The Pervasive Impact of Ignorance

Lara Kirfel*
Stanford University

Jonathan Phillips
Dartmouth College

Abstract

Norm violations have been demonstrated to impact a wide-range of seemingly non-normative judgments. Among other things, when agents' actions violate prescriptive norms they tend to be seen as having done those actions more *freely*, as having acted more *intentionally*, as being more of a *cause* of subsequent outcomes, and even as being less *happy*. The explanation of this effect continue to be debated, with some researchers appealing to features of actions that violate norms, and other researcher emphasising the importance of agents' mental states when acting. Here, we report the results of two large-scale experiments that replicate and extend twelve of the studies that originally demonstrated the pervasive impact of norm violations. In each case, we build on the pre-existing experimental paradigms to additionally manipulate whether the agents knew that they were violating a norm while holding fixed the action done. We find evidence for a pervasive impact of ignorance: the impact of norm violations on non-normative judgments depends largely on the agent knowing that they were violating a norm when acting. Moreover, we find evidence that the reduction in the impact of normality is underpinned by people's counterfactual reasoning: people are less likely to consider an alternative to the agent's action if the agent is ignorant. We situate our findings in the wider debate around the role or normality in people's reasoning.

Keywords: normality; knobe effect; ignorance; knowledge; norms; counterfactuals

*Corresponding author: Lara Kirfel (l.kirfel@stanford.edu), Department of Psychology, 450 Jane Stanford Way, Building 420, Stanford, CA 94305

"For there is nothing either good or bad, but thinking makes it so."
"Hamlet", Act 2, Scene 2; Shakespeare (1611)

The Puzzling Impact of Normality

A large and growing body of research has documented that norm violations influence a wide range of intuitive judgments, including judgments of *intentional action* (Knobe, 2003), *causation* (Kominsky & Phillips, 2019), *freedom* (Young & Phillips, 2011), *happiness* (Phillips, De Freitas, Mott, Gruber, & Knobe, 2017), *doing vs. allowing* (Cushman, Knobe, & Sinnott-Armstrong, 2008), *pro-/con-attitude* ascriptions (Pettit & Knobe, 2009), and *modal* judgments (Knobe & Szabó, 2013). Such normality effects are not hard to demonstrate. Consider the following situation:

> While sailing on the sea, a large storm came upon a captain and his ship. As the waves began to grow larger, the captain realized that his small vessel was too heavy and the ship would flood if he didn't make it lighter. The only way that the captain could keep the ship from capsizing was to cut the cargo loose which is weighing the ship down. The captain knows that the cargo contains his wife's expensive art collection because that is what he packed into the cargo.
>
> Fully realizing the cargo contains his wife's expensive art collection, the captain cut the cargo loose and it fell into the sea. While the cargo containing his wife's expensive art collection sank to the bottom of the sea, the captain was able to survive the storm and returned home safely.

Was the captain forced to throw his wife's cargo overboard? Intuitively, 'Yes.' (Phillips & Knobe, 2009). Now consider a variant in which the captains' actions violate a moral norm (changes in italics):

> While sailing on the sea, a large storm came upon a captain and his ship. As the waves began to grow larger, the captain realized that his small vessel was too heavy and the ship would flood if he didn't make it lighter. The only way that the captain could keep the ship from capsizing was to cut the cargo loose which is weighing the ship down. The captain knows that the cargo contains his wife's expensive art collection because that is what he packed into the cargo. *However, he also learned that a number of illegal passengers have hidden in the cargo boxes before the ship left the harbor.*
>
> Fully realizing the cargo contains *passengers*, the captain cut the cargo loose and it fell into the sea. While the cargo containing the *illegal passengers* sank to the bottom of the sea, the captain was able to survive the storm and returned home safely.

In cases like this, people judge the captain to have been much less forced to throw the cargo overboard (Phillips & Knobe, 2009). This is just one example of a much broader phenomenon. A norm-violating agent is judged as acting more intentionally (Knobe, 2003),

having more pro-attitudes (Pettit & Knobe, 2009), being more causal (Icard, Kominsky, & Knobe, 2017), being *less* happy (Phillips et al., 2017) and as making (vs. allowing) an outcome to occur, compared to an agent performing the same action but abiding to the norm (Pettit & Knobe, 2009; Phillips, Luguri, & Knobe, 2015). Hence, at this point there is little debate over *whether* norms influence a wide-range of judgments. Instead, discussion largely centers on why such an effect occurs.

Specific vs. General Explanations of Normality's Impact

Specific accounts

For many of these judgment types, researchers have put forward specific proposals aiming to explain why norm violations influence a given type of judgment in particular (Hindriks, 2014). It is worth taking a look at some of these to appreciate how some norm effects have led to vastly different and partly competing accounts. Consider just the influence of normality on intentional action judgments—sometimes referred to as the "Knobe effect"—which has by itself sparked a variety of explanations (see Feltz, 2007, for an overview). The fact that people are more inclined to judge that agents intentionally brought about harmful vs. helpful side effects has been explained by reference to the agent's cost-benefit trade-off (Machery, 2008; Mallon, 2008), inferences about the agent's mental states and beliefs (Alfano, Beebe, & Robinson, 2012; Laurent, Reich, & Skorinko, 2019; Uttich & Lombrozo, 2010), emotional responses and blame judgements (Cova, Lantian, & Boudesseul, 2016; Hindriks, Douven, & Singmann, 2016; Nadelhoffer, 2006), and even an interaction of 'System 1' and 'System 2' reasoning (Ngo et al., 2015; Pinillos, Smith, Nair, Marchetto, & Mun, 2011). Other theories have adopted a more semantic approach, pointing to the interpretive diversity of the concept of intentionality (Guglielmo & Malle, 2010; Nichols & Ulatowski, 2007) or the role of pragmatic implicatures in these cases (Adams & Steadman, 2007; Lindauer & Southwood, 2021). Theories of this effect in domains other than intentional action are no less diverse. For example, Díaz and Reuter (2021) argue that reduced attributions of happiness to norm-violating agents are the result of how "fitting" people perceive the concept of happiness to be under such circumstances, while others argue that it arises from beliefs about the agent's "true self" (Newman, De Freitas, & Knobe, 2015), and yet others argue that it arises from something about the concept of happiness itself (Phillips, Nyholm, & Liao, 2014). Likewise, a debate around the correct explanation of why people increase ascriptions of free will to abnormally behaving agents has emerged. While some have argued that people's desire for punishment underpins their belief in the free will of a norm-deviating agent (Clark et al., 2014; Clark, Winegard, & Shariff, 2021), others have made the case that dispositional inferences about the agent's character make up this effect (Monroe & Ysidron, 2021).

General accounts

While many of these accounts have developed explanations that take into account the idiosyncratic features of the respective judgment type, some of them converge on the role of blame as a mediating factor for the impact of normality. The basic idea that unites some of these proposals is that norm-violating actions or negative consequences usually trigger a negative evaluative response, and that the respective judgment—e.g., the attribution

of causality, free will or intentional action—is increased in order to justify this evaluation (Alicke, 2000). Accordingly, all of these heterogeneous norm effects are merely a by-product of a more general moral cognitive process, and may be subsumed under some kind of unified account according to which participants are "motivated" to negatively evaluate norm-deviating behavior (Clark et al., 2021).

Other researchers offer a competing unified explanation for the effects of normality (Bernhard, LeBaron, & Phillips, 2022; Knobe, 2010; Phillips & Knobe, 2018; Phillips et al., 2015). This account suggests that the influence of normality on all these judgments—from intentional action to happiness—is driven by people's reasoning about alternative possibilities, or "counterfactual reasoning" (Knobe, 2022; Phillips & Knobe, 2018; Phillips et al., 2015). Both counterfactual and motivated cognition accounts have an advantage over more specialised theories by being able to each explain a large proportion of the effects of normality across judgment domains. Which family of accounts, however, provides the more accurate explanation of the impact of norms on non-moral judgments is a matter of ongoing debate.

Arguably, the most thoroughly investigated case study of the impact of normality is focused on judgments of causation (Willemsen & Kirfel, 2019). Across a now large body of research, the debate has centered on which of these two theoretical accounts can best account for the sensitivity of people's causal judgments to normality (Alicke & Rose, 2012; Kominsky & Phillips, 2019; Samland & Waldmann, 2016; Sytsma, 2020a). A recent major focus of this discussion concerns the finding that the influence of normality on causal judgments is impacted by the agent's *knowledge* of the norm that was violated (Samland, Josephs, Waldmann, & Rakoczy, 2016; Samland & Waldmann, 2016). While both types of unified accounts have claimed to be able to account for this finding, our focus does not primarily concern which of these general accounts is correct. Rather, we want to use the recently discovered effect of ignorance to help adjudicate between specific and general approaches to explaining normality's impact across judgment domains. Specifically, we want to ask whether the impact of ignorance extends to the effects of normality in other judgment domains. To the extent that it does, we would have evidence that the effect of norms across domains are likely all part of a single unified phenomenon, which, consequently, should be explained in a unified way. In other words, we want to pit specific vs. general accounts of the impact of norms across domains by leveraging the prior finding that the influence of norms on causal judgments is sensitive to the epistemic state of the norm-violating agent.

Before proceeding, it will be important to be clearer about the original effect of ignorance we will be seeking to extend. In the following section, we describe how normality affects causal judgments, and explain the recent debate that led to the discovery of the moderating effect of ignorance.

Cause, Norm, and Ignorance

At the broadest level, the general phenomenon is that when some event is one of multiple necessary conditions for a given outcome, the more abnormal that event is, the more people judge it to be the cause of the outcome (Knobe & Fraser, 2008). If two cars crash in the middle of an intersection—one of which ran a red light and one of which did not—the *cause* of the accident is intuitively the driver who violated the traffic norm, not the one who didn't (for the origin of this example, see Kahnemann & Tversky, 1982). But

why? One family of accounts emphasises that people's causal judgments are simply a form of moral responsibility judgment in disguise. In fact, the influence of prescriptive normality in causal cognition is argued to be specific to the concept of "causation". According to this account, there is a semantic ambiguity in the term 'cause' between bringing about an event vs. being morally responsibility for it (Samland & Waldmann, 2016), with some even arguing that these two concepts are used interchangeably (Sytsma, 2020a). In this respect, it is of course not very surprising that they are influenced by prescriptive norm violations (Alicke, 2000; Livengood, Sytsma, & Rose, 2017; Samland & Waldmann, 2016). In other words, when people say that the driver who ran the red light was the cause of the accident, they simply mean or express that he should be blamed or held morally responsible for the accident. Evidence for this claim comes from studies showing that the effect of norm violation on causal judgments decreases if the causal outcome is of a good, rather than bad, nature (Alicke, Rose, & Bloom, 2011; Schwenkler & Sytsma, 2020).

In a challenge to this approach, researchers have pointed out that *descriptive* norm violations—e.g., events that occur despite being very unlikely—exhibit a remarkably similar pattern in intuitive causal judgments (Gerstenberg & Icard, 2020; Kominsky, Phillips, Gerstenberg, Lagnado, & Knobe, 2015). If a forest fire starts in the presence of oxygen, dry leaves, and a lightning strike, people tend to judge that the most statistically abnormal of these events—the lightning strike—was the cause. Responsibility-based accounts are difficult to extend in a way that naturally covers the impact of descriptive norm violations, since these events often do not even involve intentional agents who can be held responsible or blamed.

