The Knobe effect as an instance of a severity effect

Jan García Olier, University of Zurich
Markus Kneer, University of Zurich

1. Introduction

Following Joshua Knobe (2003), an avalanche of empirical studies in the last two decades have found an asymmetry in people's attributions of intentionality, also known as the 'Knobe effect' or 'side-effect effect' (cf. Knobe, 2003a, 2003b, 2004a; Cushman & Mele, 2008; McCann, 2005; Knobe & Mendlow, 2004; Knobe & Burra, 2006; Nadelhoffer, 2004a, 2004b, 2005; Leslie, Knobe & Cohen, 2006; Pettit & Knobe, 2009; for detailed reviews see Feltz, 2007; Nichols & Ulatowski, 2007; Cova, 2016). In Knobe's (2003) original experimental design, participants were presented with either a *harm* or a *help* condition of a CHAIRMAN scenario. In the scenario, the chairman of a company is told by one of his advisors that implementing a new program will increase their profits but also bring about harmful/helpful side effects on the environment. The chairman responds that he only cares about profits, not the environment, and implements the new program. As expected, profits increase, and the environment is harmed/helped. Participants then assessed whether the agent in said scenario acted intentionally with regards to such harmful or helpful side effects. This is where the asymmetry arises: the foreseen harmful side effects are deemed intentional by the majority of people, whereas the foreseen helpful side effects are deemed unintentional.

The Knobe effect has proven robust and pervasive. It has been replicated using scenarios other than the CHAIRMAN scenario (cf. *inter alia* Knobe & Mendlow, 2004; Adams & Steadman, 2004a; Nadelhoffer, 2004a, 2004b; Knobe, 2007) and on different age groups (e.g., Leslie, Knobe & Cohen, 2006, replicating the Knobe effect on pre-school children) and cultures (Knobe & Burra, 2006; Cova & Naar, 2012; for a study on cross-cultural variability see Robbins, Shepard & Rochat, 2017). Importantly, the Knobe effect has been found to affect the attribution of a wide range of mental states such as knowledge (cf. Beebe & Buckwalter, 2010; Beebe & Jensen, 2012) belief (Beebe, 2013), desire (Tannenbaum, Ditto, & Pizarro, 2007), and other psychological properties like being *in favor of* (Pettit & Knobe, 2009) and *in order to* attain a goal (Knobe, 2004a).

More recently, Kneer & Bourgeois-Gironde (2017) conducted an empirical study on the relation between an action's outcome and intentionality ascriptions. Instead of assigning participants to a harm or a help condition, Kneer & Bourgeois-Gironde ("K&B") assigned participants (in this occasion legal

E-mail addresses: jan.garciaolier@uzh.ch (J. García Olier), markus.kneer@gmail.com (M. Kneer)
University of Zurich, Center for Ethics, Zollikerstrasse 117, 8008 Zurich, Switzerland

professionals[1]) to either a *somewhat bad* or a *very bad* condition of a BEACH TOWN scenario. In the *somewhat bad* condition, the mayor of a beach town is told by one of his advisors that building a new highway connection will make traffic more efficient but there will also be minor adverse effects on the environment. The mayor responds that he only cares about making car traffic efficient, not the environment, and decides to build the highway connection. As expected, there are minor adverse effects on the environment. In the *very bad* condition, everything is identical except for the fact that the expected and materialized side effects on the environment are severe. Participants then had to assess whether the mayor intentionally harmed the environment. In this case, the average ascription of intentionality for the *very bad* condition was significantly higher than for the *somewhat bad*. The more harmful the foreseen outcome of an action, the more inclined people were to say that the action was intentional.[2]

The severity effect can be perceived as a challenge to standard explanations of the Knobe effect. The latter have focused on intentionality ascriptions for harmful v. helpful (or neutral) outcomes. The Knobe effect has thus been conceived as a binary, absolute effect: foreseen harmful side-effects are deemed intentional, whereas foreseen helpful — or at least not harmful — side-effects are deemed nonintentional. Nevertheless, the severity effect findings suggest that things might be somewhat more complex. As hinted at by K&B, the Knobe effect seems to be only a "special case" of a broader phenomenon. Rather than being of binary nature, the relation between an action's outcome and intentionality attributions appears to be a matter of *degrees*. This, as we will further discuss in detail, is at odds with a plethora of current explanations of the Knobe effect.[3]

The research conducted by K&B, however, suffers from a shortcoming. K&B focused only on intentionality ascriptions for graded *harmful* outcomes (somewhat bad v. very bad), but graded *helpful* outcomes were not tested. While there is empirical evidence concerning the relation between graded harmful outcomes and intentionality ascriptions, the relation between graded helpful outcomes and intentionality ascriptions remains unstudied. In this paper, we want to address this lacuna so as to work towards a clearer understanding of the relation between outcomes and intentionality ascriptions more generally.

