirically tested). For example, farting in a full lift just before exiting it (impure act) or pressing all the floors in the lift just before exiting it (harmful act).

Unlike Parkinson & Byrne (2017) and Chakroff et al. (2013), Kupfer et al. (2020) employed a linear mixed-effects model (LMM) with random intercepts and slopes for both participant and act or "scenario." This addresses issues of stimulus sampling and generalizability (Judd et al., 2012), which have already undermined classic effects in moral cognition research (McGuire et al., 2009). However, using a random factor with 14 levels imposes a significant limitation on power (see *Sensitivity and sample size simulations*). For example, one rule of thumb for researchers using LMMs suggests that random factors should have at least 20 levels (Singmann & Kellen, 2019). Again, we could take Kupfer et al.'s null finding as prima facie evidence for the confounded-stimulus-category hypothesis, with the

proviso that the study may be underpowered, and, as with Parkinson and Byrne, we cannot truly infer support for a theory or hypotheses predicting a null effect from a single nonsignificant result: absence of evidence is not evidence of absence.

The Present Research

Here, we reexamine the different roles of intent across moral domains by testing between two competing hypotheses that could explain the inconsistent experimental findings detailed above. One explanation, the confounded-stimulus-category hypothesis, holds that the intent × domain effect is a methodological artifact arising from stimulus-category confounds (i.e., context, target, severity, and weirdness of the act) in the stimuli used to manipulate the act's moral domain. A competing explanation, the reduced-magnitude hypothesis, holds that the intent × domain effect is a real effect but is reduced in magnitude when statistically adjusting for, and/or employing stimuli that control for, the context, target, severity, and weirdness of the act. We test these competing hypotheses by conducting a pre-registered, close replication of Kupfer et al. (2020).

Following Kupfer et al. (2020), we address the potential context, target, and severity confounds by employing mundane, harmful and impure other-targeted acts performed in similar contexts. While these confounds in Kupfer et al. were either objectively controlled (e.g., use of only other-targeted acts) or statistical analysis suggested that they were controlled (e.g., no significant main effect of domain on severity of moral judgment), it was less clear whether the weirdness confound was controlled. Therefore, here we added a measure of weirdness to examine whether the harmful and impure acts vary in terms of their weirdness. In addition, we perform meta-analysis of our findings and those from the aforementioned studies that allow us to test our hypotheses.

We took a comprehensive analytic approach to test our hypotheses in four respects: 1) the estimation of effect sizes and their precision and meta-analysis, 2) testing support for the null hypothesis, 3) employing LMMs to address stimulus sampling and generalizability issues and 4) conducting simulations in order to plan for sufficient statistical power with a practical balance of participants and scenarios (see Brysbaert & Stevens, 2018). In addition, we estimated all models in both frequentist and Bayesian frameworks. This allowed us to utilise the strengths and address some of the weaknesses of the respective frameworks (see Nicenboim & Vasishth, 2016; Vasishth & Nicenboim, 2016; Wagenmakers et al., 2017).

Pre-Registered, Close Replication of Kupfer et al. (2020; Experiment 1)

To test our hypotheses, we decided to conduct a pre-registered, close replication of one of the studies reporting null results. Specifically, we chose to replicate Experiment 1 of Kupfer et al. (2020) with the aim of obtaining a sample size with which to have both sufficient power (95%) to detect an effect greater than our smallest effect size of interest (SESOI; $d \geq .20$[1]) and precision in our estimates to test support for the null hypothesis (i.e., the confounded-stimulus-category hypothesis).

Method

We pre-registered our replication attempt. This preregistration was performed prior to data collection and analysis and is available at

[1] Here we specify a Cohen's d of ± 0.2 (i.e., $d_L = -0.2$ and $d_U = 0.2$) as the SESOI. That is, 0 ± 0.2**SD* or a "small effect" ($d = 0.2$). Generally, moral cognition research is not at a stage in its development where we have the benefit of explanatory models that predict the magnitude of effects. That said, our chosen SESOI has two reasons to recommend it. First, it is approximately twice the size of an effect size ($d = 0.11$) that both proponents (Schnall et al., 2015) and critics (Landy & Goodwin, 2015) of the multiple moral modules view agree is too small to be of theoretical interest. Second, it seems untrivial given it is twice the size of a proposed rule of thumb ($d = \pm 0.1$) for SESOI when calibrating the meaning of effect sizes is difficult (Kruschke, 2018).

https://osf.io/3f6c4/?view_only=b1b1d11dd00c4e6bb60054b0ee2ca20a. Further, all materials, data, and analysis scripts are openly available at https://osf.io/csxzt/?view_only=b05e41155b3849b586630ada963acb46. Our pre-registered method closely follows that reported in the original article and its supplemental materials[2]. We contacted the first author who very helpfully supplied us with a copy of the Qualtrics project that he employed for the original experiment. Any differences from the originally reported study and our pre-registration are clearly explained in the following sections. All measures, manipulations, and exclusions in the study are disclosed (see below) and in accordance with the pre-registration plan. The method of determining the final sample size is provided below, and all data analysis was conducted after completion of data collection, in accordance with the pre-registration plan. Before participation in the online experiment, participants were given information about the general nature of the experiment without giving details of the research questions or hypotheses. Participants were informed that participation was voluntary and that they could exit the study at any time without this affecting their payment. They were given details of how their data would be handled and contact details if they had concerns. The information sheet and informed consent are available in the OSF repository as part of the Materials at https://osf.io/csxzt/?view_only=b05e41155b3849b586630ada963acb46. The materials used in the experiment had been granted ethical approval from the University of Exeter Psychology ethics committee (ethics application ID: 3546062).

[2] The Supplementary Materials associated with Experiment 1 were missing the text for the harmful version of scenario 12 (Susan public defecation/bottle smashing); exploring the OSF repository enabled us to locate the missing scenario in an original pre-registered version of the project.

Participants and Design

234 participants (122 women, 109 men, two nonbinary, and one who preferred not to say) were recruited via Prolific. Following Kupfer et al. (2020), participants were required to be fluent in English and native to the U.S. Ages ranged from 18 to 76 years (M = 40.92, SD = 12.13). The experiment took approximately 45 minutes to complete. Participants were therefore paid £5.50 to complete the study at an hourly rate of £7.33. As in Kupfer et al., the experiment employed a 2 (intent: intentional vs. accidental) x 2 (domain: harm vs. purity) within-subjects design, with participants randomly assigned to one of the four conditions/versions of each scenario.

Sensitivity and sample size simulations

Given the lack of analytic methods for computing sample sizes for designs with multiple fixed and random factors, we ran simulations before data collection in order to attain a required sample size (Brysbaert & Stevens, 2018). We were able to reproduce Kupfer et al.'s (2020) results by fitting a model with the maximal random-effects structure (i.e., all intercepts and slopes for the by-participant and by-scenario random effects and their correlations), finding a small, nonsignificant intent × domain effect, b = 0.51, 95% CI [−0.03, 1.06], p = .079. While the maximal model did not always produce warnings, it sometimes did so for different versions of R, lme4, and operating systems (i.e., Mac vs PC).[3] Therefore, our simulations were based on the estimates from a reduced model, dropping the correlation between the by-participant intercept and slope for the intent × domain effect. The

[3] Further investigation of all models (via PCA inspection) consistently provided evidence of overfitting (for details, see "Overfitting problems for the maximal model in Kupfer, Inbar, & Tyber [2020]" in Supplementary Materials at https://osf.io/csxzt/?view_only=b05e41155b3849b586630ada963acb46)

model had by-scenario and by-participant variance components that accounted for 34% and 37% of variance (not explained by the fixed effects), respectively.

First, we conducted a sensitivity analysis for Kupfer et al. (2020) with our SESOI for the intent × domain effect, $d = 0.2$. The simulation revealed that the maximum attainable power was 25%, 95% CI [.21, .28]. With an effect size of $d = 0.2$ and only 14 scenarios, statistical power asymptotes at $N_{(participants)} = 250$[4]. At this point, adding more participants does not increase statistical power (see Figure 1 and for the R Markdown file, see "Sims for Close Replication of Kupfer et al. [2020; Experiment 1]" in Supplementary Materials at https://osf.io/csxzt/?view_only=b05e41155b3849b586630ada963acb46). Researchers have become increasingly aware of the need for a greater number of participants to ensure adequate statistical power. However, those employing LMMs, where participants and scenarios or items are treated as random factors, may be less aware that low numbers of scenarios or items can have a serious detrimental effect on statistical power (Judd et al., 2017; Westfall et al., 2014). We suggest that researchers use freely available simulation tools, so that they can plan for sufficient statistical power with a practical balance of participants and scenarios or items (see Brysbaert & Stevens, 2018).

Ideally, close replications try to follow experimental methods exactly, trying to reproduce findings in a new sample from the same population of interest. Further, a sufficient

[4] In the original version of this paper, we mistakenly reported 34%, 95% CI [.32, .37] power to detect an effect of $d = 0.2$. Similarly, power for our replication was also mistakenly reported as 96% power, 95% CI [.94, .97]. This was due to an error in our computation of sigma, where we used the sample standard deviation. For comparison purposes and to be consisted with the standardised effect size used throughout the paper we now compute the pooled standard deviation (sigma) using the square root of all the variance terms in the mixed model, as outlined in Judd et al. (2017). Further, initially the simulations were based on a reduced model that dropped the by-participant random slope for the target effect. We now, instead, drop the correlation between the by-participant intercept and slope for the target effect.

number of new participants (usually more than in the original study) are required to ensure a high-powered replication. This poses obvious problems for researchers trying to replicate studies where samples of participants respond to samples of stimuli where the number of stimuli employed in the original experiment place an undesirable limit on the maximum attainable power. Westfall et al. (2015) suggest that a solution in such cases is either to employ an entirely new but larger stimulus sample or expand the original stimulus sample with a new but comparable sample. We choose to expand the original stimulus sample with new but comparable stimulus sets: specifically, stimuli that control for the severity, weirdness, context, and target of the harmful or impure act[5].

Simulation revealed that 224 participants and 70 scenarios were required to detect an effect of $d = 0.2$ with 80% power, 95% CI [.76, .83] and an α-level of 0.05 (see Figure 2 and for the R Markdown file, see "Sims for Close Replication of Kupfer et al. [2020; Experiment 1]" in Supplementary Materials at https://osf.io/csxzt/?view_only=b05e41155b3849b586630ada963acb46). This is the same number of participants but five times as many scenarios as originally sampled by Kupfer et al. (2020). We thus aimed for a sample of 235 participants, which is 5% greater than the sample size required in order to take into account participants excluded from our planned analysis due to failing the attention check or reporting previously taking part in Kupfer et

[5] The use of new stimuli in close replications may strike the reader as counterintuitive. However, if one accepts that we are sampling stimuli in the same way as we are sampling participants, then enlarging, or indeed replacing, the stimulus set with comparable stimuli is analogous to sampling new but comparable (i.e., from the same population of interest) participants – something a researcher replicating an experiment would not think twice about (see Westfall et al., 2015).

al.'s original experiment (see below and Pre-Registration at https://osf.io/3f6c4/?view_only=b1b1d11dd00c4e6bb60054b0ee2ca20a).

Materials and Procedure

Participants made moral judgements of the 14 acts taken directly from Kupfer et al.'s (2020) Qualtrics materials and 56 additional acts generated from the original 14. That is, we made an additional 4 versions of each of the original 14 harmful and impure actions, taking care to keep the structures of the acts as similar as possible but the content different enough to be easily distinguishable. For example, an agent ("Frank") either intentionally (vs. accidentally) pulls an emergency brake when exiting a subway car (i.e., a harmful act), or intentionally (vs. accidentally) farted when exiting the subway car (i.e., an impure act). For full details, see the Materials at https://osf.io/csxzt/?view_only=b05e41155b3849b586630ada963acb46.

Participants judged the moral wrongness of the acts on a 7-point scale, anchored at (0) "*not at all wrong*" to (6) "*extremely wrong*." Following Kupfer et al. (2020), participants also made judgments of the target's intention on a 7-point scale, anchored at (0) *"definitely not intentional*" to (6) "*definitely intentional*." And judgments of how harmful the target's behavior had been on a 7-point scale, anchored at (0) *"not at all harmful*" to (6) "*extremely harmful.*" Following Gray & Keeney (2015), we also asked participants to judge how impure (i.e., involving sinfulness, indecency, dirtiness) the act was on a 7-point scale, anchored at (0) *"not at all impure*" to (6) "*extremely impure*." And how atypical (i.e., weird, strange, unusual) the act was on a 7-point scale, anchored at (0) *"not at all* atypical" to (6) "*extremely* atypical." We conceptualize these measures as direct and discriminant (in the case of weirdness) manipulation checks. Accordingly, we dropped two items measuring anger and disgust that were included in Kupfer et al. that do not constitute direct manipulation checks.