An alternative family of counterfactual approaches has argued that norms influence causal judgments because causal judgments rely on counterfactual possibilities (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2021; Icard et al., 2017; Lewis, 1974; Pearl, 2009), and norms are well-known to influence counterfactual thought (Icard et al., 2017; Kahneman & Miller, 1986). On this account, people are inclined to judge that the driver who ran the red light was the cause of the accident because they are more inclined to think about what would have happened if he had not run the red light, and are correspondingly less inclined to think about what would have happened if the other driver had stopped at a green light (Phillips et al., 2015). Moreover, this approach can be easily extended to descriptive norms: it is more relevant to consider possibilities in which lightning strikes do not occur than possibilities in which there are no dry leaves in the forest. And thus, it is not difficult to see why both kinds of norm violations will influence causal judgments in similar ways. Crucially, counterfactual accounts are not limited to predicting an influence of normality on causal judgments specifically. Rather, according to counterfactual accounts, norm-violations will exert an effect on any judgment type that relies on reasoning about alternatives (Phillips et al., 2015).

However, a recent criticism of these counterfactual accounts has focused on cases in which an agent *unknowingly* violates a prescriptive norm (Samland & Waldmann, 2016). Samland and Waldmann (2016) argued that in such cases, because a norm is violated, counterfactual accounts should predict that whether or not the agent knows they are violating a norm will not change normality's influence on causal judgments. They went on to show that in fact, participants' causal judgments were extremely sensitive to changes in the agents' knowledge: An agent who unknowingly violated a norm was *not* judged as more

causal than a norm-abiding agent. While this work clearly demonstrated an important relationship between what agents know and the effect of normality on causal judgments, it did not provide decisive evidence against counterfactual explanations of these effects. As Kominsky and Phillips (2019) went on to show, participants' judgments of counterfactual relevance were also highly sensitive to the agents' mental states. People found it less relevant to consider what the agent could have done differently if the agent was ignorant about the norm-violation. Further, these counterfactual relevance judgments precisely predicted the differences in causal judgments. Moreover, recent work by Kirfel and Lagnado (2021a) demonstrated that an important and related effect occurs in the case of descriptive norm violations and causal judgements. Specifically, agents are judged to be more causal of an eventual outcome when they did a statistically unlikely action that was necessary for that outcome, but this effect only occurs when the agent knows the action leads to the outcome.

Moving beyond Causation: The Impact of Ignorance

For our purposes, a critical upshot of this growing body of research is that the effect of norms on judgments of causation are broadly sensitive to agents' epistemic states. More specifically, ignorance reduces or eliminates the influence of an action's abnormality on causal judgements: abnormal actions done by ignorant agents are not perceived as more causal than normal actions. Given that causal judgments are just one type of judgments that are sensitive to normality, the question arises whether agents' epistemic states similarly affect the impact of norms across the wide range of different judgments. More precisely, it remains an important but unanswered question whether ignorance reduces the impact of normality on people's attributions of intentional action, freedom, happiness, and so on. This question is of theoretical interest because a systematic moderating effect of epistemic states on the impact of norms across a variety of judgments would provide empirical (rather than merely theoretical) evidence for a common mechanism by which normality affects judgments across domains. That is, evidence for a moderating effect of ignorance across all domains of non-moral judgments would lend weight to a single more unified account, rather than various specialised accounts of normality's impact. The present study aims to answer this question.

In addition to investigating whether epistemic states play a systematic role for the influence of normality, we also aim to undertake a first attempt towards uncovering *how* they might do so. Central to counterfactual accounts is the idea that people engage in reasoning about what could have gone differently (Kominsky & Phillips, 2019). Some studies in the domain of causal cognition show that people's counterfactual reasoning about alternatives to the causal agent's action are significantly impacted by the agent's epistemic state (Gilbert, Tenney, Holland, & Spellman, 2015; Kirfel & Lagnado, 2021b; Kominsky & Phillips, 2019; Spellman & Gilbert, 2014). If the causal agent lacks knowledge about a certain aspects of a situation (e.g. norms, causal structure, etc.), people are less inclined to consider what the agent could have done differently (Gilbert et al., 2015; Kirfel & Lagnado, 2021b; Kominsky & Phillips, 2019). A promising way forward is hence to investigate whether people's reluctance to imagine alternatives to the actions of an ignorant agent equally applies in judgment domains other than causality. Kominsky and Phillips (2019) suggest that the perceived normality of an action might depend on the agent's awareness of whether their action violates a norm ("Expectation-based normality account"). On the contrary, proponents

of a responsibility-based framework have argued that reasoning about alternatives is itself driven by moral evaluations (Sytsma, 2020b). As a way into the question how epistemic states might moderate the influence of normality, we will also investigate people's reasoning about what a knowing vs. ignorant agent could have done, or known.

The Present Study

The experiments we report make a novel contribution by considering the pervasive impact of normality as a whole (using a meta-analytic approach) and asking whether agents' epistemic states moderate the impact of normality for each of the different judgments that have previously been shown to be impacted by normality. More specifically, we both (1) attempt to replicate the effect of normality that has been previously demonstrated, and then (2) compare the replicated effect to minimally modified versions of the original materials that allow us to ask whether the effect of normality is sensitive to changes in the agents' epistemic states. First, we will ask, in each individual study, whether we replicated the originally observed effect. In all cases where we are able to replicate the original effect, we will then do the same statistical test, but replace the original knowledgeable norm-violation condition with the new ignorant norm-violation condition. Drawing on recent work on the impact of ignorance on normality effects in causal judgments (Kirfel & Lagnado, 2021a; Samland & Waldmann, 2016), we predict a similar moderating effect of ignorance on judgments of intentional action, freedom, happiness, etc. Our hypothesis is that the effect sizes for the second analysis (ignorant norm violation vs. neutral) will be smaller than those for the first analysis (norm violation vs. neutral). That is, we predict that the difference in judgments about an agent who does not violate a norm and an agent who unknowingly violates a norm (here analysed as effect sizes) will be smaller than the difference between an agent who does not violate a norm and one who knowingly violates a norm. Returning to Phillips and Knobe (2009)'s ship scenario to illustrate more concretely, we predict that the difference in force judgments between the immoral but ignorant condition (captain is *unaware* of passengers) and the neutral condition (captain is aware of cargo) will be smaller than the difference between the immoral and knowledgeable condition (captain is aware of passengers) and the neutral condition (captain is aware of cargo). In the current study, we will solely focus on the impact of *prescriptive* normality.

The results of this large-scale experiment will not only be informative for debates that have sought to explain the impact of normality in each separate case, but also will inform the broader question of whether researchers should seek a unified explanation of the pervasive impact of normality. In the General Discussion, we will return to the question of whether and how the two theoretical accounts outline above—responsibility vs. counterfactuals—can provide an explanation of our findings.

Methods

In this replication study and experimental meta-analysis, we selected 12 studies published between 2003 and 2019, containing 6 different paradigms with in total 29 statistical effects taken to indicate the influence of norm violations on different types of judgments. The replication study was carried out across two separate studies, "Study 1" (*cf.* Kirfel & Phillips, 2021), and "Study 2".

Selected Studies

To be included in this large-scale experiment, studies needed to have investigated the impact of prescriptive norm violations, broadly construed (violating a conventional norm, causing harm, etc.), on seemingly non-normative judgments. We identified 6 different judgment domains for which this was the case: *causation* (Kominsky & Phillips, 2019), *doing vs. allowing* (Cushman et al., 2008), *freedom* (Phillips & Knobe, 2009), *happiness* (Phillips et al., 2017), *mental state ascriptions* (Pettit & Knobe, 2009) and *modal judgments* (Knobe & Szabó, 2013).

Causation. A series of studies finds that if the actions of two agents are necessary for an outcome to occur (a "conjunctive causal structure"), people judge the agent who violated a norm to be more of a cause of the outcome than the agent who acted according to the norm. In contrast, if both agents' actions are independently sufficient to bring about the outcome (a "disjunctive causal structure"), the agent who acted immorally is judged to be less of a cause of the outcome. We selected four scenarios ("battery", "bridge", "motion detector", "computer"), which had both a conjunctive and disjunctive version from Kominsky and Phillips (2019). As in the original study by Kominsky and Phillips (2019), we added two comprehension checks for each scenario in study 2. As a result, the scenarios from Kominsky and Phillips's 2019 study in our second study were the only conditions where participants were excluded from the analysis based on comprehension checks. In addition, we also selected the "pen" scenario from Knobe and Fraser (2008), which only had a conjunctive version.

Doing vs. allowing. Work on the 'doing/allowing' distinction shows that morally bad behavior is more likely to be construed as actively 'doing' than as passively 'allowing'. We selected the "Dr. Bennet" scenario from Cushman et al. (2008) for our study, in which a doctor removes a homeless man from life support.

Freedom. As discussed in the introduction, studies on people's judgements about the freedom to act show that agents who acted immorally (vs. neutrally) are more thought to have acted freely (i.e., were not forced to do that action). Young and Phillips (2011) found that this effect is also affected by the moral focus of the force judgment: People agree more with the active form of the sentence "X forced Y to act" than the passive form "Y was forced by X to act", specifically when the act violates a norm. For our study, we used the original "ship" vignette from Phillips and Knobe (2009) as described above, as well as the *active* and *passive* version of the "ship" and "doctor" scenario from Young and Phillips (2011).

Happiness. Previously, research found that even when an agent is described as satisfying all of the psychological criteria for happiness (high positive affect, low negative affect, high life satisfaction), participants are disinclined to rate the agents as being "happy" when they believe the agent to be living an immoral life (though not when living morally good or neutral life). We selected the "nurse" scenario from Phillips et al. (2017) as paradigmatic test case of this effect for our study.

Mental State Ascriptions. This line of research, also known as the "side-effect effect", shows that an agent who brings about a side effect is judged as having intended this side effect to a greater extent when this effect is bad vs. good. Subsequent studies have shown that this pattern occurs for the attribution of other mental states (e.g. desire) as well Pettit and Knobe (2009), and find an inverse effect for attribution of opposition: People

judge the agent to have opposed the effect *less* when the side effect is morally bad vs. good. We selected the original "chairman" scenario for testing "intentionality" from Knobe (2003) and "decision/desire" from Pettit and Knobe (2009). Additionally from Pettit and Knobe (2009), we included the "manager" scenario (which tested "advocate / in favour of"), the "CEO" scenario (testing "opposed to") and the "bomb" scenario (testing "intended to"). In addition, we selected the "gizmo", "scrubs" and "truck trailers" scenarios from Uttich and Lombrozo (2010) who tested the attributions of intentional action to agents violating conventional norms. Throughout, we subsume the various effects in this area under the term "mental states ascriptions".

Modal Proxies. Knobe and Szabó (2013) demonstrated that the effect of norm violations found in previous research on force, intention, causation extended to 'modal proxies' of these judgments. For example, just as people would say an agent was more forced to do a morally neutral action than an immoral action, they more agreed with the sentence "Given the circumstances, the agent had to do that action" when the action was morally neutral than when the action was immoral. We selected the "captain", "pen" and "bulls-eye" vignettes for replication and extension.