---

[1] Kneer et al. (in prep.) replicate this study with laypeople and legal professionals from different cultural backgrounds.

[2] A recent challenge to the severity effect research has been raised. According to research conducted by Prochownik, et. al (2019), the asymmetrical ascriptions of intentionality across the somewhat bad and very bad conditions of Kneer & Bourgeois-Gironde's (2017) experimental design are not the result of a difference in outcome severity. The outcome in the somewhat bad condition, Prochownik, et. al (2020) argue, is seen as not harmful by a considerable number of participants. This suggests that the severity effect findings by Kneer & Bourgeois-Gironde's (2017) might just be another instance of the Knobe effect.

[3] In all fairness, the severity effect was introduced after most of the debate around the Knobe effect had already taken place. Hence, most of the accounts of the Knobe effect were not conceived to take the severity effect results into consideration.

2. The shape of the curve tracing the relation between intentionality ascriptions and graded outcomes

The introductory discussion has already drawn a distinction between *bivalent* v. *graded* views of the relation between outcomes and intentionality. According to the '*bivalent*' view, the feature that influences intentionality ascriptions is the *valence* of the outcome (positive/helpful v. negative/harmful). Bivalent accounts predict that foreseen, yet undesired outcomes will be judged intentional if and only if they are harmful. No matter how harmful a foreseen outcome is, people will say that it was intentional. Conversely, if the foreseen outcome is helpful (or perhaps simply not harmful), it will be judged as nonintentional. If we were to plot outcome severity on a 7-point scale (where 1 = very bad, and 7 = very good) against a proposed intentionality attribution (where 1 = complete disagreement, and 7 = complete agreement), we could expect something of the sort depicted by **Figure 1**.[4] Intentionality ascriptions, the bivalent view predicts, are above the scale's midpoint for harmful outcomes, and below the midpoint for helpful ones. Alternatively, one could hypothesize a broadly linear relation between intentionality ascriptions and outcome severity (or lack thereof). The propensity to ascribe intentionality would be positively correlated with the degree of harm of an outcome and negatively correlated with the degree of help of an outcome. **Figure 2** illustrates this '*graded-linear*' view. The more harmful a foreseen outcome, the higher the propensity to judge it intentional. Conversely, the more helpful the foreseen outcome, the lower the propensity to consider it intentional.

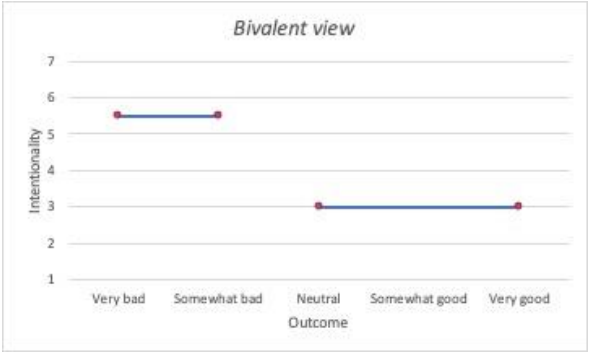

Figure 1. The *bivalent* view — red dots denote the Knobe effect.

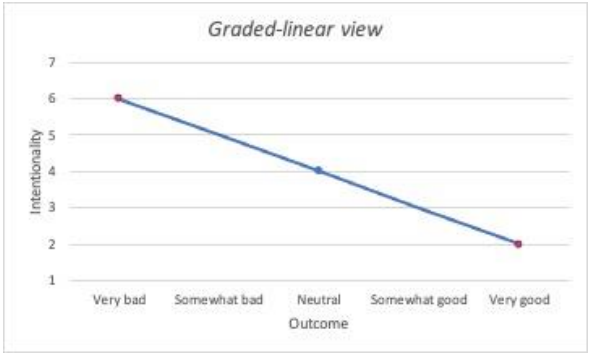

Figure 2. The *graded-linear* view — red dots denote the Knobe effect.

On one interpretation of the *Graded-linear* view, outcome valence (negative v. positive) does not play a central role. It might just happen to be the case that the shift from actions deemed intentional to nonintentional happens to be the point where valence switches from negative to positive. It is, however, also possible that the inflection point of outcome valence does play a central role. According

[4] For simplicity, and as displayed in Figure 1, in the bivalent view we group neutral outcomes with helpful ones, though this will be explored in more detail below.