We employed the same attention check questions as Kupfer et al. (2020). These included asking, “How many bunnies do you see above?” with an image above the item showing 2 bunnies and 2 kittens; “What color is the sky on a clear day?” with participants choosing from among 5 responses “red,” “blue,” “pink,” “silver,” and “green,”; and asking participants “How seriously did you take participating in this survey? (Please be honest - your answer won't affect your payment or Prolific rating)” with three responses: “seriously,” “somewhat seriously,” and “not at all seriously.”

As we could copy the original Qualtrics project from the primary author, the experimental materials were almost identical. There were only three main differences. First, we employed an extra item after all of the original items, “Did you participate in the same study a couple of years ago” (Please be honest - your answer won't affect your payment or Prolific rating) with participants choosing from three responses: “yes, I definitely remember taking part in the same study,” “yes, I think I took part but I’m not 100% sure,” “no, I don’t think I’ve taken part in this study previously.” Second, the direct and discriminant manipulation checks were presented in a random order following the primary dependent variable. Finally, we also randomized the order of the scenarios. The only other differences from the initial experimental materials were the correction of a few typos and non-American English spellings, the removal of the original author’s institutional affiliation and logo in the Qualtrics template, and the relevant differences in the study information, informed consent, and debrief given to participants. Everything else was identical.

Statistical Analysis

We attempted to specify maximal LMMs with intent, domain and their interaction added as fixed factors and participant and scenario as random factors. While we attempted to specify the random-effect structures of models as close to maximal as possible (Barr et al.,

2013), at the same time we aimed to balance Type I error and power by removing random effects that did not improve model fit, assessed by likelihood ratio test (see Matuschek et al., 2017). This is essentially the same as the analysis conducted by Kupfer et al. (2020), except we reported models with unstandardized and standardised effect sizes and their 95% confidence intervals and performed an in-depth examination of models for signs of non-convergence and overfitting. Another difference with the original authors' analytic approach was that we ran the same LMMs in the Bayesian framework, with 95% credible intervals (CrI) and the probability for parameters computed from the posterior distribution. We employed weakly informative priors for all Bayesian models, setting a 95% prior probability that the true intent × domain effect size lies between ±2 SDs. Assuming a similar SD as Kupfer et al. for moral judgments, this corresponds to a 95% prior probability that the intent × domain effect size lies between − 4.32 and 4.32 on the raw 7-point scale[6].

The final difference in our analytic approach was that we also tested support for the null hypothesis (confounded-stimulus-category hypothesis) with the two one-sided t-tests (TOST) test of equivalence (Lakens, 2017) for our frequentist models and the posterior highest density interval and region of practical equivalence (HDI+ROPE) procedure (Kruschke, 2018), probability computed from the posterior distribution, and Bayes Factors for our Bayesian models (for more details, see "Extended Details of Analytic Approach" in Supplementary Materials at https://osf.io/csxzt/?view_only=b05e41155b3849b586630ada963acb46). We specified a

[6] This Gaussian prior centered on 0 is weakly informative given 1) the 7-point scale, 2) a mean difference of 2.85 for the simple main effect of intent for harmful acts in Kupfer et al., and 3) that the literature shows that the simple main effect of intent for impure acts is of smaller magnitude, regardless of its direction. In any case, we performed prior sensitivity checks.

Cohen's *d* of ± 0.2 (i.e., *d*L = − 0.2 and *d*U = 0.2) as the SESOI or region of practical equivalence. Since, unlike Kupfer et al., we included a measure of the perceived weirdness of the act, we were also able to fit the models adjusting for weirdness, if we found a significant main effect of domain on this measure. Specifically, we fit models also including our weirdness measure and the interaction between this measure and our intent manipulation.

Consistent with Kupfer et al. (2020), we excluded data from participants who answered either of the attention check questions incorrectly or reported taking the survey "not at all" seriously. In addition to the original exclusion criteria, we also excluded participants who reported that they "definitely remember taking part in the same study" (see Pre-Registration Plan at https://osf.io/3f6c4/?view_only=b1b1d11dd00c4e6bb60054b0ee2ca20a). In addition to standard group-level analysis, we also examine the intent × domain effect at the within-person or single-subject level.

Results and Discussion

Six participants failed one of the attention checks, and two reported "definitely" taking part in the previous study. This resulted in a final sample for analysis of N = 226. We performed all analyses with packages in the R programming language (Team, 2023).

Descriptives, correlations, reliability, and (a little) validity for measures

Table 1 shows the means, standard deviations, zero-order correlations, and reliabilities for our measures of moral wrongness and the direct and discriminant manipulation checks. On average, scores were just over the midpoint (3) of the scale for all measures. There were moderate to strong positive correlations between moral wrongness and the manipulation checks (*r*s .54 to .83). Judging acts as more impure was positively correlated with perceptions of the weirdness of the act (*r* =.74) and of its harmfulness (*r* =.78). We computed coefficients omega total (ω_t), omega hierarchical (ω_h), and Cronbach's alpha (α) for each of our measures

in a confirmatory factor analysis (CFA) framework. Since participants were randomly allocated to one of the four conditions for each of the 70 items, the raw item scores reflect residual noise and both true score and systematic variance due to experimental manipulations. However, there was not enough data ($p > n$) to support a full structural equation model (SEM) accounting for the impact of the manipulations. Therefore, we fitted a simple one-factor CFA model and supplemented this by fitting an approximation of a one-parameter linear item response theory (IRT) model, extended to include the experimental manipulations and their interaction as predictors. This allows us to examine the stability of latent constructs using a by-participant random intercept, conceptually analogous to the latent factor score in CFA, while adjusting for experimental manipulations.

All reliability coefficients suggested good to excellent (.78 to .97) reliability for all measures (for details of the advantages of coefficient omega, see Flora, 2020). While the absolute model fit for one-factor models fitted to each of our measures suggested excellent to adequate fit (RMSEAs .041 to .054), relative or "incremental" fit indices suggested poor fit (CFIs = .33 to .80 and TLIs = .31 to .79). Therefore, some caution should be taken when interpreting the reliability (and assuming the convergent and structural/factorial validity) of the measures. The large number of indicators, sample size, and experimental manipulations seemed to explain this discrepancy between relative and absolute fit (for full details, see "Descriptives, Correlations, and Reliabilities" in Supplementary Materials at https://osf.io/csxzt/?view_only=b05e41155b3849b586630ada963acb46). Taken together, these results provide some evidence for the reliability and convergent and structural/factorial validity of our measures of moral wrongness, and the direct and discriminant manipulation checks.

Mixed-effects models

Frequentist and Bayesian LMMs were fitted using *lme4* (Bates et al., 2015) and *brms* (Bürkner, 2017), respectively. We conducted model checks using *performance* (Lüdecke et al., 2021). Inspection of influence statistics did not reveal any significant, influential cases. However, visual inspection of residual plots revealed some deviations from normality and homoscedasticity for most models. Comparisons of the observed data to simulated data generated from the models (i.e., posterior predictive checks) revealed multimodal distributions, with the observed data clustering at the bounds (0 and 6), where it protruded significantly above and below the normal distributions. To illustrate these patterns, a representative set of model checks are shown in Figure 3 (for full details, see "Model Checks" in Supplementary Materials at https://osf.io/csxzt/?view_only=b05e41155b3849b586630ada963acb46). Strictly speaking, the model checks suggest that most (possibly all) of the LMMs may be misspecified, they may not entirely satisfy model assumptions.

The model checks are consistent with two possible (related) threats to valid statistical inference: modeling ordinal (i.e., Likert) items as interval data and ceiling and floor effects (CFEs). Linear models assume that the increment in the psychological attribute of, say, for example, moral wrongness between "(0) not at all wrong" and "(1)" is the same as that between "(5)" and "(6) extremely wrong." Treating ordinal items as interval can lead to distorted (even inverted) effect size estimates and inflated Type I and II error rates (Liddell & Kruschke, 2018). Šimkovic & Träuble (2019) have shown that estimates of uncrossed interactions (such as the intent × domain effect) from common parametric tests may be especially unreliable when CFEs are present. Therefore, we also fitted frequentist and Bayesian cumulative link mixed-effects models (CLMMs) with *ordinal* (Christensen, 2020) and *brms* (Bürkner, 2017), respectively. Such models are more appropriate for ordinal data and provide robust estimates when CFEs and multimodality are present (Harrell, 2015).

This allows us to test LMM assumptions (Liddell & Kruschke, 2018) and the robustness of our findings to possible CFEs (Šimkovic & Träuble, 2019). For full details of the CLMMs, see "Robustness Checks" in Supplementary Materials at https://osf.io/csxzt/?view_only=b05e41155b3849b586630ada963acb46.

Our factors were deviation (sum) coded – domain: $-.5$ harm, .5 purity; intent: $-.5$ intentional, .5 accidental, allowing for ease of interpretation of the coefficients. The coefficients for the "main effects" can be interpreted as mean differences or "changes" on the original scale, and the coefficient for the intent × domain effect can be interpreted as the difference of these mean differences. For brevity, familiarity, and to aid comparison with previous effect sizes, we provide estimates and 95% CIs for the unstandardized and standardized effect sizes (i.e., Cohen's *d*)[7], and *p*-values for all frequentist models below. When describing the magnitude of effects (e.g., "small," "medium," and "large") we simply adopt Cohen (1992) rules of thumb. When describing effects as "significant," we simply mean that an effect of at least the observed magnitude would be unlikely were the null hypothesis true. For the test of our main hypotheses, we also report Bayes factors for a point null hypothesis ($H_0 : d = 0$ vs. $H_1 : d > 0$) and for a one-sided interval null hypothesis ($H_0 : d \leq 0.2$ vs. $H_1 : d > 0.2$), denoted BF_{10}^{point} and $BF_{10}^{\text{interval}}$, respectively. In addition, we report the

[7] The computation of standardized effect sizes is more complex than often assumed. There are a range of mean difference (*d*) and variance explained (η^2) effect sizes that are often conflated and which are difficult to compare since they vary as a function of design, number of items, etc. Further, no widely agreed way of computing standardized effect sizes for mixed-effects models exists. We adopt the approach outlined in Judd et al. (2017), where we compute *d* using the square root of all the variance terms in the model (with the random slopes multiplied by our sum deviation coding) as the pooled standard deviation (sigma). This can be considered equivalent to classic Cohen's *d* for a between-subjects design. We choose this method as it is the most conservative, given the experimental design, and it affords direct comparison with the (few) interpretations of effect size in moral psychology (see Landy & Goodwin, 2015; Schnall et al., 2015).

posterior probabilities summarizing the posterior distribution of d, namely P($d > 0$ | D, M), P($d \in [-0.2, 0.2]$ | D, M), and P($d > 0.2$ | D, M), computed from the corresponding Bayesian models. When describing a Bayes Factor as providing support for a hypothesis, we mean that the evidence supports this hypothesis over the competing hypothesis (i.e., the point null or a one-sided null interval). Finally, when describing a posterior probability as providing support for the existence or magnitude of an effect, we mean that, given the data and model (| D, M), the effect is probably in the correct direction ($d > 0$), practically equivalent to zero ($d \in [-0.2, 0.2]$), or greater than our SESOI ($d > 0.2$). For full details of the LMMs in the classical (frequentist) and Bayesian frameworks, see "Main Analysis (LMMs)" and "Main Analysis (Bayesian LMMs)", respectively, in Supplementary Materials at https://osf.io/csxzt/?view_only=b05e41155b3849b586630ada963acb46.

Manipulation checks

Intent. We found a large-sized, significant effect of the intent manipulation on judgments of the agent's intention, $b = 3.49$, 95% CI [3.23, 3.75], $p < .001$, $d = 1.95$, 95% CI [1.60, 2.31]. As expected, intentional actions were seen as more intentional than accidental actions (see Figure 4A). All other fixed effects were not significant, ps > .09. The TOST indicated support only for the null effect of the domain manipulation ($p = .005$; see Figure 5A).