Pre-replication procedure

Each of the 29 selected scenarios included two experimental conditions (see the "ship" scenario from Phillips and Knobe (2009) from the introduction): One "Normal" condition in which the agent's action is morally good or neutral, and one "Norm Violating" condition in which the agent acts morally bad. We created a third experimental condition for each of the 29 scenarios that matched the "Norm Violation" condition in all aspects, except for the agent's epistemic state about the normality of their action. In the "Ignorant Norm-Violation" condition, the agent's action violates a norm (e.g., causes harm) but the agent is unaware that their action violates a norm.[1] To illustrate, here is the "Ignorant Norm Violation" condition that extends the scenario first tested in Phillips and Knobe (2009)'s ship scenario (differences from previous condition again indicated by italics):

> While sailing on the sea, a large storm came upon a captain and his ship. As the waves began to grow larger, the captain realized that his small vessel was too heavy and the ship would flood if he didn't make it lighter. The only way that the captain could keep the ship from capsizing was to cut the cargo loose which is weighing the ship down. The captain *thinks* that the cargo contains his

[1] What exactly about the norm-violating action must the agent be aware of in order for normality to have an impact on people's non-moral judgements about that agent? Given that researchers find a similar epistemic effect for statistical norms (Kirfel & Lagnado, 2021a), one might question whether it is really necessary that the agent knows the norm to be of explicitly moral or prescriptive nature, or whether knowledge about their action deviating from *some kind of norm* is sufficient. While our study manipulates knowledge about prescriptive norms, broadly construed, it does not settle whether an agent's knowledge about the moral significance of the norm, its status as a norm, the agent's inferences about the consequences of a norm violation, or other aspects are the essential epistemic component that drive the influence on non-moral judgments. While there is an ongoing discussion in philosophy about the necessary epistemic conditions that an agent needs to meet in order to justify a negative moral judgment (Sher, 2009), future research will need to study people's understanding of what it means for an agent to know about a norm, and which aspects of this knowledge are essential for the role of norms in non-normative cognition.

> wife's expensive art collection because that is what he packed into the cargo. However, *completely unbeknownst to the captain*, a number of illegal passengers have hidden in the cargo boxes before the ship left the harbor.
>
> *Without* realizing the cargo contains passengers, the captain cut the cargo loose and it fell into the sea. While the cargo containing the illegal passengers sank to the bottom of the sea, the captain was able to survive the storm and returned home safely.

In order to match all three conditions and to allow for a consistent manipulation of the agent's knowledge of the normality of their action, we were often required to modify aspects of some of the original vignettes. We similarly adapted the precise phrasing of the dependent variable questions of some studies and standardised the rating scales across studies (see below). Accordingly, our replications varied from a relatively direct replication in some cases, to something closer to that of a conceptual replication in other cases (Hendrick, 1990; Lynch Jr, Bradlow, Huber, & Lehmann, 2015). Importantly, the focus of this study centers on investigating the moderating role of epistemic states on the effects of morality, rather than the strict replicability of the original effects.

Procedure

Combined Analysis of Study 1 and 2. This replication study combines data from two separate studies: "Study 1" (N = 1554), also described in Kirfel and Phillips (2021), and "Study 2" (N = 1328). Given the comparatively low replication rate of the original norm-effects observed in Study 1 (62%), we decided to run an additional study, Study 2, with an identical experimental design but experimental methods and materials that more precisely matched the original studies being replicated and extended. In addition, Study 2 included new counterfactual dependent variables not included in Study 1. Given that the replication rate of Study 2 remained at 62%, we decided to proceed by combining the data from Study 1 and 2 for all questions that were identical across studies. (We discuss the replicability of these prior findings in more detail in Discussion Part 1.)

Participants & Design

Combining study 1 and 2, we recruited 2880 participants via Amazon Mechanical Turk (M_{age} = 41.28, SD_{age} = 12.85, N_{female}= *1423,* $N_{non\text{-}binary}$= *11,* $N_{not\text{-}disclosed}$= 8). 100 participants were filtered out from study 2 for not answering one or more comprehension check questions for the scenarios in Kominsky and Phillips (2019) correctly, leaving a final sample of N = 2780. Participants were payed \$1 for ∼5 minutes of study participation.

Our study employed a 3 *normality* (normal vs. norm violation vs. ignorant norm violation) × 29 *prior study* design. Note that we specifically decided against using a 2 (norm violation) × 2 (epistemic state) study design that would include a condition describing the agent as ignorant of the fact that their action did *not* violate a norm. Such a description leads to the pragmatic implication that the agent acted recklessly or negligently, and thus actually *did* knowingly violate a norm (see also Discussion Part I). For simplicity and clarity, we opted to include only the additional ignorant norm violation condition, which still allows us to compare the size of the relevant effect when agents knowingly vs. unknowingly violate

a norm. Both norm and prior study were manipulated within participants. That is, each participant saw one example of each of the normality conditions in randomised order, and for each normality condition, the prior study that condition was drawn from was randomly sampled from the 29 different studies included in the study. After reading each scenario, participants first responded to either three (Study 1) or five (Study 2) different questions in the following order.

Original Dependent Variable. A rating of the key dependent variable used in prior work, which following the original study, was sometimes adapted to the normality condition ("Did the chairman intend to help [harm] the environment?") and sometimes not. For the ship scenario, for example, the dependent variable was an agreement rating with the same statement in all three norm conditions: "The ship captain was forced to cut the cargo loose and let it fall into the sea." on a 7-point Likert scale (1-"strongly disagree", 7-"strongly agree").

Action & Epistemic State Counterfactual [only Study 2]. In Study 2, we added two questions targeting people's counterfactual reasoning process. Participants were prompted to engage in thinking about an alternative course of events for the respective experimental scenario they saw: "Now suppose that some people are discussing this scenario and wondering how things could have been different. In thinking about what could have happened differently, please tell us whether it would be relevant or irrelevant to focus on the following: [...]". This prompt was followed by two statements, one focusing on the agent's action, "... what [agent] could have done differently." (*"Action-focused Counterfactual"*), and one focusing on the agent's epistemic state, "... what [agent] could have known."(*"Epistemic State-focused Counterfactual"*). While the Action-focused Counterfactual is designed to target people's thinking about the norm-violating or harm-producing action, the Epistemic State-focused Counterfactual captures a wide array of situations in which the agent acquires knowledge about the norm/harm, including actions of the agent that lead to their knowledge acquisition ("epistemic actions"), (Kirsh & Maglio, 1994). Both statements were rated by participants with regards to their relevance on a 100-point Likert scale (0-"Not at all relevant", 100-"Highly relevant").

Knowledge Check. A knowledge check question, asking about the central agent's knowledge of the abnormality of their action (e.g. "Please rate how much you agree or disagree with the following statement: 'The captain knew that a number of illegal passengers were hiding in the cargo boxes.' ") on a 7-point Likert scale (1-"strongly disagree", 7-"strongly agree").

Should know. A question about what the agent should have known with regards to the abnormality of their action (e.g., "Please rate how much you agree or disagree with the following statement: 'The captain *should* have known that a number of illegal passengers were hiding in the cargo boxes.' ") on a 7-point Likert scale (1-"strongly disagree", 7-"strongly agree"). This question served to examine whether the manipulation of epistemic states in our experiments not only influenced people's beliefs about what the agents actually knew, but also their normative beliefs about what the agents should have known.

Analysis approach

Study materials and analyses were pre-registered at `https://osf.io/g52zs`. While the methodological approach of this paper sits somewhere in between a meta-analysis and

a large-scale, multi-part experiment, two key features make it closer to a novel large-scale experiment. First, all of the data we analyze are data that we ourselves collected from materials that were modified from the original studies. Second, and more importantly, the key effect of interest in this paper is not the effect of normality (which could simply be estimated from prior work), but rather the moderation of the effect of normality based on changes to the agent's epistemic states. That is, we are not trying to meta-analytically describe an existing set of findings, but rather are manipulating a novel variable an attempting to estimate its effect across a wide-range of scenarios, and ask whether it is likely to generalize to additional effects of normality. Our analysis approach thus employs standard experimental analysis procedures, that is, using linear mixed effect models (Sheu & Suzuki, 2001) to fit study-level effects.

Replication Analysis. For each individual study, we first tested whether we replicated the originally observed effect, i.e., whether there was significant difference in dependent variable between the "Normal" and "Norm Violation" conditions. A replication was considered successful when $p < .05$ and the effect was in the same direction as the original effect. We collected the effect sizes for those effects that were replicated (all converted to Cohen's d).

Simple Effects Analysis. To simplify our analysis, we reduced all normality effects to simple effects. That is, interaction effects such as the causal structure $\times$ normality interaction effect observed in Kominsky and Phillips (2019) or main effects averaged across different scenarios (Lombrozo & Uttich, 2010) were decomposed into two separate simple effects (by scenario). In all cases where we were able to replicate the original effect, we then performed the same statistical test, but replaced the "Norm Violation" condition with the newly created "Ignorant Norm Violation" condition and recorded the new effect size that measures the difference between these conditions. We adopted an effect-level analysis approach towards our hypothesis that norm effects are influenced by agents' epistemic states, predicting that effect sizes of the statistical tests for "Neutral vs. Ignorant Norm Violation" will be smaller than in the "Neutral vs. Norm Violation" tests.

In order to statistically evaluate this hypothesis, we first aligned all norm effects showing the direction of the effect size for those studies showing that the norm manipulation leads to a reduction in the DV rating (e.g. the agent is judged as *less* forced in the abnormal vs. neutral condition). We then built a linear mixed-effects null model including a random intercept for the study being replicated and extended (1 | study) and a random intercept and slope for the impact of epistemic states across paradigms (Epistemic State | paradigm), and compared it to a model had the same random effects structure but included a fixed effect for "Epistemic State". This factor coded for whether each effect size was a case in which the norm violation was known vs. unknown. The fixed effect was determined to be significant if the fit of the model that included the fixed factor for epistemic state differed significantly from the model including only the random effects. The same procedure was also used for both kinds of knowledge ratings.[2]

[2]All materials of the experiments, data and analysis code can be found here: `https://github.com/LaraKirfel/PervasiveIgnorance`

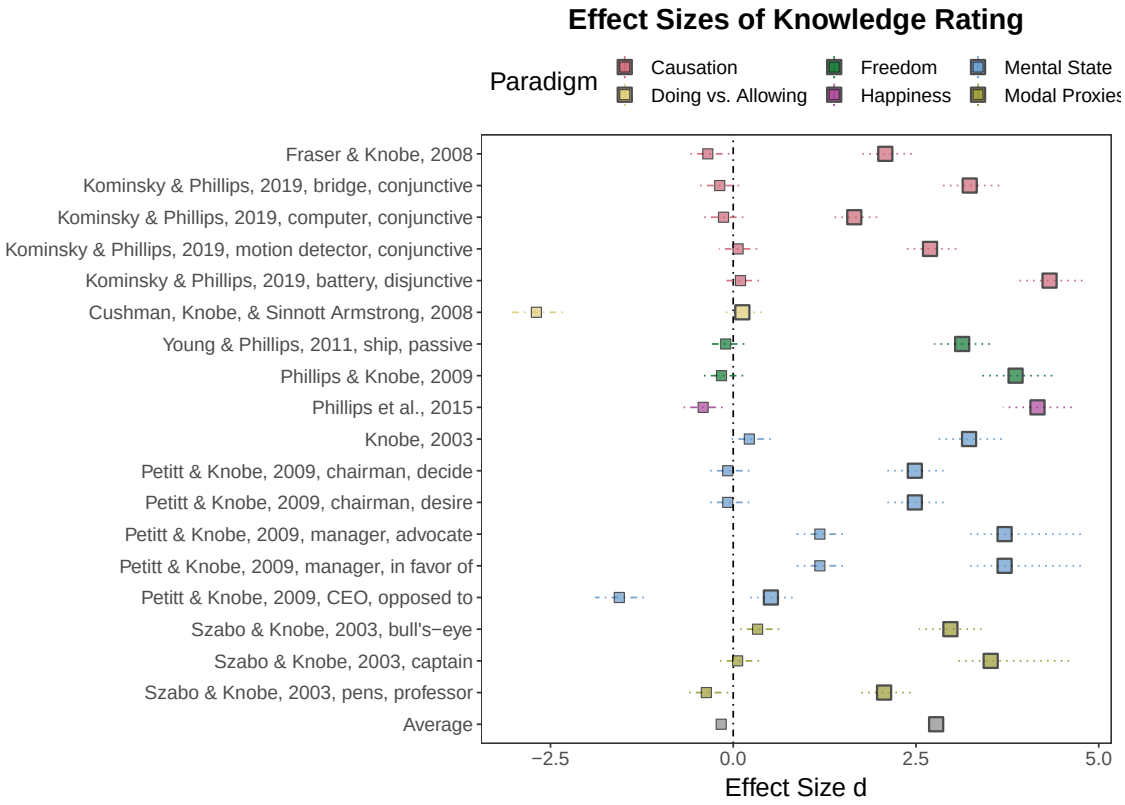

Figure 1. **Knowledge Rating Effect Sizes by Study**: Effect sizes of Knowledge rating in the original norm violation condition (neutral vs. knowing norm-violation) are marked by large squares, effect sizes in the new norm violation condition (neutral vs. ignorant norm-violation) are marked by small squares. Error bars depict 95% confidence intervals of effect sizes