to one possible hypothesis, which we will call the '*tilted-graded'* view, the propensity to agree with a proposed ascription of intentionality to increasingly harmful outcomes is more pronounced than the propensity to disagree with a proposed ascription of intentionality to increasingly helpful outcomes. In that sense, different from the '*graded-linear'* view, the gradient of the curve tracing the relation between intentionality ascriptions and outcomes is much steeper on the negative part of the spectrum than on the positive part. **Figure 3** illustrates this idea. On a different hypothesis, which we will call the '*semi-graded'* view, there would *only* be a positive correlation between the propensity to ascribe intentionality and increasingly harmful outcomes. The propensity to say that foreseen helpful outcomes are not intentional remains the same no matter the degree to which they are helpful. As illustrated in **Figure 4**, the slope of the curve tracing the relation between an action's outcome and intentionality ascriptions is steep on the negative side of the outcome-spectrum and flat on the positive side. Finally, on a '*sloppy-V'* view, the propensity to ascribe intentionality would be positively correlated with *both* increasingly harmful and helpful outcomes. However, the gradient is considerably more pronounced for the harmful part of the spectrum than on the helpful part. Importantly, however, intentionality ascriptions on the harmful part of the outcome spectrum remain above the mid-point of the scale, whereas intentionality ascriptions on the helpful part of the outcome spectrum remain below.[5] **Figure 5** illustrates this view.

Figure 3. The *tilted-graded* view — red dots denote the Knobe effect.

Figure 4. The *semi-graded* view — red dots denote the Knobe effect.

[5] In the *sloppy-V* view (and in all other models) intentionality ascriptions for foreseen helpful and neutral outcomes can be expected to remain below (or at least not significantly above) the mid-point of the scale, given the robust findings showing that helpful side effects are judged nonintentional (see Knobe, 2003a, 2003b, 2004a, 2006; McCann, 2005; Pettit & Knobe, 2009).

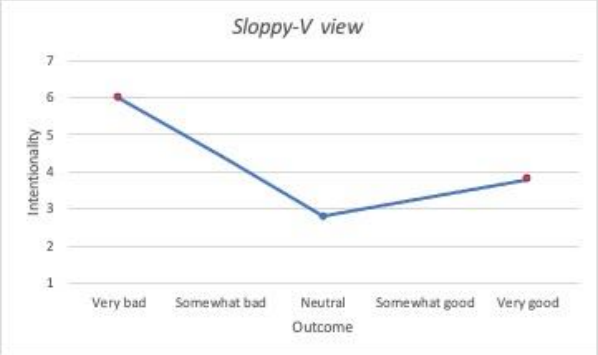

Figure 5. The *sloppy-V* view — red dots denote the Knobe effect.

Which of the former views best captures the relation between an action's outcome and intentionality ascriptions is precisely what our experiments aim to explore. Note that we are not suggesting that intentionality is a gradable concept (like e.g., "tall" or "rich"), according to which an action can be more or less intentional.[6] Rather, the intentionality scale much rather represents different levels of agreement or disagreement with a proposed ascription of intentionality, therefore measuring people's *propensity* to attribute intentionality.

3. Explaining the severity effect findings

There has been extensive debate about how to best interpret the Knobe effect. One of the discussions that dominates the literature has been whether the Knobe effect properly reflects people's competence with the concept of intentional action or whether it is the result of a bias. Another central discussion has revolved around the fact that the asymmetry also arises in the attribution of mental states other than intentionality —e.g., knowledge, desire, belief—, with some scholars arguing that an appropriate account of the Knobe effect should provide a single all-encompassing explanation (e.g., Alfano, Beebe & Robinson, 2012, who argue that all extant data on the Knobe effect can be explained by a 'single underlying asymmetry'). For the purpose of this paper, however, the debate about the plausibility or appropriateness of Knobe effect explanations is centered on a different issue: theoretically accommodating the severity effect findings by K&B.

Many scholars have conceived of the phenomenon underlying the Knobe effect in binary terms, that is as a yes/no, on/off, good/bad matter. This assumption is not surprising given that the standard research on the Knobe effect has focused on intentionality ascriptions for harmful outcomes v. helpful (or neutral) outcomes — that is, given that there are standardly but two data points. However, as hinted at by K&B, it might well be of import to explore a large spectrum of

[6] In clear contrast to this asssertion, some scholars argue that the adjective "intentional" and its components, e.g., desire and knowledge or belief, might be gradable. See, for instance, Egré, P. (2014) or (Egré, P. & Cova, F. 2015)

differently graded outcomes and people's corresponding attributions of intentionality. By limiting the *explanandum* to merely two data points (out of many), the research on the Knobe effect might have led some scholars to miss the correct explanation of the underlying phenomenon, including that of the two Knobe effect data points actually examined. It is one thing to argue why people ascribe intentionality to harmful outcomes but not to helpful ones. It is quite another, for instance, to explain why people's propensity to ascribe intentionality is commensurate with the outcome's degree of harm. In this section, we will briefly survey a few Knobe effect accounts and discuss their plausibility *vis-à-vis* the severity effect findings and, in general, gradeability. (For a more complete review see Feltz, 2007; Nichols & Ulatowski, 2007; and Cova, 2016).