Domain. We found a medium-sized, significant effect of the domain manipulation on judgments of the act's impurity, $b = -1.19$, 95% CI [−0.99, −1.38], $p < .001$, $d = 0.63$, 95% CI [0.50, 0.76]. As expected, impure acts were seen as more impure than harmful acts. Intentional acts were also seen as more impure than accidental acts, $b = 1.60$, 95% CI [1.43, 1.77], $p < .001$, $d = 0.85$, 95% CI [0.71, 0.99]. This finding mirrors Kupfer et al.'s (2020) finding that intentional acts were judged more disgusting (their proxy term for impure) than

accidental ones. This effect of intention varied as a function of domain, $b = 0.37$, 95% CI [0.14, 0.60], $p = .002$, $d = 0.2$, 95% CI [0.07, 0.32] (see Figure 4B). The effect was greater for harmful ($b = 1.79$, 95% CI [1.49, 2.08], $p < 0.001$, $d = 0.95$, 95% CI [0.78, 1.11]) than impure ($b = 1.42$, 95% CI [1.18, 1.66], $p < 0.001$, $d = 0.75$, 95% CI [0.62, 0.88]) acts. The TOST failed to support the null for any of the effects (ps > .47; see Figure 5B).

In keeping with Kupfer et al. (2020), judgments of how harmful acts were did not differ significantly as a function of domain, $b = 0.02$, 95% CI [−0.17, 0.21], $p = .81$, $d = 0.01$, 95% CI [−0.09, 0.12]. As per Kupfer et al., intentional acts were judged as more harmful than accidental acts, $b = 1.00$, 95% CI [0.88, 1.11], $p < .001$, $d = 0.56$, 95% CI [0.45, 0.67]. The effect of the intent manipulation was significantly larger for harmful vs. impure acts, $b = 0.15$, 95% CI [0.01, 0.28], $p = .04$, $d = 0.08$, 95% CI [0.004, 0.16] (see Figure 4C). However, the TOST supported the null effect for both the interaction and domain main effect ($p = .001$ and $p < .001$, respectively; see Figure 5C).

Weirdness. We found a small-to-medium-sized significant effect of the domain manipulation on judgments of the act's weirdness, $b = −0.72$, 95% CI [−0.90, −0.55], $p < .001$, $d = 0.40$, 95% CI [0.29, 0.52]. As suggested by Gray & Keeney (2015), impure acts were seen as more weird than harmful acts (see Figure 4D). We found a large-sized effect of intention on weirdness, with intentional acts seen as more weird than accidental acts, $b = 1.68$, 95% CI [1.49, 1.86], $p < .001$, $d = 0.94$, 95% CI [0.76, 1.12]. The interaction effect was not significant, $p > .07$. The TOST failed to support the null effect for the interaction ($p = .084$; see Figure 5D).

The manipulation checks indicate that we successfully manipulated intent. Our manipulation of the moral domain is more nuanced. Impure acts were judged as more impure than harmful acts. However, consistent with Kupfer et al.'s (2020) results, harmful and

impure acts were judged equally harmful. This may be because many of the harmful acts employed both here and in the original experiment (e.g., cutting in line) focus on psychological "harm" that seems intuitively equivalent to the aversive nature of many of the impure acts (e.g., showing dinners a festering wound). Consistent with Gray & Keeney (2015), our discriminant manipulation check indicated that impure acts were judged as weirder than harmful acts. Surprisingly, the act's intent had a greater effect on judgments of weirdness. Taken together, the data suggest that we successfully manipulated intent and moral domain, at least in an equivalent manner to Kupfer et al.

Moral judgment

Basic model. Consistent with Kupfer et al.'s (2020) results, a large-sized, significant effect indicated that intentional acts were seen as more morally wrong than accidental acts, $b = 2.14$, 95% CI [1.93, 2.35], $p < .001$, $d = 1.18$, 95% CI [0.96, 1.40]. In addition, impure acts were judged as more morally wrong than harmful acts, $b = -0.18$, 95% CI $[-0.33, 0.02]$, $p = .03$, $d = 0.10$, 95% CI [0.01, 0.19]. Although Kupfer et al. did not find a significant effect of domain, their estimate was comparable.

Results revealed a small-sized, significant intent × domain effect, $b = 0.49$, 95% CI [0.25, 0.74], $p < .001$, $d = 0.27$, 95% CI [0.13, 0.41]; $BF_{10}^{\text{point}} = 49.80$, $BF_{10}^{\text{interval}} = 1.63$, $P(d > 0 \mid D, M) > .999$, $P(d \in [-0.2, 0.2] \mid D, M) = .16$, and $P(d > 0.2 \mid D, M) = .84$. The effect of intent was greater for harmful ($b = 1.07$, 95% CI [0.84, 1.27], $p < 0.001$, $d = 1.31$, 95% CI [1.06, 1.57]) than impure ($b = 0.92$, 95% CI [0.78, 1.07], $p < 0.001$, $d = 1.04$, 95% CI [0.83, 1.26]) acts (see Figure 6, A). It should be noted that although the Bayes Factor against the point null provided strong evidence in favor of the intent × domain effect, the Bayes Factor against the null interval indicates that the data are indecisive. The TOST failed to support practical equivalence for the intent × domain interaction ($p = .851$), but we could accept it for

the main effect of moral domain (p = .01; see Figure 7A). Furthermore, prior sensitivity checks revealed that the Bayes Factor against the null interval became decisive when using both principled ($BF_{10}^{\text{interval}}$ = 3.39) or informative ($BF_{10}^{\text{interval}}$ = 3.02) priors (for details, see "Prior Sensitivity Checks" in Supplementary Materials at https://osf.io/csxzt/?view_only=b05e41155b3849b586630ada963acb46).

Adjusting for weirdness. Since we found a significant effect of our manipulations on the perceived weirdness of the act, we fitted a model attempting to adjust for weirdness. Originally, we had planned to add only the weirdness (centered) and the intent × weirdness interaction to the basic model. However, we found that intent had a bigger effect on weirdness than domain. Therefore, our adjusted model included both the intent × weirdness and domain × weirdness interactions (for a discussion of adjustment for interaction effects, see Yzerbyt et al., 2004). Results revealed a small-sized, significant intent × domain effect, b = 0.39, 95% CI [0.21, 0.56], $p < .001$, d = 0.26, 95% CI [0.14, 0.39]; BF_{10}^{point}= 121.40, $BF_{10}^{\text{interval}}$ = 0.49, P($d > 0$ | D, M) > .999, P($d \in [-0.2, 0.2]$ | D, M) = .40, and P($d > 0.2$ | D, M) = .60 (see Figure 6B). As with the basic model, the Bayes Factor against the null interval indicated that the data are indecisive. The TOST failed to support practical equivalence for the intent × domain effect (p = .848), but we could accept it for the domain main effect, and for the intent × weirdness and domain × weirdness interaction effects (p = .003, $p < .001$, and $p < .001$, respectively; see Figure 7B). However, prior sensitivity checks revealed that the BF against the null interval remained indecisive when using both principled ($BF_{10}^{\text{interval}}$ = 1.03) or informative ($BF_{10}^{\text{interval}}$ = 0.87) priors (for details, see "Prior Sensitivity Checks" in Supplementary Materials at https://osf.io/csxzt/?view_only=b05e41155b3849b586630ada963acb46).

Robustness checks. We fitted CLMMs for the basic and adjusted models with probit links[8]. The basic CLMM showed a larger, significant intent × domain effect, d_{latent} = 0.39, 95% CI [0.18, 0.60], $p < .001$; BF_{10}^{point}= 77.52, $BF_{10}^{\text{interval}}$ = 6.77, P($d > 0$ | D, M) = .999, P($d \in [-0.2, 0.2]$ | D, M) = .05, and P($d > 0.2$ | D, M) = .95. The pattern of results with the adjusted CLMM was similar, d_{latent} = 0.36, 95% CI [0.19, 0.53], $p < .001$; BF_{10}^{point} = 342.71, $BF_{10}^{\text{interval}}$ = 8.44, P($d > 0$ | D, M) > .999, P($d \in [-0.2, 0.2]$ | D, M) = .04, and P($d > 0.2$ | D, M) = .96. The TOST failed to support practical equivalence for the intent × domain effect, p = .96. Compared to our LMMs, the CLMMs showed improved model checks and better model comparison, and stronger evidence for the intent × domain effect (see "Robustness Checks" in Supplementary Materials at https://osf.io/csxzt/?view_only=b05e41155b3849b586630ada963acb46). It appears that the CLMMs provide a better description of the data, with their estimates being more appropriate than those of the LMMs (Liddell & Kruschke, 2018).

The results from both LMMs and CLMMs in the frequentist framework supported the existence of the intent × domain effect and its non-equivalence to our SESOI (d = 0.2). Bayesian LMMs supported the existence of the effect, but posterior probabilities and BFs provided weaker evidence for its non-equivalence with our ROPE/SESOI and non-positive or at most trivially positive magnitude, respectively. However, prior sensitivity analysis suggested that with principled or informative priors, the BF against the null interval for the basic, but not adjusted, model provided support against a non-positive or trivially positive effect. Our robustness checks supported and strengthened these inferences with the BFs

[8] Using probit links for the CLMM means that the estimates can be interpreted as standardized effect sizes on the underlying continuous latent construct (Cohen's d_{latent}).

against the null interval for both the basic and adjusted models providing support against a non-positive or trivially positive intent × domain effect.

Taken together, the results of our close replication of Kupfer et al. (2020; Experiment 1) support the reduced-magnitude hypothesis, which holds that the intent × domain effect is a real effect but is reduced in magnitude when statistically adjusting for and/or employing stimuli that control for the context, target, severity, and weirdness of the act. Our failure to find support for the null hypothesis (i.e., the confounded-stimulus-category hypothesis) suggests that the intent × domain effect is not merely a methodological artifact arising from stimulus-category confounds.

Model and Robustness Checks for Kupfer et al. (2020; Experiment 1)

The model checks for LMMs in our replication were consistent with possible (related) threats to valid statistical inference arising from modeling ordinal items as interval data and from CFEs. Therefore, we conducted the same model and robustness checks (i.e., fitting frequentist and Bayesian CLMMs with probit links) as above on the data for the (basic) model fitted by Kupfer et al. (2020). Figure 8 shows the model checks. As in our replication, residual plots revealed some deviations from normality and homoscedasticity, and posterior predictive checks showed the observed data bunching at the bounds (0 and 6). In addition, influence statistics revealed three influential cases. These model checks suggest that Kupfer et al. (2020) faced similar threats to valid statistical inference (i.e., modeling ordinal items as interval data and CFEs) to those we encountered in our replication.

CLMMs fitted to Kupfer et al.'s (2020; Experiment 1) data revealed a significant intent × domain effect, d_{latent} = 0.48, 95% CI [0.02, 0.94], p = .04; BF_{10}^{point} = 1.86, $BF_{10}^{\text{interval}}$ = 1.73, P(d > 0 | D, M) = .94, P($d \in [-0.2, 0.2]$ | D, M) = .18, and P(d > 0.2 | D, M) = .81. The TOST failed to support practical equivalence for the intent × domain effect, p = .88.

From a frequentist perspective, unlike the LMM, the CLMM supported the existence of the intent × domain effect in the original Kupfer et al. study. From a Bayesian perspective, the posterior probabilities support both the existence of the effect and that its magnitude exceeds our SESOI. However, the Bayes Factors indicated that the data are indecisive. Prior sensitivity checks revealed that the Bayes Factor against the point null became decisive when using informative (BF_{10}^{point} = 3.39), but not principled (BF_{10}^{point} = 2.11), priors. However, Bayes Factors remained indecisive against the null interval for both principled (BF_{10}^{point} = 2.07) and informative (BF_{10}^{point} = 1.99) priors (for details, see "Model and Robustness Checks for Kupfer et al." in Supplementary Materials at https://osf.io/csxzt/?view_only=b05e41155b3849b586630ada963acb46). Taken together, our (re)analyses of Kupfer et al.'s (2020; Experiment 1) data suggest that the original findings may not be robust to violations of LMM assumptions. We interpret the model and robustness checks for the original Kupfer et al. study as indicating the existence of the intent × domain effect, thereby providing further support for the reduced-magnitude hypothesis.