Results

Part I: Knowledge Ratings & Original Dependent Variables

Combining the data from our two studies ($N = 2781$), we successfully replicated 18 out of 29 effects of normality on non-moral judgments (62%): 5 effects on judgments of *causation* (out of 9), the effect on *doing vs. allowing*, two effects on judgments about *freedom* (out of 5), the effect on judgments about *happiness*, 6 effects on *mental state ascriptions* (out of 10) and all three effects on *modal proxies*. Because our focus was not measuring the strict replicability of original effects, but rather comparing effect sizes for a knowing vs. ignorant norm-violation across a variety of different judgment domains, we proceeded to analyse the data for only the 18 successfully replicated effects.

Knowledge Check. The experiment included two measures that tracked people's perception about the focal agent's epistemic state about the normality of their action. Our

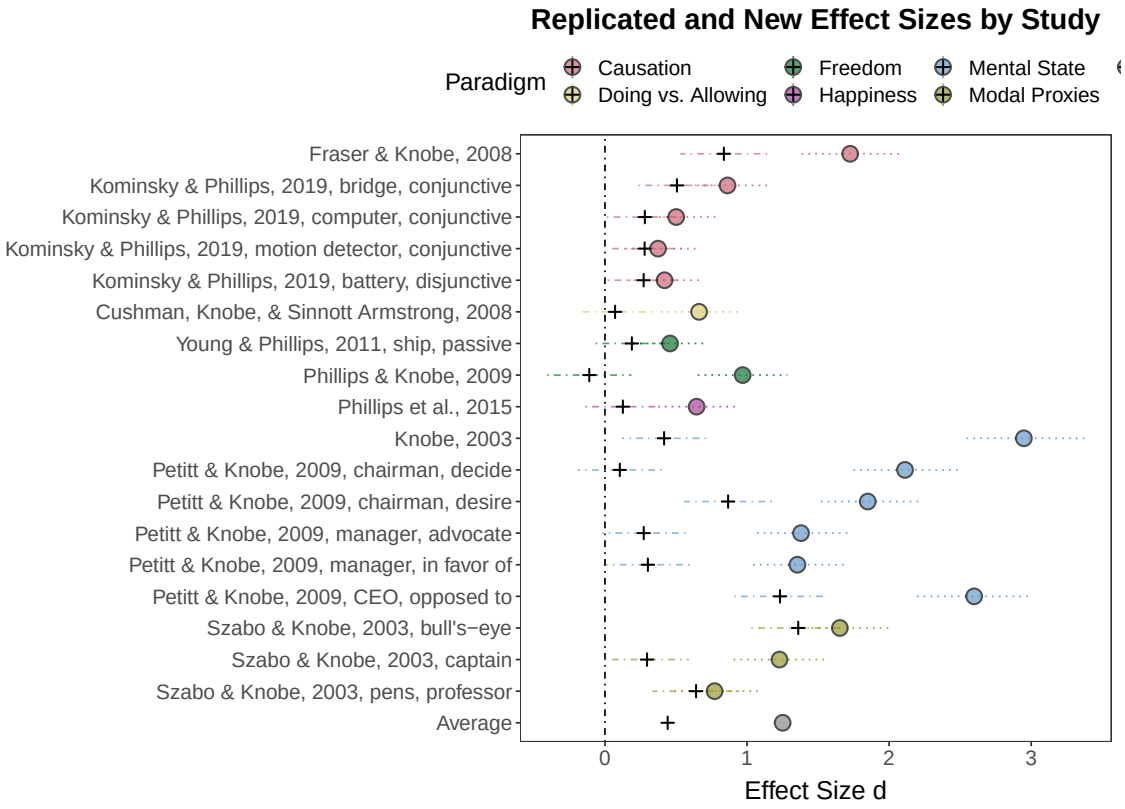

Figure 2. **DV Effect Sizes by Study**: Replication effect sizes of the original effects (neutral vs. norm-violation) are marked by dots, effect sizes of the new effect (neutral vs. ignorant norm-violation) are marked by crosses. Error bars depict 95% confidence intervals of effect sizes

analysis revealed a significant effect for whether the norm violation occurred knowingly vs. unwittingly on effect sizes of knowledge ratings $\chi^2(1) = 19.39$; $p < .001$ (b = -1.50, SE = .10, t = -14.89). The effect on ratings of the agent's knowledge about the abnormality of their behaviour was larger when the norm violation was intentional ($M = 2.78$, $SD = 1.16$) vs. ignorant (M = -0.16, SD = .86) (Figure 1). While this is not surprising, it serves as an important manipulation check, demonstrating that we successfully manipulated participants' perceptions of the agents' knowledge, and more specifically, the agent's knowledge about whether they violated a norm.

Should Know Check. Additionally, the extent to which people judged that the agent should have known that their behaviour was counter-normative was also predicted by our manipulations of the agents' knowledge, $\chi^2(1) = 8.39$; $p < .01$ (b = -.64, SE = .16, t = -4.14). Differences in whether the agent should have known were larger when the agent knowingly violated a norm ($M = .92$, $SD = .94$) than when they did so unknowingly (M = -0.35, SD = .51). This suggests that our manipulations not only successfully manipulated

participants' perceptions of whether the agents did in fact know that their actions violated a norm, but also whether they *should* have known that.

Dependent Variable. Given these results, we can now turn to the critical test of our hypothesis: whether our manipulations of the agents' epistemic states affected the pervasive impact of norms. We found that they did. Once again, the likelihood ratio test indicated that a model including a fixed effect for "Epistemic State" provided a better fit for effect sizes of the dependent variable than a model without it $\chi^2(1) = 6.28$; $p = .01$ (b = -.34, SE = .10, t = -3.37). The average replication effect size, i.e. the effect size for the original effect of norm violation was larger ($M = 1.25$, $SD = 0.77$) than the average new effect size, i.e., the effect of an ignorant norm violation ($M = 0.44$, $SD = 0.40$) (Figure 2).

Discussion Part I

We successfully replicated 18 effects demonstrating the influence of norm-violations on judgments of causation, freedom, happiness, doing vs. allowing, mental state ascriptions and modal claims (replication rate 62%). This replication rate is relatively low compared to previous work on the replicability of experimental philosophy (Cova et al., 2021). However, it is important to remember that the majority of these studies were not direct replications. In fact, our replication study and experimental meta-analysis differed from standard replication procedures in two important aspects.

First, our studies involved modifying the original experimental materials to allow for a close match between the new conditions in which the agent was ignorant of the normative status of their action. While Study 2 was conducted to partly address this issue, some of the scenarios had to be significantly adapted such that a meaningful manipulation of the agent's knowledge state was possible across all experimental conditions. Investigating the effect of norms with slightly altered original material might have led to some effects not being replicated. This renders our replication attempt closer to that of a conceptual replication.

Second, our approach to determining the size of our sample differs from typical direct-replication procedures. Because we conducted these experiments as single large-scale studies which randomly assigned participants to conditions, we were likely under-powered to detect some of the smaller effects, and over-powered to detect larger effects. Crucially, however, we enforced simple effect tests by scenario on all studies, hence reducing more complex interaction and aggregated effects to simplified t-tests between norm vs. no norm conditions. While this simplified approach allows for a direct comparison of the effect of intentional vs. ignorant norm violations, it might also account for why some of the effects did not replicate.

Our replication rate was slightly below the replication rate previously found for studies in experimental philosophy (70%) (Cova et al., 2021; Strickland & De Cruz, 2021), yet still higher than in other sub-fields of psychology (Collaboration, 2015; Nosek et al., 2022). After taking into account deviations in material, methods and analysis, the residual of non-replicated findings might simply point to the fact that a few norm effects are less robust than originally found (Machery, Grau, & Pury, 2020; Stuart, Colaço, & Machery, 2019). For transparency, we have included an overview in Appendix A with the results of our study (replicated and new effect sizes) together with data on the effect sizes from the original literature, as well as the statistical tests that were employed in the original

studies. All results from our analysis can be found here: `https://larakirfel.github.io/PervasiveIgnorance/`.

From the 18 norm effects that we replicated, participant's knowledge ratings validated an effective manipulation of knowledge about the norm in our modified scenarios. Participants attributed less knowledge to the ignorant agent about the immoral status of their behaviour and also were less inclined to judge that the ignorant agent "should have known" about the norm.

Critically, we also found that the agent's epistemic state about the norm does indeed moderate the impact of norms on the different judgment types. The effect of a norm violation on judgments about causation, freedom, etc., is reduced when the agent is ignorant about the fact that they are violating a norm. The extent of reduction varies across studies and judgment domains. Not all effect sizes in the 'ignorant norm violation' condition approximated zero, suggesting that in some studies, the ignorant agent's norm violation still exerted a weak influence on participants' responses. The consistent reduction of the impact of normality, however, demonstrates the crucial role of the agent's epistemic states for the influence of norms on people's judgments.

Given that our study did not include a condition in which the agent is norm-conforming but ignorant about their norm-conformity ('ignorant/neutral'), it might be questioned whether ignorance has a broader influence than suggested in this study. Rather than specifically moderating the influence of norm-violations, it might be that ignorance also influences judgments when the agent is not violating a norm. However, people do not seem to attend to the exact mental state of an agent if the agent is engaging in morally insignificant, harmless behavior (Alicke, 2000; Fincham, 1985; Young & Tsoi, 2013). In contrast, creating scenarios in which an agent acts in a norm-conforming manner while actually being ignorant about the moral status of their action might give rise to inferences about the agent's recklessness (Pizarro & Tannenbaum, 2012; Pizarro, Tannenbaum, & Uhlmann, 2012), or even suggest the agent might suspect they're violating a norm (Yaffe, 2018). While we refrained from including this condition in our study, future research should further investigated the broader impact of agents' ignorance on non-moral judgments.

In sum, the first part of our study confirmed the systematic influence of the agent's epistemic state for the effect of norm-violations on non-moral judgements. We now turn to the two additional measures included in Study 2 to investigate people's reasoning about alternatives in these cases.

Part II: Counterfactuals [Study 2]

Action Counterfactual. In study 2, we added two measures of counterfactual relevance. Continuing our meta-analytic approach, we analysed the effect sizes of an intentional vs. ignorant norm violation on people's relevance ratings of two different counterfactuals. The first counterfactual statement concerned a change in the focal agent's action, i.e., what the agent could have done differently. Including a factor coding for whether the norm was consciously vs. unknowingly violated provided a better fit for people's responses than a null model, $\chi^2(1) = 11.23$; $p < .001$ (b = -.40, SE = .07, t = 5.71). People found it more relevant to consider what the agent could have done differently when the agent's immoral behaviour was done knowingly (M = 1.12, SD = .53), compared to when the agent was unaware of the fact that they were acting in a morally bad way (M = 0.36, SD = 0.52) (Figure 3).