3.1.Conceptual competence account

Joshua Knobe (2006, 2010; see also Pettit & Knobe, 2009) has argued that the asymmetrical ascriptions of intentionality in Knobe effect cases result from an asymmetry in the *moral valence* of the outcome — i.e., a morally bad outcome in the harm conditions v. a morally good outcome in the help conditions.[7] For Knobe, people's intentionality ascriptions are influenced by the moral valence of the outcome because their concept of intentionality is sensitive to moral factors. People evaluate the moral goodness or badness of the outcome and make judgments about the attitude that the agent should hold towards it. These moral evaluations, in turn, are taken into consideration when assessing whether the outcome was intentional or unintentional. When the outcome is morally bad and the agent is indifferent towards bringing it about, foresight suffices for intentionality to be ascribed; when the outcome is morally good, however, a desire or pro-attitude is a necessary condition of intentionality.[8]

Knobe's account, as is, cannot provide a plausible explanation for the severity effect findings by K&B. Whether an outcome is considered intentional or unintentional, on this account, depends (at least in part) on a binary feature: whether it is morally good or bad (but not how good or how bad). More specifically, a foreseen yet undesired outcome is considered intentional if and only if morally bad. In that sense, we should have expected participants of K&B's experiment to ascribe intentionality to the outcomes in both severity effect

[7] Note that it appears as if, for some time, Knobe was not committed to saying that the side-effect effect would only arise for cases differing in *moral* valence. Instead, he suggested that an asymmetry in intentionality ascriptions would result in cases differing in outcome valence *tout* court, i.e., outcomes good v. bad "in some respect" (see Knobe & Mendlow, 2004; see also Knobe, 2004b). In any case, Knobe's account has always relied on a binary *valence* structure.

[8] In his paper *Person as scientist, person as moralist* (2010), Knobe argues that we *expect* people to have a 'pro-attitude' towards morally good actions and a 'con-attitude' towards morally bad actions. Showing indifference to a morally bad action is sufficient for it to count as intentional, indifference towards a morally good action is far from being sufficient for it to count as intentional.

conditions —somewhat bad v. very bad— because they are both morally bad (i.e., a harm to the environment). Contrary to the former prediction, not only did K&B find that participants' ascriptions of intentionality to the outcomes in the severity effect conditions were significantly different, but also that participants presented with the very bad outcome condition were (on average) willing to ascribe intentionality while participants presented with the somewhat bad outcome condition were not.[9] It is implausible then to assert that a concept of intentionality that is sensitive to the moral valence of the outcome, as Knobe argues, can account for the severity effect findings. Rather, if one were to forward a conceptual account, one would have to argue that the concept of intentionality is determined (in part, again) not by the moral valence of the outcome but its *degree of moral badness*[10] — whether an outcome counts as intentional or unintentional does not depend on whether it is perceived as morally good or bad, but how bad it is. This, however, would be a substantial revision to Knobe's account.

3.2.Norm-based accounts

A group of scholars have argued that the Knobe effect does not stem from the moral valence of the outcome; rather, they argue, it stems from a more general binary feature: the *normative status* of an action (or outcome). The asymmetry in intentionality ascriptions, so regarded, results from the fact that the outcome in the harm condition is norm-violating (or results from a norm-violating action) while the outcome in the help condition is norm-conforming (or results from a norm-conforming action). For instance, Holton's (2010) account of the Knobe effect, which relies on said normative status, rests on the following two (related) claims: First, the scholar argues, to intentionally violate a norm one only needs to do it knowingly. To intentionally conform to a norm, however, one needs to be 'counterfactually guided by it'.[11] Second, in making attributions of intentional action (and of other action-relevant mental states), people consider (among other things) whether the agent intentionally violated or conformed to a norm when performing the action. As Holton puts it, "If in performing an action the agent intentionally violates a norm against an outcome, then that is a factor in the outcome being brought about intentionally." (2010, p. 4) Since in the harm condition of Knobe effect cases the agent intentionally violates a norm against an outcome (by doing so knowingly), people are willing to judge the outcome as

[9] In K&B's severity effect experiment, the average intentionality ascription for the very bad condition was significantly above the midpoint of the Likert scale. The average intentionality ascription for the somewhat bad condition, by contrast, was below the midpoint of the Likert scale (2017, p. 143).

[10] We would consider an even more plausible move to forward a more general badness account rather than a gradable moral badness account. As some scholars have put it (e.g. Machery, 2008; Alfano, Beebe & Jensen, 2012), the relation between intentionality ascriptions and outcomes can (and should) be explained without necessarily resorting to morality since the Knobe effect also arises in non-moral cases, e.g. the Sales Vignette by Knobe & Mendlow (2004).