Meta-Analyses and Replication Success

We performed a series of meta-analyses on the results from the replication and Kupfer et al. (2020). We specified fixed and random effects meta-analytic models in both classical (frequentist) and Bayesian frameworks. We present the classical fixed-effects models here. However, there is little difference between the pooled estimates from random- and fixed-effects models in either framework, although our Bayesian random-effects models recovered greater heterogeneity across studies (see "Meta-Analysis and Replication Success" in the Supplementary Materials at https://osf.io/csxzt/?view_only=b05e41155b3849b586630ada963acb46).

Meta-analyses revealed a significant small-sized, pooled estimate, $d = 0.28$, 95% CI [0.15, 0.40], $p < .001$; $BF_{10}^{\text{point}} = 142.98$, $BF_{10}^{\text{interval}} = 2.96$, P($d > 0$ | D, M) > .99, P($d \in [-0.2, 0.2]$ | D, M) = .12, and P($d > 0.2$ | D, M) = .88 (see Figure 9A). Given the data and model, pooling across studies provides evidence for the intent × domain effect, suggesting it is likely larger than our SESOI ($d = .20$). We also conducted a meta-analysis using the estimates from the robustness checks (CLMMs) fitted to the data from both studies. One could argue that these are more appropriate for meta-analysis. Results revealed a medium-sized (on the latent scale), pooled estimate, $d_{latent} = 0.41$, 95% CI [0.22, 0.60], $p < .001$; $BF_{10}^{\text{point}} = 539.01$, $BF_{10}^{\text{interval}} = 8.82$, P($d > 0$ | D, M) > .99, P($d \in [-0.2, 0.2]$ | D, M) = .02, and P($d > 0.2$ | D, M) = .98 (see Figure 9B). The meta-analyses offer support for the existence and practical significance of the intent × domain effect. It seems likely that the effect exceeds our SESOI. This further supports the reduced-magnitude hypothesis.

The effect size estimates across the original and replication studies were very similar. Indeed, the meta-analytic *Q*-test of the between-study heterogeneity was not significant. Thus, we failed to reject the null hypothesis of compatibility in effect sizes, $Q(1) = .02$, $p = .88$. TOSTs failed to support practical equivalence for the intent × domain effect in the original and replication studies, $p = .73$ and $p = .84$, respectively (see Figure 9A). It is challenging to define replication "success" when the original study is nonsignificant. However, one can interpret the results of the TOSTs as suggesting that we have successfully "replicated" (there was no test of support for a null region in the original study) a failure to find support for a (null) region of practical equivalence of $d \pm 0.2$. Of course, we cannot infer compatibility of effect sizes in a NHST framework from failing to establish incompatibility. However, the posterior probability that the difference between the random intercepts was within our ROPE provided positive evidence that the effect sizes for the original and replication studies were compatible, P($d \in [-0.2, 0.2]$ | D, M) = .95.

General Discussion

The intent × domain effect provides converging behavioral, neural, and cross-cultural evidence that our moral mind/brain may be composed of distinct, domain-specific cognitive systems. Recent null results suggested that a confounded (by severity, weirdness, context, and target of the act) stimulus category may explain the effect (Kupfer et al., 2020; Parkinson & Byrne, 2017). Other evidence suggested that controlling or adjusting for potential confounds may reduce, but not eliminate, the effect (Chakroff et al., 2013). Closely replicating Kupfer et al., we tested the confounded-stimulus-category and reduced-magnitude hypotheses by fitting LMMs in both frequentist and Bayesian frameworks, employing a wide range of statistical indices. We found evidence for a small-to-medium-sized effect, supporting the reduced-magnitude hypothesis. We failed to find support for the confounded-stimulus-category (null) hypothesis. Specifically, we did not find evidence that the effect was equivalent to our SESOI or region of practical equivalence, $d \in [-0.2, 0.2]$. Results were consistent when attempting to adjust for the act's weirdness. MFT was partly motivated by the notion that “harmless” acts can still be perceived as morally wrong (Graham et al., 2012; Haidt & Joseph, 2004). Therefore, finding an intent × domain effect when impure and harmful acts were perceived as equally harmful (and impure acts more impure) might be considered a conservative test of the reduced-magnitude hypothesis and a liberal test of the confounded-stimulus-category hypotheses.

Model checks revealed potential threats to valid statistical inference (e.g., ceiling and floor effects, modeling ordinal data as interval, and deviations from normality and homoscedasticity). Therefore, we fitted CLMMs to test the robustness of our findings. Results from these models provided increased support for the reduced-magnitude hypothesis. Moreover, model checks for a LMM fitted to the data from Kupfer et al. suggested similar threats to valid statistical inference. Indeed, fitting a CLMM (a more appropriate model for

ordinal data with potential ceiling and floor effects) to the original study's data yielded a small, statistically significant intent × domain effect. Meta-analyses of the replication and original study revealed a significant, small (d = 0.28, 95% CI [0.15, 0.40], with LMM estimates) to medium-sized (d_{latent} = 0.41, 95% CI [0.22, 0.60], with CLMM estimates) effect. Although the original and replication studies differed in statistical significance, they produced consistent, practically equivalent effect-size estimates. The main difference lies in the much greater power and precision of the replication, which is itself explained by a greater number of levels in the by-scenario random factor.

Taken together, our results support the reduced-magnitude hypothesis. That is, the intent × domain effect is real, but its magnitude is reduced when controlling and/or adjusting for the context, severity, weirdness, and target of the act. Our study helps explain the "inconsistent" findings regarding the role of intent across different moral domains. In doing so, it highlights some crucial lessons for valid statistical and scientific inference in experimental social psychology.

Valid Statistical Inference in Experimental Social Psychology

In this section, we outline differences between our approach to statistical and scientific inference and that commonly used in experimental social psychology, including Kupfer et al.'s (2020) original study. To foreshadow, we discuss the importance of sample-size simulations, testing support for the null, effect size, precision, model checks, and (robust) ordinal regression for Likert-type responses.

Simulations to Determine Sample Size for LMMs

The difference between the (classical) statistical inference reported in Kupfer et al.'s (2020) original study and that reported in our replication primarily stems from the respective statistical power of the two studies. Kupfer et al. did not report a sample size calculation for

the LMM used to test their hypothesis in Experiment 1. Despite a highly inflated Type I error rate and an inability to generalize findings beyond the sampled stimuli when simply averaging across items rather than treating them as a random factor (Judd et al., 2012), adopting LMMs in experimental social psychology is not yet (sadly) common practice. As adoption of LMMs increases, it is essential to recognize that treating a small number of scenarios or items as random factors can substantially limit power and precision (Judd et al., 2017; Westfall et al., 2014). We recommend that researchers fitting LMMs use sample-size simulations; our R code is provided in the Supplementary Materials (see also Brysbaert & Stevens, 2018). In cases where researchers cannot run simulations (which can be time-consuming and require some coding skill), a suggested rule of thumb from cognitive psychology is that random factors should have at least 20 levels (Singmann & Kellen, 2019). Given that average effect sizes in social psychology are smaller than those in cognitive psychology (Open Science Collaboration, 2015) and, arguably, by-item variability is greater, this number could easily be doubled or tripled for social psychology.

It may be surprising that fitting an LMM to Kupfer et al.'s study with N = 237 and a within-subjects design does not provide enough data to draw strong, unanimous inferences across classical and Bayesian frameworks and their associated statistical indices. Adequate power and precision, especially when planning to test support for the null hypothesis, require more data than might be expected (e.g., considering both the number of stimuli or items and the number of participants). The present study demonstrates that this may be especially important when making statistical and scientific inferences about small effects, which are the norm in experimental social psychology (Open Science Collaboration, 2015; Lovakov & Agadullina, 2021).

Effect Size, Precision, and Support for the Null

Kupfer et al. (2020) reported that the effect of intent did not vary as a function of moral domain, taking their nonsignificant interaction as support for the strong version of their null hypothesis of intent invariance across moral domains. Although a common practice, it is not valid to infer that there is "no difference" merely from a nonsignificant *p*-value. These errors can lead to the dismissal of crucial effects, undermining scientific progress and wasting public resources. This fact has led hundreds of scientists and statisticians to openly call for an end to NHST and the adoption of estimation statistics (i.e., effect sizes and precision) in frequentist or Bayesian frameworks (see Amrhein et al., 2019), with others calling for the use of Bayes Factors. Our approach included these alternatives to NHST, as well as equivalence testing, within the classical NHST framework (i.e., TOST). Specifying a meaningful SESOI is essential to testing support for the null via equivalence tests (Lakens, 2017). We are sure that for some our chosen SESOI ($d > 0.2$) might strike them as too small or too large. However, we believe that our chosen SESOI is a reasonable choice for two reasons. First, there is agreement among moral psychology researchers from diverging perspectives that meaningful effects should exceed $d = 0.11$ (Landy & Goodwin, 2015; Schnall et al., 2015). Second, empirically derived estimates suggest that small effect sizes in social psychology are around $d = 0.15$ (Lovakov & Agadullina, 2021). Researchers are welcome to disagree with us. However, we hope that researchers, editors, and reviewers will agree that the present study demonstrates the need to engage with, and require reporting of, SESOI, the precision of estimates, and testing support for the null hypothesis in experimental social psychology.

Model checks and (robust) ordinal regression models

Across the replication and original study model checks (i.e., posterior predictive checks) were consistent with two related threats to valid statistical inference: ceiling and floor effects (i.e., errors of measurement) and modelling ordinal data as interval data. Although modeling ordinal (i.e., Likert-style) responses as interval data is common practice in social

psychology, it can lead to inaccurate effect size estimates and inflated Type I and II error rates (Liddell & Kruschke, 2018). CLMMs offer a robust and flexible modeling approach for Likert-style measures and are more robust than LMMs to clustering at the bounds (see Bürkner & Vuorre, 2018; Harrell, 2015; Liddell & Kruschke, 2018). Given that fitting a CLMM to the original data from Kupfer et al. (2020) revealed a significant effect, we speculate that modeling ordinal responses as interval data may contribute to social psychology's lower replication rate compared to cognitive psychology (Open Science Collaboration, 2015), where using ordinal responses is less common. Future work should examine the extent of this threat to valid statistical inference.

Our robustness checks using CLMMs were not preregistered. However, preregistered analyses (across the basic and adjusted models) supported the reduced-magnitude hypothesis and failed to support the confounded-stimulus-category hypothesis. In hindsight, we should have carried out model checking, beyond examining convergence and overfitting, in LMMs fitted to data from Kupfer et al. (2020) before preregistration. We suggest that researchers involved in replication efforts conduct such model checks using the original study's data (if available) to inform preregistration.

As a segue to the next section, we want to make it clear that we did not begin our inquiries into the moral mind/brain with the intention of becoming methodologists advocating for improved methods and practices. When our inquiries began, we were unaware that the case for such efforts had been made convincingly long before the replication crisis (e.g., Meehl, 1978/2006; see Waller et al., 2006). Since then, many reforms (e.g., Munafò et al., 2017) have been implemented to varying degrees. However, even with the adoption of better methods and practices, psychologists must still confront what we see as the more complex challenge of developing explanatory theories of the mind/brain itself.

Developing Explanatory Theories of the Moral Mind/Brain

Our findings are not compatible with strongly modular (Fodor, 1985) or partially modular interpretations of the moral mind/brain (early explications of MFT could be read this way, see Haidt & Joseph, 2004) but remain compatible with weakly modular interpretations (e.g., Graham et al., 2012). The problem is that our findings could equally well be interpreted as consistent with a single moral-cognitive system (e.g., Schein & Gray, 2018) with domain-dependent modulation. It is easier to test for strong modularity than to distinguish between weakly modular and singular systems with domain-dependent modulation. Recent imaging work suggests that the computations underpinning moral judgments of acts reflecting harm, purity, fairness, authority, loyalty, and liberty involve both shared (across moral foundations) and unique activations (Hopp et al., 2023). However, the moral foundation vignettes employed suffer from the same stimuli-category confounds examined here (e.g., weirdness, severity). Indeed, Hopp et al. reported that the severity of moral judgments varied significantly across moral domains. Interestingly, the authors suggested that the shared activation for judgments across moral modules (vs. social norms) reflected a "domain-general" (ToM) network, which they interpreted as consistent with a weakly modular account. This shared network of brain regions is preferentially recruited across a broad range of studies on moral cognition (Eres et al., 2018), suggesting that it reflects general computations within the moral domain. Research has shown that activity in these brain regions maps tightly onto experimental manipulations of agent intent and the (causal) outcomes of harmful acts (Koster-Hale et al., 2017; Young & Saxe, 2009). This is consistent with formal models of the structured representations and principles that characterize the posited (causal, intentional, and moral) computations being carried out by the moral mind/brain in the case of harmful acts (Mikhail, 2007, 2009; see also, Levine et al., 2018).