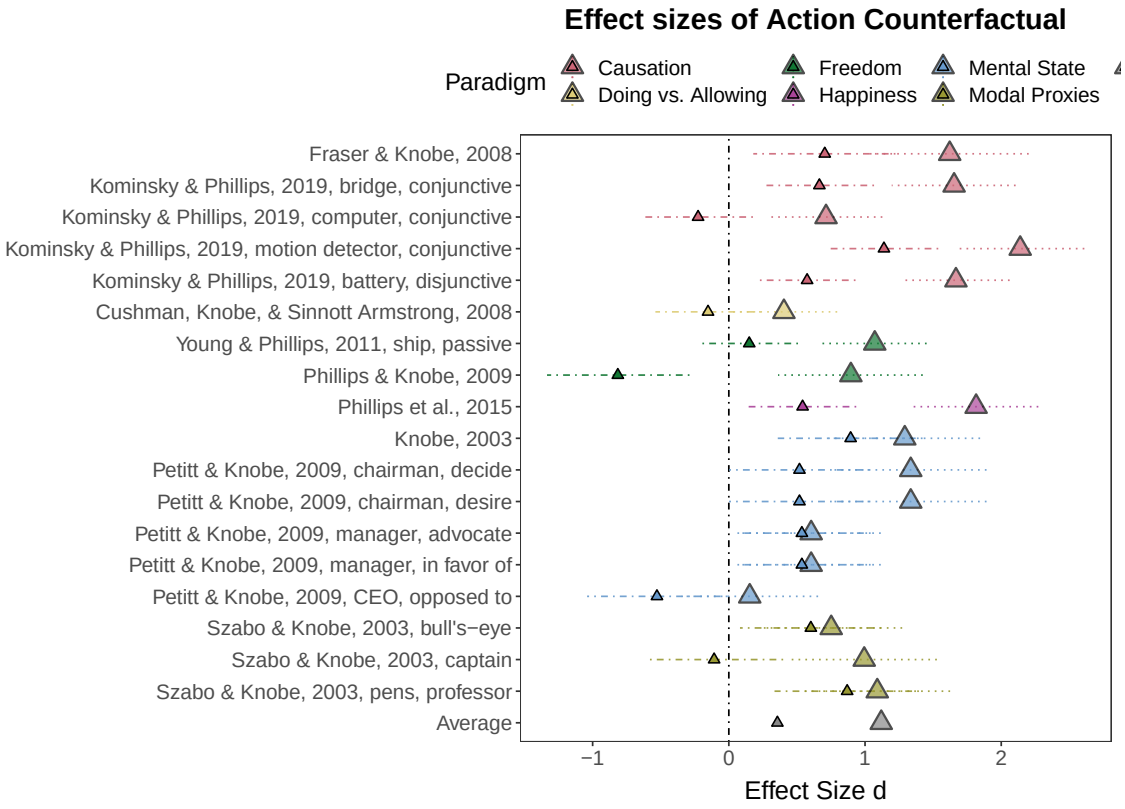

Figure 3. **Action Counterfactual Effect Sizes by Study**: Effect sizes of Relevance ratings of the Action Counterfactual in the original norm violation condition (neutral vs. knowing norm-violation) are marked by large triangles, effect sizes in the new norm violation condition (neutral vs. ignorant norm-violation) are marked by small triangles. Error bars depict 95% confidence intervals of effect sizes

Put differently, people perceived the possibility of a change in the agent's (norm-violating) action as less relevant if the agent was ignorant of violating the norm.

Epistemic Counterfactual. The second counterfactual statement concerned a change in the agent's epistemic state, broadly construed. Here, people evaluated how relevant they considered a change in the agent's knowledge state, that is, what the agent could have known. Despite a general trend towards increased counterfactual relevance ratings in the ignorant norm violation condition ($M = 0.80$, $SD = 0.61$) compared to the deliberate norm violation condition ($M = 0.39$, $SD = 0.37$), this difference in effect sizes was not significant, $\chi^2(1) = 1.50$; $p = .22$ (Figure 4).

Discussion Part II

In our second study, we added two response measures targeting people's counterfactual reasoning process. These ratings focused on changes in two different variables in

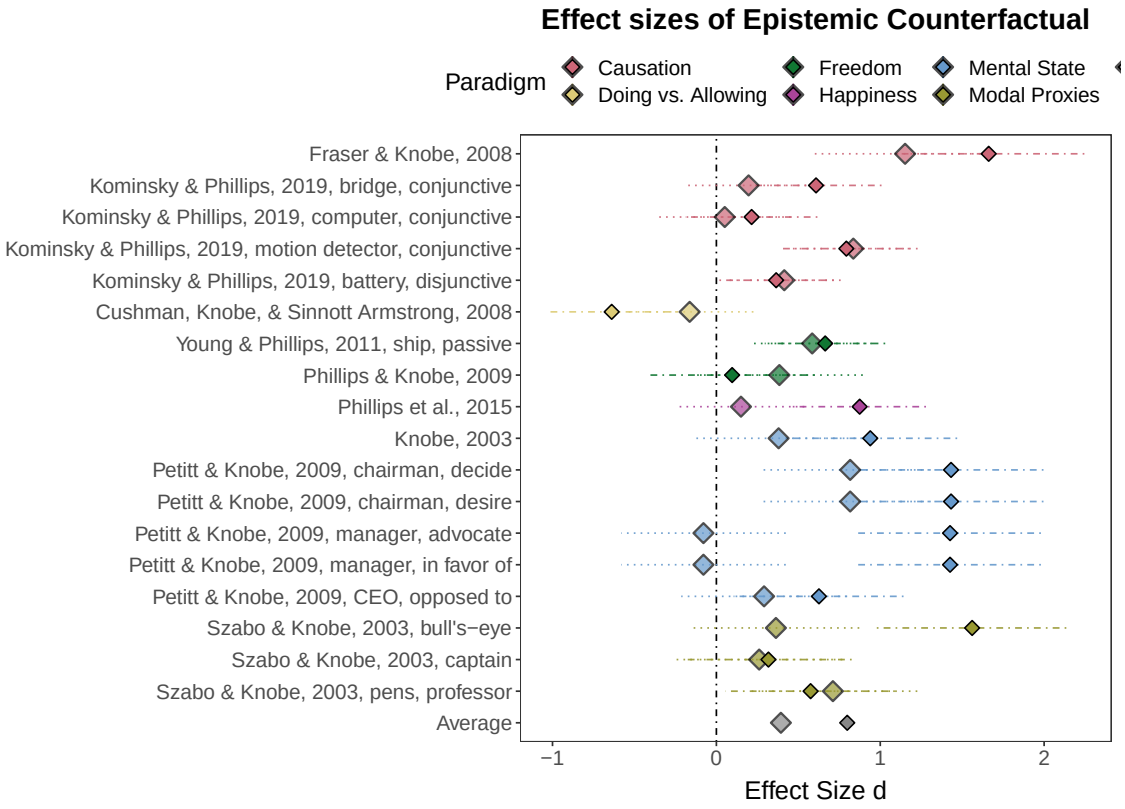

Figure 4. **Epistemic Counterfactual Effect Sizes by Study**: Effect sizes of Relevance ratings of the Epistemic State Counterfactual in the original norm violation condition (neutral vs. knowing norm-violation) are marked by large rhombuses, effect sizes in the new norm violation condition (neutral vs. ignorant norm-violation) are marked by small rhombuses. Error bars depict 95% confidence intervals of effect sizes

counterfactual possibilities: the agent's action, and the agent's knowledge state. When thinking about what could have gone differently, we found that people rated it highly relevant to consider a change in the agent's action when the agent was knowledgeable about the norm, and less relevant when they were ignorant. The epistemic state factor hence proved to be crucial for engaging in counterfactual reasoning about alternatives to the agent's action.

In contrast, we did not find such an effect on people's counterfactual reasoning about the agent's knowledge state. People found it relevant to consider what the agent could have known, independent of whether the agent knew about the moral status of their action, or was ignorant about it. Given that the manipulation of normality was centered on the focal agent's action, the fact that the epistemic condition selectively affected counterfactual reasoning about actions might not come as a surprise. In line with Kominsky and Phillips (2019)'s expectation-based normality account, the agent's ignorance of the immoral status of their behaviour might have diminished people's perception of the abnormality of the

action, rendering an alternative scenario in which the ignorant agent refrains from acting less relevant. The perceived normality of the epistemic state in contrast, that is, whether it is normal or abnormal for the agent to not know about the norm, was less clear-cut. People might not have had the intuition that the agent's lack of knowledge about the norm was abnormal, and hence not made a difference in what they thought a knowing vs. ignorant agent could have known.

Moreover, the experimental scenarios in our study were rather unspecific with regards to what the relevant possibilities were in which the agent could have learned that their action would violate a norm. Consider again the "ignorant norm-violation" condition from Phillips and Knobe (2009)'s ship scenario described previously. The vignette leaves open whether there was an opportunity for the captain to learn about the secret passengers in the cargo compartment. This has likely led to individual assumptions about the mutability of the agent's ignorance (Kirfel & Lagnado, 2021b), whether the agent could have done something to acquire the relevant knowledge (Kirsh & Maglio, 1994) or whether the agent was in some sense negligent (Murray, Krasich, Irving, Nadelhoffer, & De Brigard, 2022; Murray, Murray, Stewart, Sinnott-Armstrong, & De Brigard, 2019). All these assumptions will have affected how people responded to the counterfactual relevance question about the agent's epistemic state ("... what the agent could have known."), and potentially led to unaccounted for variation in these ratings. A more fine-tuned manipulation of the conditions that led up to an agent's ignorance could provide further insight into how people reason counterfactually about epistemic states.

General Discussion

Studies show that norm violations influence a wide range of domains, including judgments of causation, freedom, happiness, doing vs. allowing, mental state ascriptions, and modal claims. A continuing debate centers on why normality has such a pervasive impact, and whether one should attempt to offer a unified explanation of these various effects (Hindriks, 2014). In this study, we found evidence that the epistemic state of norm-violating agents plays a fundamental role in the impact of norms on non-normative judgments. Across a wide range of intuitive judgments and highly different manipulations of an agents' knowledge, we found that the impact of normality on non-normative judgments was diminished when the agent did not know that they were violating a norm. More precisely, the agent's knowledge of the norm violation determined the extent to which abnormal actions increased judgments of causation, decreased attribution of force, increased attributions of intentional action, and so on. In other words, the impact of ignorance appears to be as pervasive as the impact of normality itself. In addition, our study showed that the agent's epistemic state also influenced to what extent people engage in reasoning about alternatives to the agent's action. If the agent was ignorant when they violated a norm, people were less inclined to consider what the agent could have done differently.

At the broadest level, the current results provide evidence that the pervasive impact of normality likely warrants a unified explanation at some level: we considered a specific feature that had been shown to moderate the impact of normality in one domain (causation) and demonstrated that this same feature of the impact of normality can be found across a wide range of other domains. This finding suggests that the impact of norms arises from a shared underlying mechanism that is recruited across domains. Specific accounts may,

of course, seek to incorporate agents' epistemic states into their respective theory of how normality influences judgments in one particular domain. However, such an approach will miss out on a generalisation and will necessarily be less parsimonious. Accordingly, we turn now to considering two broad approaches to offering a unified account of the pervasive impact of ignorance.