[11] According to Holton, to be 'counterfactually guided by a norm' means to be willing to modify one's behavior in order to conform to such norm (2010).

intentionally brought about. In the help condition, by contrast, the agent is not taken to intentionally conform to a norm (since his actions are performed with indifference towards the norm), therefore people are not willing to judge the outcome as intentionally brought about.

Uttich & Lombrozo's (2010) account of the Knobe effect, although slightly different from Holton's, also relies on the normative status of actions. In particular, the scholars argue, a behavior that violates a norm (e.g., a moral norm, but not exclusively) implies that the agent had a strong reason for acting — strong enough not to conform to the norm. That the agent had a strong reason for acting against a norm, in turn, is informative of an intentional mental state. A norm-conforming behavior, by contrast, does not tells us anything about the agent's reasons for acting or his mental state. Since the agent's behavior in the harm condition of Knobe effect cases violate norms such as "one should protect the environment", people then take this norm-violating status as evidence on which to support an intentionality ascription. The norm-conforming behavior of the agent in the help conditions, however, does not provide any clues as to the agent's mental state. This asymmetry regarding the normative status of the actions in the harm and help conditions, according to Uttich & Lombrozo, thus explains the Knobe effect.

Neither Holton's nor Uttich & Lombrozo's accounts of the Knobe effect can provide a plausible explanation for the severity effect findings. The normative status, central to these accounts (and others), is a binary rather than a gradable feature. A difference in intentionality ascriptions should only arise in pairs of cases contrasting *norm-violating* v. *norm-conforming* actions or outcomes, regardless of the degree of an outcome's severity. There is, however, a potential fix for this type of account when it comes to explaining the severity effect findings: arguing for gradable features of norms or (more specifically) norm-violation, instead of a binary normative status. On the one hand, it seems possible to argue that a difference in the willingness to ascribe intentionality results from a gradual difference in features that (roughly) refer to the *type of norm* violated. For instance, the explicitness of the norm, the salience of the norm,[12] the interest protected by the norm (e.g., life, the environment, traffic normality, etc.), the strength with which the norm demands a certain conduct (i.e., how prohibited or prescribed it is),[13] whether deviance from the norm is sanctioned and how

[12] Robinson, Stey, and Alfano have argued, for instance, that "the [Knobe] effect is driven not by the violation of a norm as such, but by the violation of a salient norm (of any kind)" (2015, p. 180). The problem with salience, however, is that it cannot explain the severity effect findings, since the somewhat bad and very bad conditions do not vary on this respect. Rather, both severity effect conditions involve actions that violate *a same norm* (to protect the environment) with no other competing norms in place.

[13] Malle (forthcoming), for instance, finds that people reliably identify grades of norm strength - as put by the author "people consistently and consensually distinguish between deontic expressions that denote grades of

severely, the relevance of the norm for the society, and so on. Problematically, however, this is substantially different to the aforementioned norm-based accounts (and, in general, all other binary norm-based accounts). Moreover, it cannot explain the severity effect findings, since the somewhat bad and very bad outcome conditions of K&B's experiment do not vary on any of those respects. Rather, both conditions involve actions that violate *the same norm* (to protect the environment). On the other hand, one could also argue that a difference in the willingness to ascribe intentionality results from a gradual difference in features pertaining to *norm-violation* per se. For instance, to what extent does the action or the outcome deviates from what is expected (as prescribed or prohibited), or the extent to which the interest protected by the norm is affected. Nevertheless, even when this could potentially explain the severity effect findings by K&B, there is still a substantial difference between the aforementioned norm-based accounts (and, in general, all other binary norm-based accounts) and such gradable conceptualization of norm-violation. In any case, turning to the former gradable features of norm-violation or norms, although being a plausible move, would already imply a major theoretical revision to binary norm-based accounts.

3.3. Blame-driven bias accounts

For a group of scholars, the Knobe effect arises due to a systematic performance error. On a highly influential version of this type of accounts, in particular, the Knobe effect is the result of a *blame-driven bias* affecting people's intentionality judgments (cf. Alicke, 2000, 2008; Alicke & Rose, 2010; Nadelhoffer, 2004a, 2004b, 2005, 2006)[14]. People, the argument follows, ascribe intentionality to side effects in order to support *blame (or praise)* attributions in cases where the agent's actions arouse strong negative (or positive) reactions. Put differently, it is our blame attributions that shape our ascriptions of intentionality (and, potentially, other mental states). In that sense, participants presented with the harm condition of Knobe-effect cases are willing to say that the harmful side effects where intentionally brought about because they perceive the agent as deserving much blame. In the help condition, by contrast, the agent that brings about foreseen helpful side effects is not seen as deserving much blame (or praise even) and, consequently, intentionality is not ascribed.[15] If people perceived the agent in the help condition as deserving considerable blame or even praise

prohibition (e.g., frowned upon < unacceptable < forbidden) and grades of prescription (e.g., called for < expected < required)".