In contrast, we have little idea about any (domain-specific) structured representations and principles governing moral judgments of impure acts and acts that engage other posited moral-cognitive modules. We suggest that MFT researchers and others (e.g., Curry et al., 2019) concerned with delineating and decomposing the moral mind/brain into multiple moral-cognitive modules would do well to focus on developing theoretical accounts of structured representations and principles for these moral domains. Such computational-representational properties are considered the "signal properties" of modularity (Barrett & Kurzban, 2006). This suggestion to focus on developing formal computational-representational models of the moral mind/brain echoes wider calls for reforms to theory development in the psychological sciences (e.g., Borsboom et al., 2021; Guest & Martin, 2021; Muthukrishna & Henrich, 2019; Oberauer & Lewandowsky, 2019; Rooij & Baggio, 2021).

Probabilistic models defined over structured representations have been usefully employed to formally model theories of core knowledge (Ullman & Tenenbaum, 2020) and language (Yang, 2004). Similarly, we believe that formal and computational modeling of theories of the moral mind/brain (see Crockett, 2016; Cushman, 2023) in conjunction with precise experimental, developmental, and comparative work offers researchers the best chance of successfully delineating and decomposing the moral mind/brain. For example, drift diffusion modeling of response time and moral judgment has already provided valuable evidence in favour of a single moral-cognitive process (see Baron & Gürçay, 2017; Gürçay & Baron, 2016) over a (default-interventionist) dual process theory of moral judgment (Greene, 2007; Greene et al., 2001). While this work focused on distinguishing cognitive operations as a function of the number and temporal order of processing styles (i.e., automatic vs controlled) rather than the number of functionally-specialized cognitive modules, similar approaches could be applied to examining whether the moral mind/brain is constituted by

multiple moral-cognitive modules (vs. a singular system with domain-dependent modulation).

In a recent review of the purity moral domain, Gray et al. (2023) concluded that there is insufficient evidence for a purity-based cognitive module. Indeed, the authors concluded that purity is not a coherent psychological construct. This resonates with broader concerns (we share) about measurement in psychology (Flake et al., 2017; Flake & Fried, 2020) and more specific discussions on the role of technical definitions of morality (see Dahl, 2023). Ultimately, we believe that morality is whatever our best scientific theory of morality says it is. That said, definitions of the phenomenon of interest can serve as useful placeholders, alongside other aspects of the meta-theoretical frameworks (e.g., Marr, 1982) within which scientific theories and models are developed (see Rooij & Baggio, 2021). Focusing on formal and computational models and meta-theoretical frameworks helps us address the toothbrush problem: in psychology, theories are like toothbrushes; nobody wants to use anyone else's (Mischel, 2008). We cannot think of many cases in which competing predictions derived from different models or theories of the moral mind/brain have been experimentally tested in rigorous, valid, and sustained ways. By making predictions precise, formal models and theories make it easier and more worthwhile for researchers (with different a priori perspectives) to engage with one another's models and theories. We do worry that a simplistic focus on definitions as the "explanation" for all disagreements among researchers in the field risks isolating research programs.

We urge researchers in the field to adopt the statistical methods and practices advocated here (e.g., mixed-effects models, sample-size simulations, testing the null, reporting effect size and precision, model checks, and (robust) ordinal regression for Likert-type responses) in order to identify a set of reliable and robust causal effects that can be accounted for by explanatory theories (Sweetman & Newman, 2020). That said, not every

causal effect on moral judgment is equally in need of explanation. As Rooij & Baggio (2021) aptly put it, “trying to build theories on collections of effects is much like trying to write novels by collecting sentences from randomly generated letter strings. Indeed, each novel ultimately consists of strings of letters, and theories should ultimately be compatible with effects. Still, the majority of the (infinitely possible) effects are irrelevant for the aims of theory building, just as the majority of (infinitely possible) sentences are irrelevant for writing a novel.” Without the development of (ideally formal) explanatory theories and models within (meta-theoretical) frameworks, we risk expending an inordinate amount of effort for little explanatory progress.

Open Practices

This experiment in this article has received the badges for Open Data, Open Materials, and Preregistration. Materials, data and all scripts are available at https://osf.io/csxzt/?view_only=b05e41155b3849b586630ada963acb46. The preregistration is available at https://osf.io/3f6c4/?view_only=b1b1d11dd00c4e6bb60054b0ee2ca20a. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

References

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, *567*(7748), 305–307. https://doi.org/10.1038/d41586-019-00857-9

Baron, J., & Gürçay, B. (2017). A meta-analysis of response-time tests of the sequential two-systems model of moral judgment. *Memory & Cognition*, *45*(4), 566–575. https://doi.org/10.3758/s13421-016-0686-8

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Barrett, H. C., Bolyanatz, A., Crittenden, A. N., Fessler, D. M. T., Fitzpatrick, S., Gurven, M., Henrich, J., Kanovsky, M., Kushnick, G., Pisor, A., Scelza, B. A., Stich, S., Rueden, C. von, Zhao, W., & Laurence, S. (2016). Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment. *Proceedings of the National Academy of Sciences*, *113*(17), 4688–4693. https://doi.org/10.1073/pnas.1522070113

Barrett, H. C., & Kurzban, R. (2006). Modularity in cognition: Framing the debate. *Psychological Review*, *113*(3), 628–647. https://doi.org/10.1037/0033-295x.113.3.628

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Bechtel, W. (2012). *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. Psychology Press. (New York). https://doi.org/10.4324/9780203810095

Borsboom, D., Maas, H. L. J. van der, Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science*, *16*(4), 756–766. https://doi.org/10.1177/1745691620969647

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*(1), 1052–21. https://doi.org/10.5334/joc.10

Bürkner, P.-C. (2017). brms : An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1). https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C., & Vuorre, M. (2018). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, *2*(1), 77–101. https://doi.org/10.1177/2515245918823199

Chakroff, A., Dungan, J., Koster-Hale, J., Brown, A., Saxe, R., & Young, L. (2016). When minds matter for moral judgment: intent information is neurally encoded for harmful but

not impure acts. *Social Cognitive and Affective Neuroscience*, *11*(3), 476–484. https://doi.org/10.1093/scan/nsv131

Chakroff, A., Dungan, J., & Young, L. (2013). Harming ourselves and defiling others: What determines a moral domain? *PLoS ONE*, *8*(9), e74434-12. https://doi.org/10.1371/journal.pone.0074434

Christensen, R. H. B. (2020). Cumulative link models for ordinal regression with the R package ordinal. *Comprehensive R Archive Network*. https://cran.r-project.org/web/packages/ordinal/vignettes/clm_article.pdf

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716–aac4716. https://doi.org/10.1126/science.aac4716

Crockett, M. J. (2016). How formal models can illuminate mechanisms of moral judgment and decision making. *Current Directions in Psychological Science*, *25*, 85–90. https://doi.org/10.1177/0963721415624012

Curry, O. S., Chesters, M. J., & Lissa, C. J. V. (2019). Mapping morality with a compass: Testing the theory of "morality-as-cooperation" with a new questionnaire. *Journal of Research in Personality*, *78*, 106–124. https://doi.org/10.1016/j.jrp.2018.10.008

Cushman, F. (2023). Computational social psychology. *Annual Review of Psychology*, *75*(1), 625–652. https://doi.org/10.1146/annurev-psych-021323-040420

Dahl, A. (2023). What we do when we define morality (and why we need to do it). *Psychological Inquiry*, *34*(2), 53–79. https://doi.org/10.1080/1047840x.2023.2248854

Dungan, J. A., & Young, L. (2019). Asking 'why?' enhances theory of mind when evaluating harm but not purity violations. *Social Cognitive and Affective Neuroscience*, *14*(7), 699–708. https://doi.org/10.1093/scan/nsz048

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research. *Social Psychological and Personality Science*, *8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Flora, D. B. (2020). Your coefficient alpha is probably wrong, but which coefficient omega Is right? A tutorial on using R to obtain better reliability estimates. *Advances in Methods and Practices in Psychological Science*, *3*(4), 484–501. https://doi.org/10.1177/2515245920951747

Fodor, J. A. (1985). Précis of the modularity of mind. *Behavioral and Brain Sciences*, *8*, 1–42. https://doi.org/10.1017/s0140525x0001921x

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, *116*(2), 439–453. https://doi.org/10.1037/a0015251

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S., & Ditto, P. (2012). Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology*, *47*, 55–130. https://doi.org/10.1016/b978-0-12-407236-7.00002-4

Gray, K., DiMaggio, N., Schein, C., & Kachanoff, F. (2023). The problem of purity in moral psychology. *Personality and Social Psychology Review*, *27*(3), 272–308. https://doi.org/10.1177/10888683221124741

Gray, K., & Keeney, J. E. (2015). Impure or just weird? Scenario sampling bias raises questions about the foundation of morality. *Social Psychological and Personality Science*, *6*(8), 859–868. https://doi.org/10.1177/1948550615592241

Greene, J. D. (2007). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.), *Moral Psychology, Vol. 3: The Neuroscience of Morality: Emotion, Disease, and Development* (pp. 35–79). MIT Press. https://doi.org/10.7551/mitpress/7504.003.0004

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105–2108. https://doi.org/10.1126/science.1062872

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*(4), 789–802. https://doi.org/10.1177/1745691620970585

Gürçay, B., & Baron, J. (2016). Challenges for the sequential two-system model of moral judgement. *Taylor & Francis*, 1–33. https://doi.org/10.1080/13546783.2016.1216011

Haidt, J., & Joseph, C. (2004). Intuitive ethics: how innately prepared intuitions generate culturally variable virtues. *Daedalus*, *133*(4), 55–66. https://doi.org/10.1162/0011526042365555

Harrell. (2015). *Regression modeling strategies with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer. https://doi.org/10.1007/978-3-319-19425-7

Hopp, F. R., Amir, O., Fisher, J. T., Grafton, S., Sinnott-Armstrong, W., & Weber, R. (2023). Moral foundations elicit shared and dissociable cortical activation modulated by political ideology. *Nature Human Behaviour*, *7*(12), 2182–2198. https://doi.org/10.1038/s41562-023-01693-8

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69. https://doi.org/10.1037/a0028347

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology*, *68*(1), 601–625. https://doi.org/10.1146/annurev-psych-122414-033702

Koster-Hale, J., Richardson, H., Velez, N., Asaba, M., Young, L., & Saxe, R. (2017). Mentalizing regions represent distributed, continuous, and abstract dimensions of others' beliefs. *NeuroImage*, *161*, 9–18. https://doi.org/10.1016/j.neuroimage.2017.08.026

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270–280. https://doi.org/10.1177/2515245918771304

Kupfer, T. R., Inbar, Y., & Tybur, J. M. (2020). Reexamining the role of intent in moral judgements of purity violations. *Journal of Experimental Social Psychology*, *91*, 104043. https://doi.org/10.1016/j.jesp.2020.104043

Lakens, D. (2017). Equivalence tests. *Social Psychological and Personality Science*, *8*(4), 355–362. https://doi.org/10.1177/1948550617697177

Landy, J. F., & Goodwin, G. P. (2015). Does incidental disgust amplify moral judgment? A meta-analytic review of experimental evidence. *Perspectives on Psychological Science*, *10*(4), 518–536. https://doi.org/10.1177/1745691615583128

Levine, S., Leslie, A. M., & Mikhail, J. (2018). The mental representation of human action. *Cognitive Science*, *42*(4), 1229–1264. https://doi.org/10.1111/cogs.12608

Lewontin, R. (1998). The evolution of cognition: Questions we will never answer. In D. N. Osherson, D. Scarborough, & S. Sternberg (Eds.), *An invitation to cognitive science* (Vol. 4, pp. 107–132). An invitation to cognitive science. (Cambridge; MA). https://doi.org/10.7551/mitpress/3967.001.0001

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328–348. https://doi.org/10.1016/j.jesp.2018.08.009