Motivated Cognition, Counterfactuals, and Ignorance

Motivated Moral Cognition

On the one hand, blame-based accounts may try and use this discovery to their advantage by arguing that an agent's knowledge is directly relevant to whether they should be blamed (Cushman et al., 2008; Cushman, Sheketoff, Wharton, & Carey, 2013; Laurent, Nuñez, & Schweitzer, 2015; Yuill & Perner, 1988), and thus that these effects reflect that the impact of normality arises from the motivation to blame or hold agents responsible for their actions (Alicke & Rose, 2012; Livengood et al., 2017; Samland et al., 2016; Samland & Waldmann, 2016). For example, the tendency to report that agents who bring about harm acted intentionally may serve to corroborate people's desire to judge the agent's behaviour negatively (Nadelhoffer, 2004; Rogers et al., 2019). Motivated accounts differ in terms of exactly which moral judgment is argued to be at stake, i.e. whether norm-violations elicit a desire to punish (Clark et al., 2014), to blame (Alicke & Rose, 2012; Hindriks et al., 2016), to hold accountable (Samland & Waldmann, 2016) or responsible (Sytsma, 2020a), and whether its influence works in form of a cognitive bias (Alicke, 2000), or a more affective response (Nadelhoffer, 2004). Common to all, however, is the assumption that it is the impetus to morally condemn the norm-violating agent that underlies exaggerated attributions of specific properties, from free will to intentional action.

Our study puts an important constraint on how the normative judgment that motivated reasoning accounts assume might work. To account for our findings, motivated accounts cannot generally appeal to whether an agent's action violated a clear norm, but have to take into account whether people would all-things-considered blame the agent (Driver, 2017). In that sense, the mere violation of a norm must not, itself, suffice to trigger the relevant blame response. Rather, the perception of this norm violation must occur in conjunction with an assessment of the epistemic state of the agent such that the relevant motivated reasoning is only elicited when the agent is *aware* of the immorality of their action. For example, Alicke and Rose's 2012 Culpable Control Model holds that immediate negative evaluative reactions of an agent's behaviours often cause people to interpret all other agential features in a way that justifies blaming the agent. Such accounts face a challenge. On the one hand, they seem committed to the idea that people should discount the agent's ignorance to support their immediate negative evaluation of the harm causing actions. On the other hand, they need to account for the fact that people seem to be sensitive to fine-grained epistemic features of the agent when forming their negative evaluation of the harm causing action.

Importantly, studies show that different types of moral judgments are differently sensitive to the epistemic states factor (Cushman, 2008; Cushman et al., 2013). At minimum, a unified motivated account needs to specify the *type* of moral judgment they take to modulate non-moral cognition across a vast array of domains, and under which conditions motivated

reasoning is elicited.

In addition, the motivated reasoning account still faces the challenge of explaining the similar impact of descriptive rather than prescriptive norms (Gerstenberg & Icard, 2020; Kominsky et al., 2015; Morris, Phillips, Gerstenberg, & Cushman, 2019): Agents who deviate from typical or usual behaviour are judged as having acted more intentionally (Uttich & Lombrozo, 2010), are seen as more causal (Icard et al., 2017; Kirfel & Lagnado, 2021a; Kominsky et al., 2015), are attributed more free will (Bernhard et al., 2022; Monroe & Ysidron, 2021) or negative emotions such as regret (Fillon, Lantian, Feldman, & N'Gbala, 2022; Kahneman & Miller, 1986), and so on. Critically, more recently, the effect of ignorance has also been shown to moderate the effect of descriptive norm violations on causal judgments in an identical fashion (Kirfel & Lagnado, 2021a). Some have suggested that statistical norm-deviances and atypical behaviour might likewise influence responsibility judgements (Sytsma, 2020a; Sytsma, Livengood, & Rose, 2012). In that sense, the attribution of causality to a normally or abnormally acting agent simply tracks people's responsibility judgements (Sytsma, 2020a, in press), often because such behaviours also covary in or give raise to inferences about more moral features of that action (Livengood et al., 2017; Sytsma, 2019, 2020a). Such accounts however, still face the challenge to explain the role of agent ignorance in case of descriptive norms, such as the finding that people do *not* increase causal attributions to atypically acting agents who at the same time are ignorant about the typicality of other people's behaviour (Kirfel & Lagnado, 2021a).

Counterfactuals

An alternative approach would be to extend a unified counterfactual-based account to explaining the pervasive impact of normality (Phillips & Knobe, 2018; Phillips et al., 2015). At the heart of this proposal lies the assumption that in the course of making a judgment about a certain state of affairs or action, people consider alternative events (or "counterfactuals", Lewis, 1974) in which certain aspects of the actual situation occurred differently. Judgments in the aforementioned domains are influenced by the degree to which people regard certain alternative possibilities as relevant (Phillips et al., 2015). That is, judging whether an agent is forced to act, or whether an agent caused an outcome is shaped by the alternatives people consider as relevant, e.g., an alternative scenario in which the agent refrains from acting. Norms or normality, it is argued, influence the perceived relevance of certain counterfactuals (Hitchcock & Knobe, 2009; Kahneman & Miller, 1986; Lewis, 1974). People tend to consider alternative scenarios in which things go "normally", broadly construed (Bear & Knobe, 2017). This tendency subsequently affects the degree to which they see a norm-violating agent's action as intentional, forced, etc. (Phillips et al., 2015). While counterfactual accounts have been spelled out in slightly different frameworks (Icard et al., 2017; Knobe & Szabó, 2013; Morris et al., 2018; Phillips, Morris, & Cushman, 2019), the key point is that norms do not influence judgments via a specifically *moral* cognitive mechanism, but by affecting the relevance of alternative possibilities (Phillips & Knobe, 2018), and thus also their likelihood of coming to mind (Phillips et al., 2019). This distinction in what underlies the influence of norms marks the key difference between counterfactual and motivated cognition accounts. A norm-violating agent is judged as less forced to act than a norm-abiding agent not because of people's desire to blame the former, but because of their tendency to represent alternative scenarios in which the agent acts

normally (Bernhard et al., 2022).

Counterfactual accounts provide a perfectly general account for the influence of a variety of different kinds of norms, such as descriptive norms (Gerstenberg & Icard, 2020; Icard et al., 2017; Kominsky et al., 2015) and rational norms (Phillips, Young, & Gerstenberg, 2022). One challenge for these theories, however, is to explain how and why these epistemic states play the correct role in shaping counterfactual reasoning. Some progress on this has been made in the domain of causal judgments (Kirfel & Lagnado, 2021b; Kominsky & Phillips, 2019). Kominsky and Phillips (2019) argue that the epistemic state of a norm-violating agent modulates the perceived abnormality of their action. If the agent lacks the knowledge that their action is norm-violating, the action is not perceived as abnormal in the first place. Put another way, people may take into account what would be a normal action, given only what the agent knows about the world. On such an account, ignorance moderates perceptions of normality. Counterfactual accounts hold that when people make judgments that rely on modal cognition, they construct and reason over a set of possible actions (Morris, Phillips, Huang, & Cushman, 2021; Phillips & Knobe, 2018; Phillips et al., 2019). However, instead of entertaining all possible actions that an agent could have undertaken in a given situation, it is assumed that people sample from a small subset of actions (Icard, 2016; Icard et al., 2017), and that this sampling is biased toward possibilities that are statistically and/or prescriptively normal. Thus, in line with Kominsky and Phillips (2019)'s proposal, if people do not perceive an agent's action as abnormal when they unknowingly violate a norm, then they are less likely to sample a possibility in which the agent does not violate that norm. Accordingly, judgments that rely on counterfactuals should be less influenced by ignorant norm violations.

However, it is also possible that agents' epistemic states actually play broader role in reasoning about alternatives. Rather than merely affecting sampling propensity via normality, an agent's knowledge may more directly restrict the set of actions that people sample from. Current models of modal cognition propose that the first step in sampling possible actions involves partitioning the space of all possible actions, and focusing only on ones that concern the actual situation being faced (Phillips et al., 2019): We do not consider the possibility of ordering ice cream when reasoning about what could do when trying to save a ship from sinking. Thus, one possibility is that an agent's knowledge of the situation they are in may guide the way in which the space of actions is partitioned, such that we only consider actions that involve the agent's understanding of the situation (rather than the situation itself): if the agent is not aware of ballast in the bottom of their ship, we do not consider actions involving the agent throwing ballast overboard to save his ship from sinking. In this respect, the agent's knowledge of the world at a given time-point would constrain the initial set of relevant actions that people sample from. While our results show that people's counterfactual reasoning about actions is indeed sensitive to epistemic states, future research will need to investigate how exactly epistemic factors affect the counterfactual reasoning process.

A separate challenge faced by a unified counterfactual account is to explain the role of possibilities in each of the judgments where normality has been found to have an impact. While the relevant work has again been done in many of the cases—see e.g., the work by Knobe and Szabó (2013) and Phillips et al. (2015)—there are others where the connection to counterfactuals is less clear, as in the case of assessments of happiness (Phillips, Mis-

enheimer, & Knobe, 2011). A unified counterfactual framework would have to show that norm effects in such domains rely on thinking about alternative possibilities, too. Thus far, the extant literature on happiness has largely tried to explain the decrease in attributions of happiness to immoral agents without referring to alternative possibilities. Instead, these accounts have focused on the idea that immoral agent's positive feelings may not be aligned with their true 'moral' selves (Newman et al., 2015; Yang, Knobe, & Dunham, 2021), or that the concept of happiness may simply be partly evaluative (Phillips et al., 2014). To the extent that counterfactual accounts want to offer a fully unified explanation of the impact of normality, more work needs to be done to extend this approach to domains that have not been traditionally linked to counterfactual thinking.

Related Proposals and Future Directions

The function of agentive epistemic states has been addressed in theoretical frameworks outside of motivated and counterfactual reasoning. In response to the findings from our study 1 (Kirfel & Phillips, 2021), Crutchfield and Scheall (2021) have put forward the idea that an agent's "epistemic burden" constrains the expectations we as observers have towards their behaviour. An epistemic burden is defined as "... the weight, in terms of missing knowledge and capacity, that [an agent] need[s] to somehow overcome (or heft and carry) in order to bring deliberate realization of the goal entirely." (Scheall & Crutchfield, 2021). The epistemic burden of an agent towards knowing a norm determines whether we hold expectations towards their norm-compliance, and hence how strongly we judge that agent if they don't follow the norm (Crutchfield & Scheall, 2021). If the agent is completely ignorant about a norm, acting in a norm-confirming manner will be ranked low among their action preferences, and expectations towards such an action will hence be attenuated. Crutchfield and Scheall (2021)'s account has the advantage that it can capture different epistemic states and difficulty levels of knowledge acquisition, and they explicitly include non-propositional knowledge such as abilities and capacities into their account of epistemic burdens. While they take people's judgments to be driven by reasoning about alternatives, how exactly their theory of epistemic burdens can be integrated into frameworks of modal cognition still needs to be worked out. Moreover, people might still hold strong expectations towards certain norm-conforming behaviours if an agent's epistemic burden is avoidable or self-inflicted (Kirfel & Lagnado, 2021b; Peels, 2011).