[14] For other accounts that regard the Knobe effect as a performance error see Adams & Steadman (2004a, 2004b) and Malle & Nelson (2003).

[15] On this account then, the harm and help conditions of Knobe effect cases are not analogous—as Alicke & Rose put it, when presented with Knobe's CHAIRMAN scenario "social perceivers view the environment-harming executive as a major jerk (i.e., one who arouses strong negative evaluations), but view the environment-helping executive as only a minor one." (2010, p. 330)

(although to a lesser extent),[16] participants would be more likely to ascribe intentionality to the side effects of his actions. The difference between the harm and help conditions with regards to blame attributions thus explains the Knobe effect.

Blame-driven bias accounts provide a plausible explanation to the severity effect findings. Blame (and praise) is a gradable moral judgment — one does not simply ascribe blame or not; one also assigns more or less blame (or praise) depending on different factors such as the agent's mental state when acting (e.g., intentionality, knowledge, recklessness) or the action's outcome (see Cushman, 2008; Malle, 2021; see also Kneer & Machery, 2019, for an empirical study on the effects of outcome on blame judgments). As Malle puts it, blame judgments are "graded assessments that take numerous pieces of information into account" (2021, p. 19). In support of this, it is worth mentioning that it is fitting to say that someone deserves more (or a higher degree of) blame than other. By contrast, saying that someone broke a norm more than other, for instance, does not seem fitting. In any case, it seems quite plausible to argue that, when presented with the very bad condition of the severity-effect experiment, people perceive the agent as more blameworthy (than that in the somewhat bad condition) because he is indifferent to bringing about not just any harmful outcome, but a very harmful outcome. This higher degree of perceived blameworthiness, in turn, might drive people's willingness to ascribe an intentionality ascription that supports such blame attribution. Simply put, people's propensity to say that an outcome was intentionally brought about might be commensurate with the degree to which an agent is perceived as blameworthy.[17]

3.4. Belief–attribution heuristic account

For Alfano, Beebe & Robinson (2012) the Knobe effect can be explained by a "*norm-violation/belief-attribution heuristic*" (or more simply a *belief-attribution heuristic*). According to this heuristic, people expect others to reflect more and form beliefs about actions that violate (salient) norms but not about actions that conform to norms, because norm-violating actions (different than norm-conforming ones) involve higher practical costs. Since intentionality entails belief

[16] Nadelhoffer (2004b) conducted an experimental study and assigned participants to a condition where a presumably praiseworthy agent brings about some side effects. He then found that 55% of the participants were willing to say that the side effects were intentionality brought about.

[17] Precisely, K&B find some empirical support for this conclusion: the average intentionality ascriptions for the *somewhat bad* and *very bad* conditions of their severity-effect experiment were almost identical to the average blame attributions to the agents in such conditions. As the scholars put it, it seems like "An increased attribution of intentionality (i.e., volitional behavior control) might be driven by an increased desire to blame the agent." (2017, p. 144)

(and so do other mental states sensitive to the Knobe-effect),[18] it is the belief-attribution heuristic that explains the Knobe effect: People, the scholars argue, are more inclined to ascribe intentionality in the harm condition but not in the help condition, because they attribute greater degrees of reflection and belief (entailed by intentionality) to the norm-violating agent in the harm condition than to the norm-conforming agent in the help condition.

The belief-attribution heuristic account loses plausibility when it comes to explaining the severity effect findings because of the central role it gives to the action's normative status. The latter, as stated above, is a binary feature —either norm-violating or norm-conforming— that, supposedly, leads to clear verdicts of intentional action. A norm-violating action or outcome would lead people to attribute intentionality, whereas a norm-conforming one would not. On such grounds, a belief-attribution heuristic could only account for a difference in ascriptions of intentionality (and other mental states) between pairs of cases contrasting norm-violating v. norm-conforming actions (or outcomes). Problematically, the severity effect conditions do not differ with respect to such normative status: the behavior of the agents in both conditions violate a same norm (which could be defined, roughly, as "one should protect the environment").

Nonetheless, the belief-attribution heuristic can still offer a plausible explanation for the severity effect (normative status aside). Actions that are potentially very harmful, one could argue, warrant greater levels of reflection and belief than less harmful actions — because, for instance, sanctions (moral and legal) may vary according to the degree of harm produced. This, in turn, might drive people's willingness to attribute a belief-entailing mental state (e.g., intentionality or knowledge) to the agents engaging in harmful actions. When presented with the very bad condition of K&B's experiment, where the agent foresees that his actions could bring about a severe harm, participants are then more willing to agree with a proposed ascription of intentionality. In the somewhat bad condition, however, participants are not so willing to agree with a proposed ascription of intentionality because the agent in such condition only foresees that his actions could bring about a minor harm.