Lovakov, A., & Agadullina, E. R. (2021). Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology*, *51*(3), 485–504. https://doi.org/10.1002/ejsp.2752

Lüdecke, D., Ben-Shachar, M., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, *6*(60), 3139. https://doi.org/10.21105/joss.03139

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press. https://doi.org/10.7551/mitpress/9780262514620.001.0001

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

McGuire, J., Langdon, R., Coltheart, M., & Mackenzie, C. (2009). A reanalysis of the personal/impersonal distinction in moral psychology research. *Journal of Experimental Social Psychology*, *45*(3), 577–580. https://doi.org/10.1016/j.jesp.2009.01.002

Mikhail, J. (2007). Universal moral grammar: theory, evidence and the future. *Trends in Cognitive Sciences*, *11*(4), 143–152. https://doi.org/10.1016/j.tics.2006.12.007

Mikhail, J. (2009). *Elements of moral cognition*. Cambridge University Press. https://doi.org/10.1017/cbo9780511780578

Mischel, W. (2008). The toothbrush problem. *APA Observer*. https://www.psychologicalscience.org/observer/the-toothbrush-problem

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*(3), 221–229. https://doi.org/10.1038/s41562-018-0522-1

Nicenboim, B., & Vasishth, S. (2016). Statistical methods for linguistic research: Foundational ideas - Part II. *Language and Linguistics Compass*, *10*(11), 591–613. https://doi.org/10.1111/lnc3.12207

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, *26*(5), 1596–1618. https://doi.org/10.3758/s13423-019-01645-2

Parkinson, M., & Byrne, R. M. J. (2017). Judgments of moral responsibility and wrongness for intentional and accidental harm and purity violations. *The Quarterly Journal of Experimental Psychology*, *0*(0), 1–12. https://doi.org/10.1080/17470218.2016.1276942

Rooij, I. van, & Baggio, G. (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science*, *16*(4), 682–697. https://doi.org/10.1177/1745691620970604

Schein, C., & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review*, *22*(1), 32–70. https://doi.org/10.1177/1088868317698288

Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2015). Landy and Goodwin (2015) confirmed most of our findings then drew the wrong conclusions. *Perspectives on Psychological Science*, *10*(4), 537–538. https://doi.org/10.1177/1745691615589078

Šimkovic, M., & Träuble, B. (2019). Robustness of statistical methods when measure is affected by ceiling and/or floor effect. *PLoS ONE*, *14*(8), e0220889. https://doi.org/10.1371/journal.pone.0220889

Singmann, H., & Kellen, D. (2019). An introduction to mixed models for experimental psychology. In D. Spieler & E. Schumacher (Eds.), *New methods in cognitive psychology* (pp. 4–31). Psychology Press. (New York). http://singmann.org/download/publications/singmann_kellen-introduction-mixed-models.pdf

Sweetman, J., & Newman, G. A. (2020). Replicating different roles of intent across moral domains. *Royal Society Open Science*, *7*(5), 1–7. https://doi.org/10.1098/rsos.190808

Team, R. C. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Uhlmann, E. L., & Zhu, L. L. (2014). Acts, persons, and intuitions: Person-centered cues and gut reactions to harmless transgressions. *Social Psychological and Personality Science*, *5*(3), 279–285. https://doi.org/10.1177/1948550613497238

Ullman, T. D., & Tenenbaum, J. B. (2020). Bayesian models of conceptual development: Learning as building models of the world. *Annual Review of Developmental Psychology*. https://doi.org/10.1146/annurev-devpsych-121318-084833

Vasishth, S., & Nicenboim, B. (2016). Statistical methods for linguistic research: Foundational ideas - Part I. *Language and Linguistics Compass*, *10*(8), 349–369. https://doi.org/10.1111/lnc3.12201

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2017). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, 1–23. https://doi.org/10.3758/s13423-017-1343-3

Westfall, J., Judd, C. M., & Kenny, D. A. (2015). Replicating studies in which samples of participants respond to samples of stimuli. *Perspectives on Psychological Science*, *10*, 390–399. https://doi.org/10.1177/1745691614564879

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*, 2020–2045. https://doi.org/10.1037/xge0000014

Yang, C. D. (2004). Universal Grammar, statistics or both? *Trends in Cognitive Sciences*, *8*(10), 451–456. https://doi.org/10.1016/j.tics.2004.08.006

Yonce, N. G., Grove, L. J. [Ed];, Faust, W. M. [Ed];, Lenzenweger, D. [Ed];, & [Ed], M. F. (Eds.). (2006). *A Paul Meehl Reader, Essays on the Practice of Scientific Psychology*. Lawrence Erlbaum Associates Publishers. https://doi.org/10.4324/9780203759554

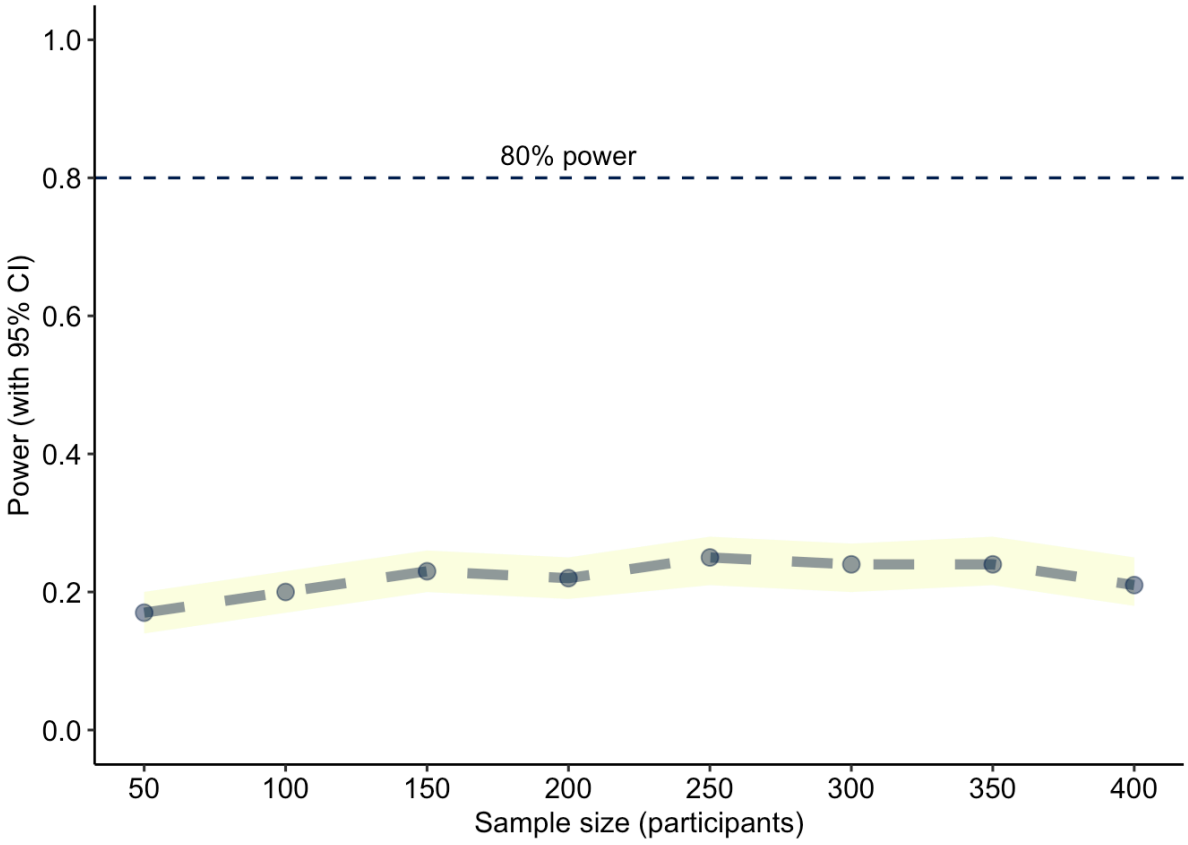
Young, L., & Saxe, R. (2009). An fMRI investigation of spontaneous mental state inference for moral judgment. *Journal of Cognitive Neuroscience*, *21*(7), 1396–1405. https://doi.org/10.1162/jocn.2009.21137

Young, L., & Saxe, R. (2011). When ignorance is no excuse: Different roles for intent across moral domains. *Cognition*, *120*(2), 202–214. https://doi.org/10.1016/j.cognition.2011.04.005

Yzerbyt, V. Y., Muller, D., & Judd, C. M. (2004). Adjusting researchers' approach to adjustment: On the use of covariates when testing interactions. *Journal of Experimental Social Psychology*, *40*(3), 424–431. https://doi.org/10.1016/j.jesp.2003.10.001

Figure 1

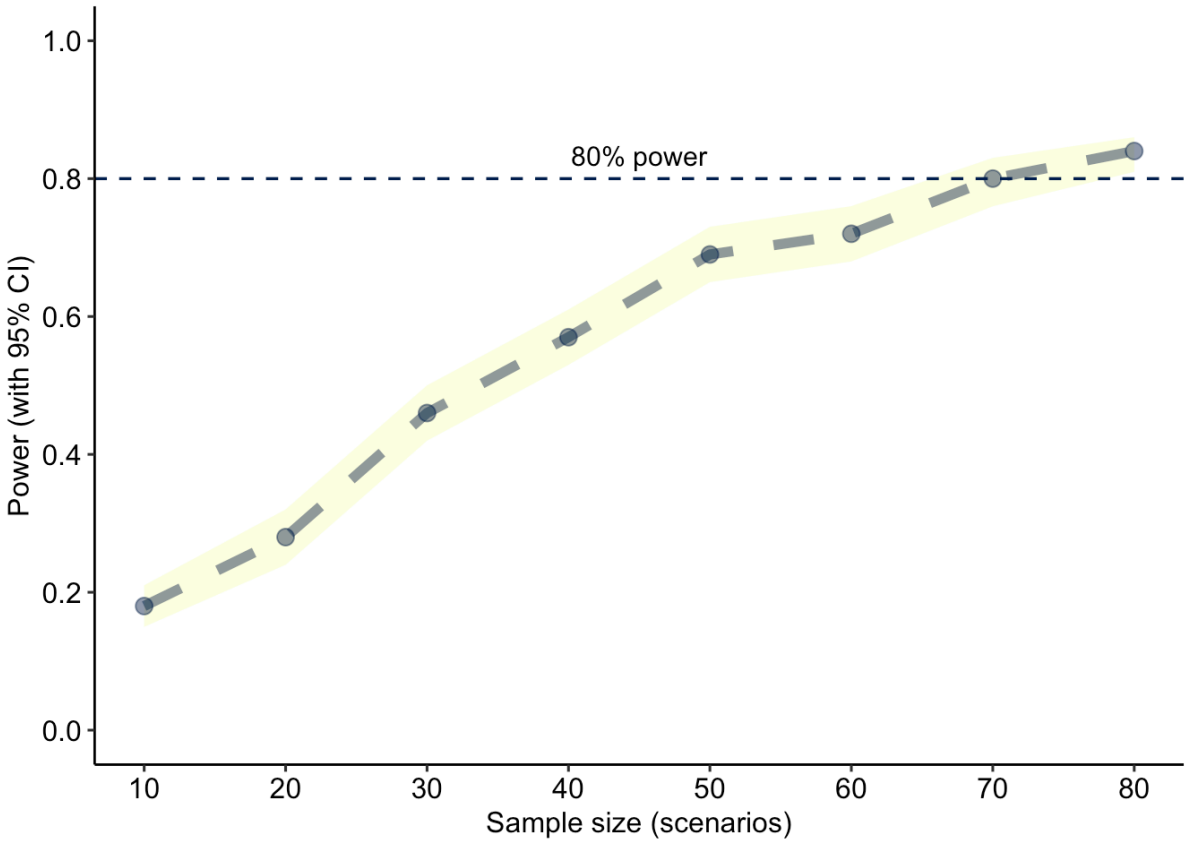
Power curve for intent × domain effect as a function of number of participants with the original 14 scenarios (Kupfer et al., 2020; Experiment 1)



Note. Error ribbon reflects 95% CIs.