Some progress in accounting for the role of epistemic states has also been made in the area of agentive modals (Carr, 2017; Mandelkern, Schultheis, & Boylan, 2017). Mandelkern et al. (2017) posit an account of ability modals (e.g. "John *can* go swimming this evening") that integrates epistemic facts. An agent is said to be able to φ if there is some practically available action such that if they tried to do that action, they would succeed in φ-ing. Crucially, the agent's epistemic state concerning their actual situation constrains what counts as a practically available action. The account captures the intuition about ability ascriptions that a claim like "John can go swimming this evening." is likely to be false if John does not know how to swim, but also if, e.g., John does not know of any available swimming pools in his area. In the latter case, Mandelkern et al. (2017)'s account holds that while there is an objective reading in which John might physically be able to swim, there is also subjective reading such that John's practical situation is limited to what he knows about his actual situation. The appeal to the notion of practical availability as suggested

in Mandelkern et al. (2017) might provide a useful starting point for accounts that seek to provide a general modal account of the impact of norms and ignorance. Ultimately, future research might benefit from refining the exact epistemic conditions that reduce the impact of normality, as well as probing their influence on further non-moral judgment domains.

Conclusion

In this large-scale experimental study, we find evidence for a pervasive impact of ignorance: the impact of norm violations on non-normative judgments depends largely on the agent knowing that they were violating a norm when acting. Moreover, we find that the reduction in the impact of normality is underpinned by people's counterfactual reasoning. While the influence of norms has given rise to a variety of judgment-specific accounts, two major theories aim to explain a large amount of the norm literature by a more unified framework. Both motivated cognition as well as counterfactual reasoning theories provide an account of the role of norms in non-moral cognition, and crucially, might hold the potential to extend their framework for the moderating role of epistemic states. Whichever of these theories turns out to be correct in the end, this work should inspire a new target on the impact of normality, since one needs not only to explain the pervasive impact of norms, but also the pervasive impact of ignorance.

Acknowledgements

We would like to thank Molly McQuoid for her help with the experiments and editing this manuscript.

References

Adams, F., & Steadman, A. (2007). Folk concepts, surveys, and intentional action. In C. Lumer & S. Nannini (Eds.), *Intentionality, deliberation, and autonomy: The action-theoretic basis of practical philosophy* (pp. 17–33). Ashgate Publishing.

Alfano, M., Beebe, J. R., & Robinson, B. (2012). The centrality of belief and reflection in knobe-effect cases: A unified account of the data. *The Monist*, *95*(2), 264–289.

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological bulletin*, *126*(4), 556-574.

Alicke, M. D., & Rose, D. (2012). Culpable control and causal deviance. *Social and Personality Psychology Compass*, *6*(10), 723–735.

Alicke, M. D., Rose, D., & Bloom, D. (2011). Causation, norm violation, and culpable control. *The Journal of Philosophy*, *108*(12), 670–696.

Bear, A., & Knobe, J. (2017). Normality: Part descriptive, part prescriptive. *Cognition*, *167*, 25–37.

Bernhard, R. M., LeBaron, H., & Phillips, J. (2022). It's not what you did, it's what you could have done. *Cognition*, *228*, 105-222.

Carr, J. (2017). Deontic modals. In T. McPherson & D. Plunkett (Eds.), *The routledge handbook of metaethics* (pp. 194–210). Routledge.

Clark, C. J., Luguri, J. B., Ditto, P. H., Knobe, J., Shariff, A. F., & Baumeister, R. F. (2014). Free to punish: a motivated account of free will belief. *Journal of personality and social psychology*, *106*(4), 501-513.

Clark, C. J., Winegard, B. M., & Shariff, A. F. (2021). Motivated free will belief: The theory, new (preregistered) studies, and three meta-analyses. *Journal of Experimental Psychology: General*, *150*(7), e22–e47.

Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Cova, F., Lantian, A., & Boudesseul, J. (2016). Can the knobe effect be explained away? methodological controversies in the study of the relationship between intentionality and morality. *Personality and Social Psychology Bulletin*, *42*(10), 1295–1308.

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., ... others (2021). Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology*, *12*(1), 9–44.

Crutchfield, P., & Scheall, S. (2021). A unified account of the effects of norm violation on various judgments. *pre-print*.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353–380.

Cushman, F., Knobe, J., & Sinnott-Armstrong, W. (2008). Moral appraisals affect doing/allowing judgments. *Cognition*, *108*(1), 281–289.

Cushman, F., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. *Cognition*, *127*(1), 6–21.

Díaz, R., & Reuter, K. (2021). Feeling the right way: Normative influences on people's use of emotion concepts. *Mind & language*, *36*(3), 451–470.

Driver, J. (2017). Wronging, blame, and forgiveness. In D. Shoemaker (Ed.), (Vol. 4, p. 206-2018). Oxford University Press.

Feltz, A. (2007). The knobe effect: A brief overview. *The Journal of Mind and Behavior*, 265–277.

Fillon, A., Lantian, A., Feldman, G., & N'Gbala, A. (2022). Exceptionality effect in agency attributions: Exceptional behaviors are perceived as higher free will than routine behaviors. *International Review of Social Psychology*, *35*(1).

Fincham, F. D. (1985). Outcome valence and situational constraints in the responsibility attributions of children and adults. *Social Cognition*, *3*(2), 218-233.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review*, *128*(5), 936.

Gerstenberg, T., & Icard, T. (2020). Expectations affect physical causation judgments. *Journal of Experimental Psychology: General*, *149*(3), 599.

Gilbert, E. A., Tenney, E. R., Holland, C. R., & Spellman, B. A. (2015). Counterfactuals, control, and causation: Why knowledgeable people get blamed more. *Personality and Social Psychology Bulletin*, *41*(5), 643–658.

Guglielmo, S., & Malle, B. F. (2010). Can unintended side effects be intentional? resolving a controversy over intentionality and morality. *Personality and social psychology bulletin*, *36*(12), 1635–1647.

Hendrick, C. (1990). Replications, strict replications, and conceptual replications: are they important? *Journal of Social Behavior and Personality*, *5*(4), 41-49.

Hindriks, F. (2014). Normativity in action: how to explain the knobe effect and its relatives. *Mind & Language*, *29*(1), 51–72.

Hindriks, F., Douven, I., & Singmann, H. (2016). A new angle on the knobe effect: Intentionality correlates with blame, not with praise. *Mind & Language*, *31*(2), 204–220.

Hitchcock, C., & Knobe, J. (2009). Cause and norm. *The Journal of Philosophy*, *106*(11), 587–612.

Icard, T. (2016). Subjective probability as sampling propensity. *Review of Philosophy and Psychology*, *7*(4), 863–903.

Icard, T., Kominsky, J., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, *161*, 80–93.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological review*, *93*(2), 136.

Kahnemann, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahnemann, P. Slovic, & A. Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases.* Cambridge University Press.

Kirfel, L., & Lagnado, D. (2021a). Causal judgments about atypical actions are influenced by agents' epistemic states. *Cognition*, *212*, 104721. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0010027721001402` doi: https://doi.org/10.1016/j.cognition.2021.104721

Kirfel, L., & Lagnado, D. (2021b, Oct). *Changing minds — epistemic interventions in causal reasoning.* PsyArXiv. Retrieved from `psyarxiv.com/db6ms` doi: 10.31234/osf.io/db6ms

Kirfel, L., & Phillips, J. (2021). The impact of ignorance beyond causation: An experimental meta-analysis. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43, p. 1595-1601).

Kirsh, D., & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive science*, *18*(4), 513–549.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*(3), 190–194.

Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences*, *33*(4), 315–329. doi: 10.1017/S0140525X10000907

Knobe, J. (2022). Morality and possibility. In M. Vargas & J. Doris (Eds.), (p. 310-333). Oxford University Press.

Knobe, J., & Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. *Moral psychology*, *2*, 441–448.

Knobe, J., & Szabó, Z. G. (2013). Modals with a taste of the deontic. *Semantics and Pragmatics*, *6*, 1–43.

Kominsky, J. F., & Phillips, J. (2019). Immoral professors and malfunctioning tools: Counterfactual relevance accounts explain the effect of norm violations on causal selection. *Cognitive science*, *43*(11), e12792.

Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D., & Knobe, J. (2015). Causal superseding. *Cognition*, *137*, 196–209.

Laurent, S. M., Nuñez, N. L., & Schweitzer, K. A. (2015). The influence of desire and knowledge on perception of each other and related mental states, and different mechanisms for blame. *Journal of Experimental Social Psychology*, *60*, 27–38.

Laurent, S. M., Reich, B. J., & Skorinko, J. L. (2019). Reconstructing the side-effect

effect: A new way of understanding how moral considerations drive intentionality asymmetries. *Journal of Experimental Psychology: General*, *148*(10), 1747-1766.

Lewis, D. (1974). Causation. *The journal of philosophy*, *70*(17), 556–567.

Lindauer, M., & Southwood, N. (2021). How to cancel the knobe effect. *American Philosophical Quarterly*, *58*(2), 181–186.

Livengood, J., Sytsma, J., & Rose, D. (2017). Following the fad: Folk attributions and theories of actual causation. *Review of Philosophy and Psychology*, *8*(2), 273–294.

Lombrozo, T., & Uttich, K. (2010). Putting normativity in its proper place. *Behavioral and Brain Sciences*, *33*(4), 344-345.

Lynch Jr, J. G., Bradlow, E. T., Huber, J. C., & Lehmann, D. R. (2015). Reflections on the replication corner: In praise of conceptual replications. *International Journal of Research in Marketing*, *32*(4), 333–342.

Machery, E. (2008). The folk concept of intentional action: Philosophical and experimental issues. *Mind & Language*, *23*(2), 165–189.

Machery, E., Grau, C., & Pury, C. L. (2020). Love and power: Grau and pury (2014) as a case study in the challenges of x-phi replication. *Review of Philosophy and Psychology*, *11*(4), 995–1011.

Mallon, R. (2008). Knobe versus machery: testing the trade-off hypothesis. *Mind & Language*, *23*(2), 247–255.

Mandelkern, M., Schultheis, G., & Boylan, D. (2017). Agentive modals. *The Philosophical Review*, *126*(3), 301–343.

Monroe, A. E., & Ysidron, D. W. (2021). Not so motivated after all? three replication attempts and a theoretical challenge to a morally motivated belief in free will. *Journal of Experimental Psychology: General*, *150*(1), e1-e12.

Morris, A., Phillips, J., Gerstenberg, T., & Cushman, F. (2019). Quantitative causal selection patterns in token causation. *PloS one*, *14*(8), e0219704.

Morris, A., Phillips, J., Huang, K., & Cushman, F. (2021). Generating options and choosing between them depend on distinct forms of value representation. *Psychological Science*, 1731–1746.

Morris, A., Phillips, J., Icard, T., Knobe, J., Gerstenberg, T., & Cushman, F. (2018). *Judgments of actual causation approximate the effectiveness of interventions.* PsyArXiv.

Murray, S., Krasich, K., Irving, Z., Nadelhoffer, T., & De Brigard, F. (2022). Mental control and attributions of blame for negligent wrongdoing. *Journal of Experimental Psychology: General*. doi: 10.1037/xge0001262

Murray, S., Murray, E. D., Stewart, G., Sinnott-Armstrong, W., & De Brigard, F. (2019). Responsibility for forgetting. *Philosophical Studies*, *176*(5), 1177–1201.

Nadelhoffer, T. (2004). Blame, badness, and intentional action: a reply to knobe and mendlow. *Journal of Theoretical and Philosophical Psychology*, *24*(2), 259–269.

Nadelhoffer, T. (2006). Bad acts, blameworthy agents, and intentional actions: Some problems for juror impartiality. *Philosophical explorations*, *9*(2), 203–219.

Newman, G. E., De Freitas, J., & Knobe, J. (2015). Beliefs about the true self explain asymmetries based on moral judgment. *Cognitive science*, *39*(1), 96–125.