[18] As Alfano et. al. put it, "Agent *a* intentionally makes it the case that *p* by φ-ing only if *a* believes that φ-ing would help to make it the case that *p*.". With regards to knowledge, also, "Agent *a* knows that φ-ing would make it the case that *p* only if *a* believes that φ-ing would help to make it the case that *p*." (2012, p. 266)

Account	References	Explanans	Summary	Binary or gradable account
Conceptual competence	Knobe (2006, 2010), Pettit & Knobe (2009)	A *morally bad* outcome.	Foresight is sufficient for an action to be intentional if and only if the outcome is morally bad.	Binary
Norm-asymmetry	Holton (2010)	*Intentional norm-violating* status of the outcome/action	To intentionally violate a norm one only needs to do so knowingly. To intentionally conform to a norm, however, one needs to be "counterfactually guided by it". Judgments of intentional action are then susceptible to whether the agent intentionally violated or conformed to a norm when acting.	Binary
Trade-off hypothesis	Machery (2008)	Conceptualization of harmful outcomes as *costs* incurred in order to obtain a benefit.	People conceptualize the harmful side effects as costs incurred in order to gain a benefit or simply, a trade-off. Since we think of the costs for gaining a benefit as intentionally incurred, harmful side effects are judged intentional. By contrast, since helpful side-effects are not conceptualized as costs but rather as additional benefits, they are not deemed intentional.	Binary
Rational scientist view	Uttich & Lombrozo (2010)	*Norm-violating* status of the action.	The norm-violating status of an action suggests that the agent had a strong reason not to conform to a norm, which in turn is evidence of intentionality.	Binary
Pragmatic implicatures	Adams & Steadman (2004a, 2004b)	Conversationally implying that the agent is *blameworthy.*	People judge the harmful side effects as intentional because they want to (conversationally) imply that the agent is blameworthy for his negative actions. Not ascribing intentionality could be interpreted as if the agent is not blameworthy. Conversely, people do not ascribe intentionality to the helpful side effects since they do not perceive the agent as blameworthy or praiseworthy, and do not want to imply so.	Binary
Multiple concepts of intentionality	Cushman & Mele (2008); see also Nichols & Ulatowski (2007), Lanteri (2012).	Majority-held 'half' concept of intentionality, where belief is a sufficient condition of intentionality if and only if the action is *morally bad*	There are 'two and a half' concepts of intentionality. First, there is a group of people that hold a concept where having a justified belief about the action is sufficient for intentionality to ascribed (a 'belief-based' concept of intentionality). Second, there is another group of people that treat desire (to perform the action) as a necessary condition of intentionality (a 'desire-based' concept of intentionality). Finally, there is a 'half' concept of intentionality, where belief is a sufficient condition of intentionality if and only if the action is morally bad. The Knobe effect is then explained by the fact that a large number of participants employ the 'half' concept of intentionality	Binary

			(although it is not explicitly endorsed by them), and thus ascribe intentionality asymmetrically to foreseen harmful and helpful side effects.	
Blame-driven bias	Nadelhoffer (2006, 2008), Alicke, (2000, 2008), Alicke & Rose (2010)	The *blame* (or praise) attributed to the agent for arousing negative reactions	People ascribe intentionality to harmful side effects because they want to support their desire to blame the agent. By contrast, they do not ascribe intentionality to the agent that brings about helpful side effects because he is not perceived as deserving much blame.	Gradable
Belief-attribution heuristics	Alfano, et. al (2012); see also Robinson, Stey, & Alfano (2015)	The reflection and *belief* state attributed to the agent	People expect others to reflect more and form beliefs about norm-violating actions but not about norm-conforming ones. Since intentionality is a belief-entailing mental state, people are willing to ascribe it to the norm-violating agent (in the harm condition) but not to the norm-conforming agent (in the help condition).	Gradable

Table 1. Brief summary of (some) Knobe effect accounts.

3.5. Binary v. gradable accounts

In the preceding sections, a distinction between two types of explanations of the Knobe effect emerged: Those which resort to a binary feature of actions/outcomes and those which invoke (at least in principle) gradable features. On *binary accounts*, as we will call them, the asymmetry in intentionality ascriptions is the result of a morally good v. bad outcome, a norm-conforming v. norm-violating action/outcome, a cost v. benefit conceptualization, and so on (see **Table 1**). On *gradable accounts*, by contrast, the Knobe effect asymmetry is driven by an underlying difference in the degrees of attribution of a single feature, e.g., blame or reflection and belief. The latter features, we said, are better suited than the former to explain K&B's findings; a propensity to ascribe intentionality that is commensurate with the *degree* of harm of an action's outcome cannot be explained by *binary* factors applicable to both harmful outcomes. However, some of the accounts that turn to binary variables can theoretically accommodate the severity effect findings if revised. One could potentially argue that the propensity to ascribe intentionality is driven by, e.g., the badness of the outcome or the force of the norm (however the latter is exactly conceptualized), rather than a rough norm-violating status or negative valence. That being said, in the next section we will report two experiments with the purpose of investigating the shape of the curve tracing the relation between intentionality ascriptions and graded outcomes.