Figure 2

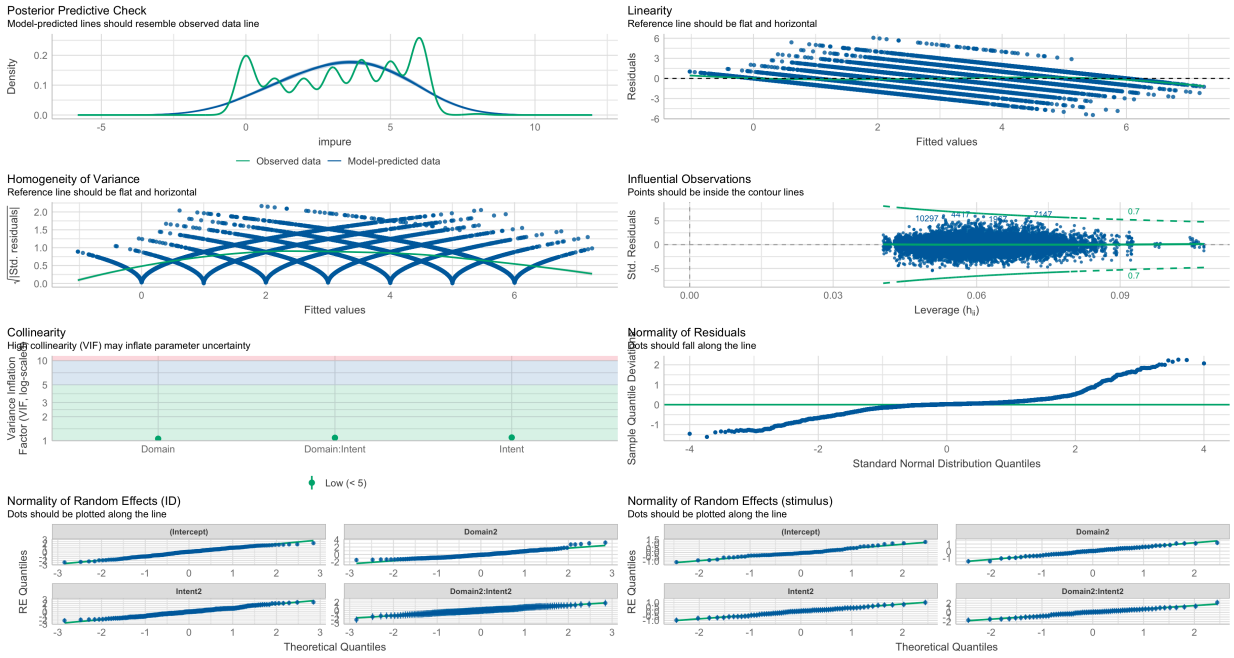
Power curve for intent × domain effect as a function of number of scenarios with the original 224 participants (Kupfer et al., 2020; Experiment 1).



Note. Error ribbon reflects 95% CIs.

Figure 3

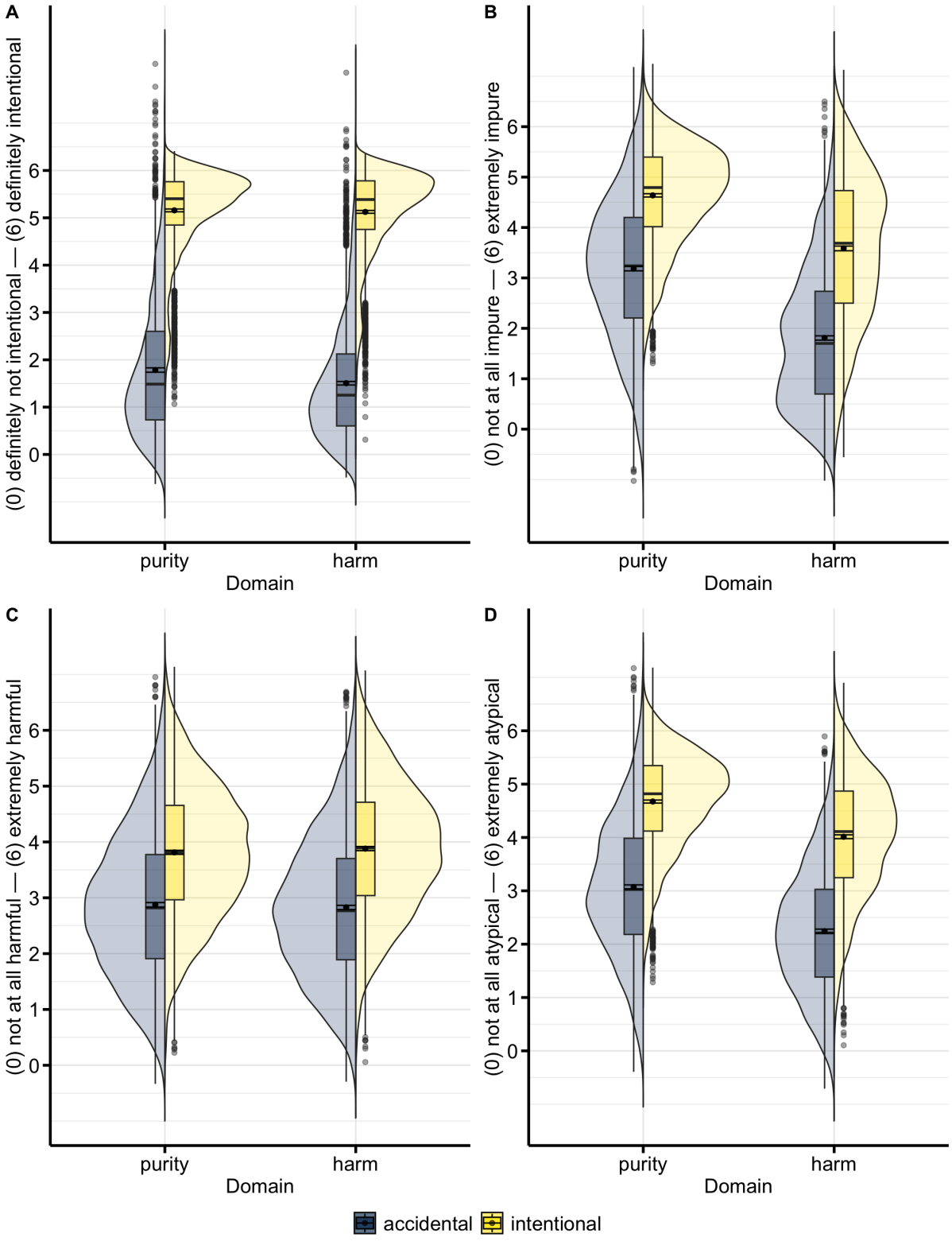
Visual model checks for the LMM intent manipulation.



Note. Shows visual model diagnostics for posterior predictive, linearity, homogeneity of variance, influential observations, collinearity, normality of residuals, and normality of random effects.

Figure 4

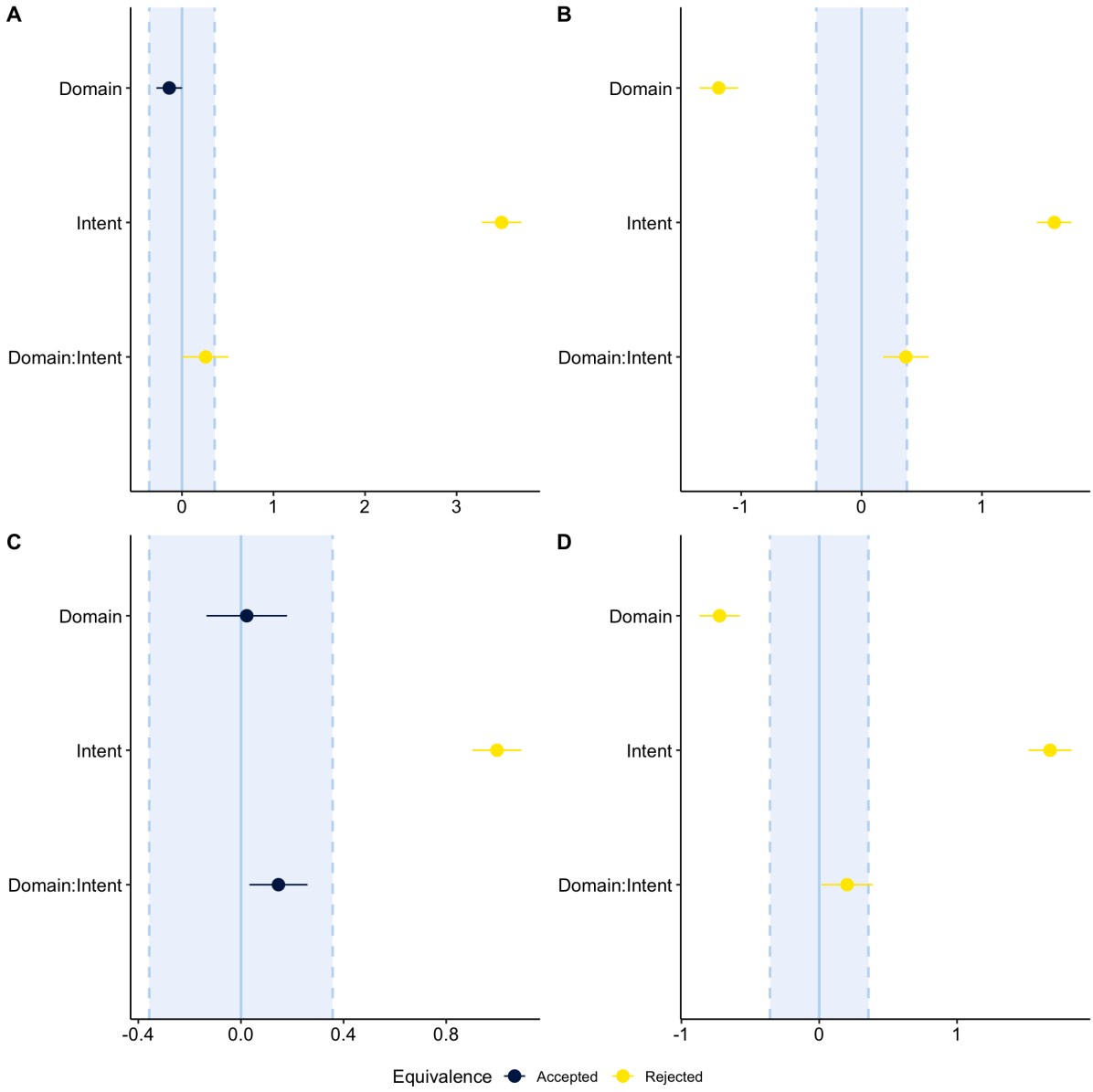
Estimated relationship between the experimental manipulations and judgments of intent (A), purity (B), harm (C), and weirdness (D).



Note. Error bars reflect 95% CIs.

Figure 5

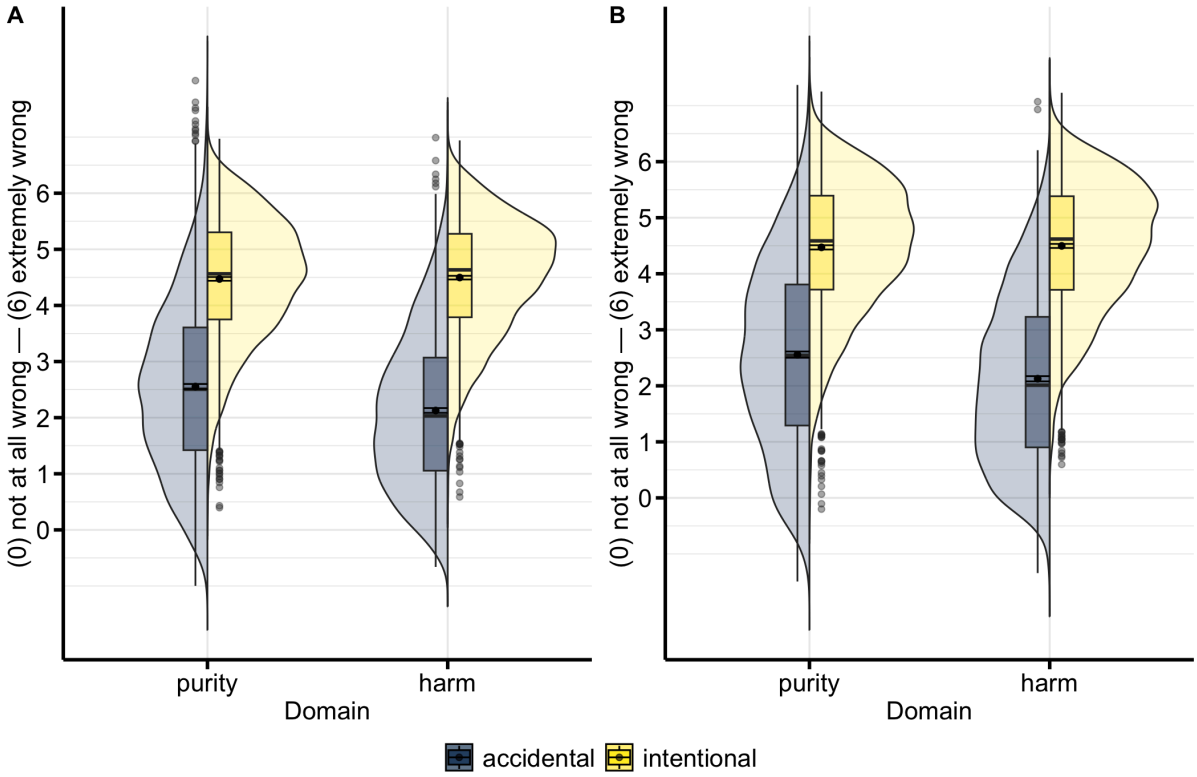
Two one-sided t-tests (TOST) for main and interaction effects on intent (A), purity (B), harm (C), and weirdness (D) judgments.