Ngo, L., Kelly, M., Coutlee, C. G., Carter, R. M., Sinnott-Armstrong, W., & Huettel, S. A. (2015). Two distinct moral mechanisms for ascribing and denying intentionality. *Scientific reports*, *5*(1), 1–11.

Nichols, S., & Ulatowski, J. (2007). Intuitions and individual differences: The knobe effect revisited. *Mind & Language*, *22*(4), 346–365.

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., ... others (2022). Replicability, robustness, and reproducibility in psychological science. *Annual review of psychology*, *73*, 719–748.

Pearl, J. (2009). *Causality*. Cambridge university press.

Peels, R. (2011). Tracing culpable ignorance. *Logos & Episteme*, *2*(4), 575–582.

Pettit, D., & Knobe, J. (2009). The pervasive impact of moral judgment. *Mind & language*, *24*(5), 586–604.

Phillips, J., De Freitas, J., Mott, C., Gruber, J., & Knobe, J. (2017). True happiness: The role of morality in the folk concept of happiness. *Journal of Experimental Psychology: General*, *146*(2), 165-181.

Phillips, J., & Knobe, J. (2009). Moral judgments and intuitions about freedom. *Psychological Inquiry*, *20*(1), 30–36.

Phillips, J., & Knobe, J. (2018). The psychological representation of modality. *Mind & Language*, 65–94. Retrieved from `http://dx.doi.org/10.1111/mila.12165` doi: 10.1111/mila.12165

Phillips, J., Luguri, J. B., & Knobe, J. (2015). Unifying morality's influence on non-moral judgments: The relevance of alternative possibilities. *Cognition*, *145*, 30–42.

Phillips, J., Misenheimer, L., & Knobe, J. (2011). The ordinary concept of happiness (and others like it). *Emotion Review*, *3*(3), 320–322.

Phillips, J., Morris, A., & Cushman, F. (2019). How we know what not to think. *Trends in cognitive sciences*, *23*(12), 1026–1040.

Phillips, J., Nyholm, S., & Liao, S.-y. (2014). The good in happiness. In T. Lombrozo, S. Nichols, & J. & Knobe (Eds.), *Oxford studies in experimental philosophy: Volume 1* (pp. 253–293). Oxford University Press.

Phillips, J., Young, L., & Gerstenberg, T. (2022). Normal causation. *unpublished manuscript*.

Pinillos, N. Á., Smith, N., Nair, G. S., Marchetto, P., & Mun, C. (2011). Philosophy's new challenge: Experiments and intentional action. *Mind & Language*, *26*(1), 115–139.

Pizarro, D. A., & Tannenbaum, D. (2012). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. In M. Mikulincer & P. R. Shaver (Eds.), *The social psychology of morality: Exploring the causes of good and evil* (p. 91–108). American Psychological Association.

Pizarro, D. A., Tannenbaum, D., & Uhlmann, E. (2012). Mindless, harmless, and blameworthy. *Psychological Inquiry*, *23*(2), 185–188.

Rogers, R., Alicke, M. D., Taylor, S. G., Rose, D., Davis, T. L., & Bloom, D. (2019). Causal deviance and the ascription of intent and blame. *Philosophical Psychology*, *32*(3), 404–427.

Samland, J., Josephs, M., Waldmann, M. R., & Rakoczy, H. (2016). The role of prescriptive norms and knowledge in children's and adults' causal selection. *Journal of Experimental Psychology: General*, *145*(2), 125.

Samland, J., & Waldmann, M. R. (2016). How prescriptive norms influence causal inferences. *Cognition*, *156*, 164–176.

Scheall, S., & Crutchfield, P. (2021). The priority of the epistemic. *Episteme*, *18*(4),

726–737.
Schwenkler, J., & Sytsma, J. (2020). Reversing the norm effect on causal attributions. *pre-print*.
Sher, G. (2009). *Who knew?: Responsibility without awareness*. Oxford University Press.
Sheu, C.-F., & Suzuki, S. (2001). Meta-analysis using linear mixed models. *Behavior Research Methods, Instruments, & Computers*, *33*(2), 102–107.
Spellman, B. A., & Gilbert, E. A. (2014). Blame, cause, and counterfactuals: The inextricable link. *Psychological Inquiry*, *25*(2), 245–250.
Strickland, B., & De Cruz, H. (2021). Replicability in cognitive science. *Review of Philosophy and Psychology*, *12*(1), 1–7.
Stuart, M. T., Colaço, D., & Machery, E. (2019). P-curving x-phi: Does experimental philosophy have evidential value? *Analysis*, *79*(4), 669–684.
Sytsma, J. (2019). The character of causation: Investigating the impact of character, knowledge, and desire on causal attributions. *pre-print*.
Sytsma, J. (2020a). Causation, responsibility, and typicality. *Review of Philosophy and Psychology*, 1–21.
Sytsma, J. (2020b). Resituating the influence of relevant alternatives on attributions. *pre-print*.
Sytsma, J. (in press). The responsibility account. In P. Willemsen & A. Wiegmann (Eds.), *Advances in experimental philosophy of causation.* Bloomsbury.
Sytsma, J., Livengood, J., & Rose, D. (2012). Two types of typicality: Rethinking the role of statistical typicality in ordinary causal attributions. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(4), 814–820.
Uttich, K., & Lombrozo, T. (2010). Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition*, *116*(1), 87–100.
Willemsen, P., & Kirfel, L. (2019). Recent empirical work on the relationship between causal judgements and norms. *Philosophy Compass*, *14*(1), e12562.
Yaffe, G. (2018). The point of mens rea: The case of willful ignorance. *Criminal Law and Philosophy*, *12*(1), 19–44.
Yang, F., Knobe, J., & Dunham, Y. (2021). Happiness is from the soul: The nature and origins of our happiness concept. *Journal of Experimental Psychology: General*, *150*(2), 276.
Young, L., & Phillips, J. (2011). The paradox of moral focus. *Cognition*, *119*(2), 166–178.
Young, L., & Tsoi, L. (2013). When mental states matter, when they don't, and what that means for morality. *Social and personality psychology compass*, *7*(8), 585–604.
Yuill, N., & Perner, J. (1988). Intentionality and knowledge in children's judgments of actor's responsibility and recipient's emotional reaction. *Developmental psychology*, *24*(3), 358.

Appendix A

Study Name	Statistic that showed Original Effect	Original Effect Size	Replicated Effect Size	Original N per Condition	N per Condition in our Study
Phillips & Knobe, 2009	$t(50) = 4.7, p < .001$	$d = 1.33$*	$d = 0.97$	ca. 26	ca. 86
Fraser & Knobe, 2008	$t(17) = 5.5, p < .001$	$d = 2.66$*	$d = 1.73$	ca. 9	ca. 89
Cushman, Knobe, & Sinnott Armstrong, 2008	$t(298) = 4.3, p < .001$	$d = 0.53$*	$d = 0.66$	ca. 150	ca. 110
Knobe, 2003	$\chi^2(1, N = 78) = 27.2, p < .001$	n.a., categorial response	$d = 2.95$	ca. 39	ca. 88
Phillips et al., 2015	$t = -5.27, p < .001$, $d = .66$	$d = 0.66$*	$d = 0.65$	ca. 50	ca. 110
Young and Phillips, 2011 (ship, active + passive)	active: $t(56) = 1.55, p = 0.128$, passive: $t(49) = 2.15, p = 0.037$	active: n.s. passive: $d = 0.61$	active: $d = 0.06$ (n.s.) passive $d = 0.46$	ca. 30	ca. 117
Young and Phillips, 2011 (doctor, active + passive)	active: t(58) = 1.93, p = 0.058, passive: $t(38) = 3.48, p = 0.001$	active = n.s passive: $d = 1.13$	active: $d = 0.15$ (n.s.) passive: $d = 0.21$ (n.s.)	ca. 25	ca. 118
Kominsky & Phillips, 2019, (battery, conjunctive + disjcuntive)	conjunctive: $t(80) = -0.95, p = 0.34$ disjunctive: $t(85) = 5.39, p < .005$	conjunctive: n.s. disjunctive: $d = 1.14$	conjunctive: $d = 0.04$ (n.s.) disjunctive: $d = 0.42$	ca. 40	ca. 104
Kominsky & Phillips, 2019, (bridge, conjunctive + disjcuntive)	conjunctive $t(41) = -11.83, p < .001$ disjunctive: $t(47) = -1.19, p = .23$	conjunctive $d = 2.69$ disjunctive = n.s.	conjunctive: $d = 0.86$ disjunctive: $d = 0.32$ (n.s.)	ca. 40	ca. 123
Kominsky & Phillips, 2019, (computer, conjunctive + disjunctive)	not included in analysis	n.a.	conjunctive: $d = 0.50$ disjunctive: $d = 0.27$ (n.s.)	ca. 40	ca. 105
Kominsky & Phillips, 2019, (motion detector, conjunctive + disjcuntive)	conjunctive: $t(73) = -5.59, p < .001$ disjunctive: $t(57) = -1.31, p = .19$	conjunctive: $d = 1.22$ disjunctive: n.s.	conjunctive: $d = 0.37$ (n.s.) disjunctive: $d = 0.13$ (n.s.)	ca. 40	ca. 110
Petitt & Knobe, 2009 (chairman, desire/decide)	Decide: $t(35) = 2.4, p < .05$, desire: $t(298) = 9.52, p < .0001$ (from Ditto & Pizarro, 2007)	decide: $d = 0.81$* desire: $d = 1.03$*	decide: $d = 2.11$ desire: $d = 1.85$	desire: ca. 17 decide: ca. 165	ca. 87
Petitt & Knobe, 2009 (assistant manager, advocate/ in favor of))	$F(1, 58) = 4.6, p < .05$	$d = .56$**	advocate: $d = -1.38$, in favour of: $d = -1.35$	ca. 16	ca. 90
Petitt & Knobe, 2009 (CEO, opposed to)	$t(54) = 2.0, p < .05$	$d = .54$*	$d = 2.60$	ca. 28	ca. 89
Petitt & Knobe, 2009 (Bomb, intended to)	$t(35) = 2.5, p < .05$	$d = .84$*	$d = 0.33$ (n.s.)	ca. 16	ca. 88
Szabo & Knobe, 2003 (captain, modal proxies)	$t(40) = 7.9$, $p < .01$	$d = 2.49$*	$d = 1.23$	ca. 21	ca. 88
Szabo & Knobe, 2003 (pens, modal proxies)	$F(1, 76) = 43.6, p < .001$	$d = 1.51$**	$d = -0.77$	ca. 22	ca. 88
Szabo & Knobe, 2003 (bull's eye, modal proxies)	$t(50) = -7.1, p < .001$	$d = 2.00$*	$d = 1.65$	ca. 26	ca. 88
Uttich and Lobrozo, 2010 (gizmos)	$F(1, 288) = 12.828, p < .01$	$d = 2.00$***	$d = 0.09$ (n.s.)	ca. 25	ca. 120
Uttich and Lobrozo, 2010 (scrubs)	$F(1, 288) = 12.828, p < .01$	$d = 2.00$***	$d = 0.12$ (n.s.)	ca. 25	ca. 117
Uttich and Lobrozo, 2010 (truck trailers)	$F(1, 288) = 12.828, p < .01$	$d = 2.00$***	$d = 0.24$ (n.s.)	ca. 25	ca. 118

Table 1

Overview of Original and Replicated Effect Sizes ** marks effect sizes estimated from test statistics, ** marks interaction effects, *** marks effect sizes averaged across multiple scenarios*