Note. Yellow estimates indicate that we can reject equivalence to the null region. Blue estimates indicate we can accept equivalence to the null region ($d = \pm 0.2$). Error bars reflect 90% CIs.

Figure 6

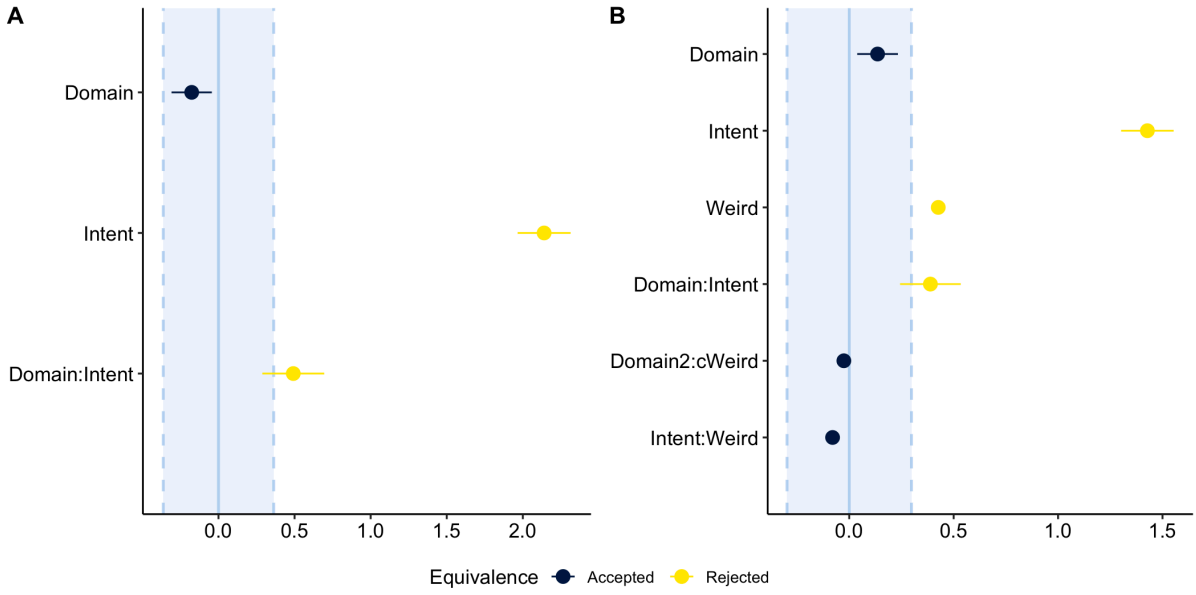
Estimates of moral judgment as a function of experimental condition (A) and after attempting to adjust for weirdness (B).



Note. Error bars reflect 95% CIs.

Figure 7

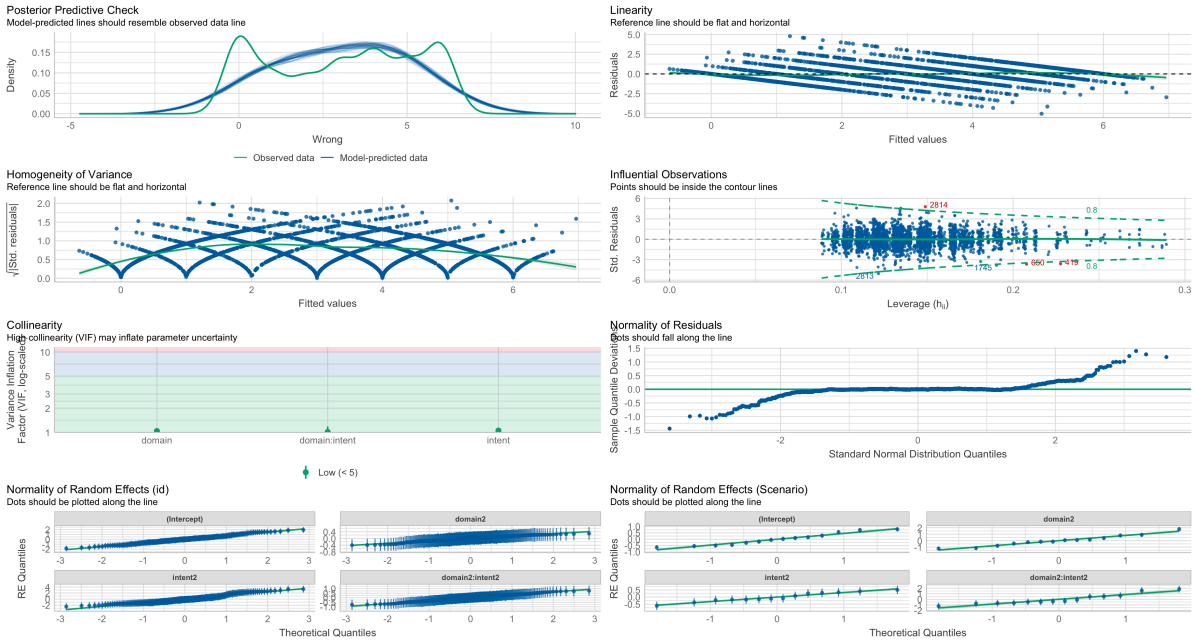
Two one-sided t-tests (TOST) for main and interaction effects on moral judgment (A) and adjusting for weirdness (B).



Note. Yellow estimates indicate that we can reject equivalence to the null region. Blue estimates indicate that we can accept equivalence to the null region ($d = \pm 0.2$). Error bars reflect 90% CIs.

Figure 8

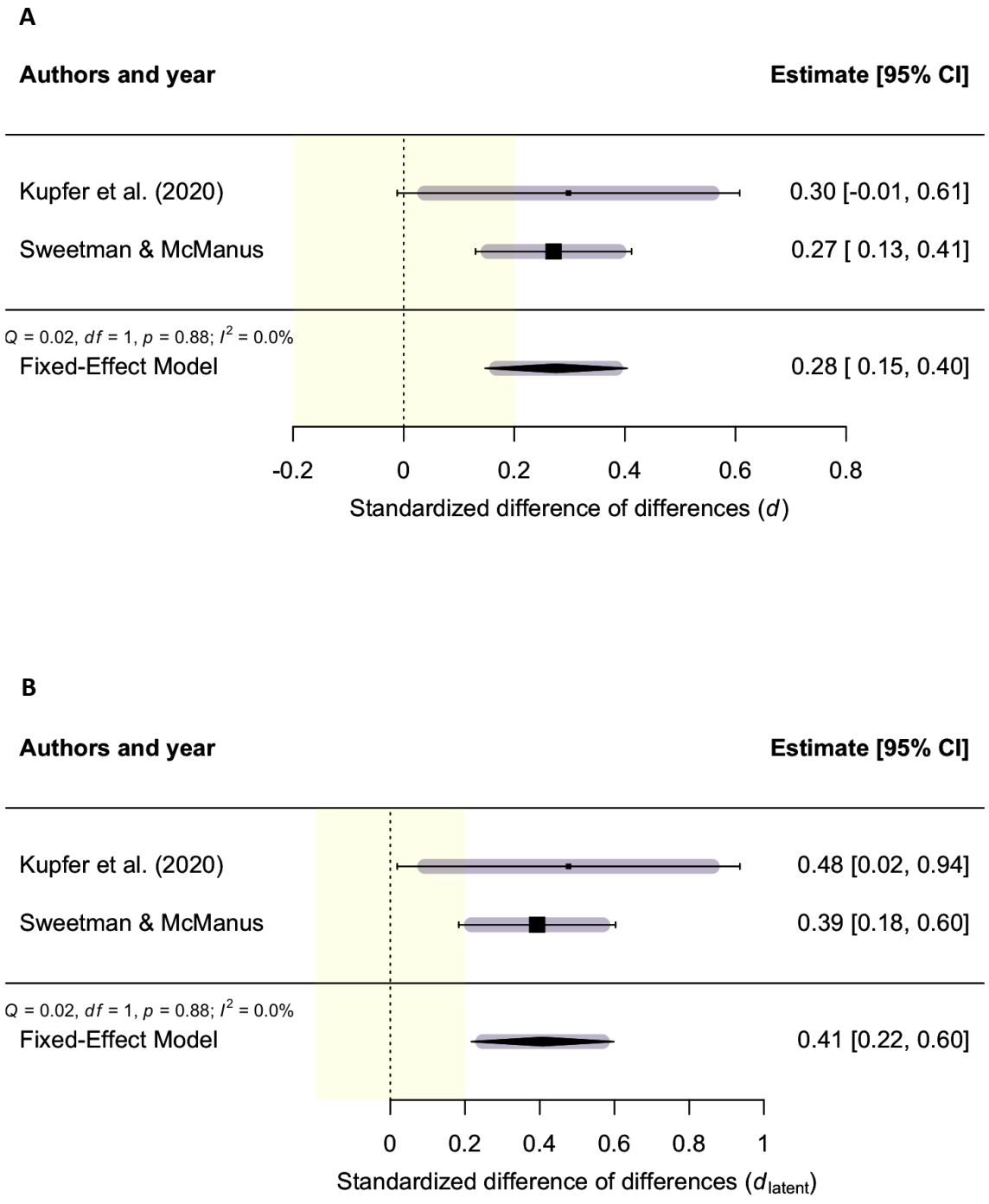
Visual model checks for the LMM for moral judgment fitted in Kupfer et al. (2020).



Note. Shows visual model diagnostics for posterior predictive, linearity, homogeneity of variance, influential observations, collinearity, normality of residuals, and normality of random effects. Red points in the influential observations figure indicate influential observation

Figure 9

Forest plot showing standardized effect sizes (A) Cohen's d and (B) Cohen's d_{latent} for the intent × domain effect from the original and replication studies for the fixed-effects model.



Note. Error bars represent 95% CIs, and the study's point estimate size reflects its weighting. The pooled estimate is represented as a diamond, with its length indicating the 95% CI. The broken line reflects a standardized mean difference of zero. The shaded grey error bars reflect 90% CIs, with the shaded yellow area representing the region of practical equivalence. Cochran's Q and I^2 reflect study heterogeneity.

Table 1

Means, standard deviations, zero-order correlations, and reliabilities for moral wrongness, and the discriminate and direct manipulation checks (n = 226)

parameter *M* (*SD*) ω_t	weird 3.41 (2.11) .94	impure 3.30 (2.14) .95	harm 3.35 (1.86) .97	intent 3.39 (2.50) .78
wrong 3.41 (2.11) .95	.72*** [.65, .78]	.83*** [.79, .87]	.81*** [.75, .85]	.54*** [.44, .62]
intent	.60*** [.51, .68]	.56*** [.47, .65]	.52*** [.41, .61]	
harm	.68*** [.60, .75]	.78*** [.73, .83]		
impure	.74*** [.67, .79]			

Note. Mean (*M*) and standard deviations (*SD*; in brackets); coefficients omega total (ω_t); Pearson's correlation coefficient, 95% CIs [in square brackets], *** $p < .001$.