4. Experiment 1

Experiment 1 explores attributions of intentionality (and knowledge) across a range of different outcomes: *very bad, somewhat bad, neutral, somewhat good and very good*. To get a better understanding of the latter, we also measured *perceived* goodness and badness of the outcome. Methodologically, this is by and large unchartered territory. In contrast to classic side-effect effect research, we are less interested in the difference between intentionality ascriptions for good and bad outcomes, but in the correlation between perceived goodness and badness of outcome on intentionality. While there is some data on the negative part of the outcome spectrum (cf. K&B, 2017; Prochownik et. al, 2020; Tobia, in preparation), the relation between differently graded positive outcomes and intentionality attributions is completely unexplored. Moreover, since the *graded-linear* view (Fig. 2) is not the only theoretically likely possibility and the *semi-graded* (Fig. 3), *tilted-graded* (Fig. 4) or the *sloppy-V* (Fig. 5) views might be more plausible than the *graded-linear* view, we decided to explore the positive and the negative part of the outcome-spectrum separately (and preregistered this decision).[19]

A further complication regards neutral outcomes, on which data is sparse and theoretical reflection limited (for notable exceptions see Kneer & Machery, 2019, and the papers cited therein). What to do with "outcome midpoint" data? In this study we proceeded in the following, preregistered fashion: The positive part of the outcome-spectrum was defined as the datasets of participants that were assigned to the very good, somewhat good, and neutral conditions, including only those participants who actually judged the outcome to be neutral or good (on a 7-point Likert scale, scores > 3 for the outcome badness/goodness variable). We thus excluded participants who judged the outcome as bad (on a 7-point Likert scale, scores < 4 for the outcome badness/goodness variable), given that for the positive part of the outcome-spectrum we were only interested in testing the relation between *perceived* outcome goodness and the mental state ascriptions. The negative part of the outcome-spectrum, on the other hand, was defined as the datasets of participants who were assigned to the very bad, somewhat bad, and neutral conditions, including only those participants who judged the outcome to be bad or neutral (on a 7-point Likert scale, scores < 5 for the outcome badness/goodness variable). We thus excluded participants who judged the outcome as good (on a 7-point Likert scale, scores > 4 for the outcome badness/goodness variable) since, for this part, we were interested in testing the relation between *perceived* outcome badness and the mental state ascriptions.

[19]Pre-registration link: https://aspredicted.org/blind.php?x=9sq6m4

Since we contend that perceived blameworthiness (or praiseworthiness), rather than outcomes per se, might drive intentionality (and potentially knowledge) attributions, we also asked participants to ascribe blame and praise to the agent. As with neutral outcomes, an issue arises with agents perceived as neither blame- nor praiseworthy. To resolve this matter, we opted for the following procedure: The praise part of the spectrum was defined as the datasets of participants assigned to the very good, somewhat good, and neutral outcome conditions, including only those participants who assigned praise or neither blame nor praise (on a 7-point Likert scale, scores > 3 for the blame/praise variable). We thus excluded participants who assigned blame (on a 7-point Likert scale, scores < 4 for the blame/praise variable). The blame part of the spectrum, on the other hand, was defined as the datasets of participants that were assigned to the *very bad*, *somewhat bad*, and *neutral* outcome conditions, including only those participants who assigned blame or neither blame nor praise (on a 7-point Likert scale, scores < 5 for the blame/praise variable). We thus excluded those participants who assigned praise (on a 7-point Likert scale, scores > 4 for the blame/praise variable).

4.1. Participants

372 participants were recruited on Amazon Mechanical Turk to complete a Qualtrics survey. We restricted the participants' IP address to the US. As preregistered,[20] non-native English speakers, participants who failed an attention check or a comprehension check, and those who responded under 10 seconds were excluded. The final sample of participants consisted of 289 participants (age M=41years, SD=13 years, range 20–78), of whom 147 were female.

4.2. Materials and procedure

In a between-subjects design, participants were randomly assigned to one of five conditions of the WASTE scenario, where the agent's actions produce either *very bad*, *somewhat bad*, *neutral*, *somewhat good*, or *very good* side effects. The scenario read (**bold** labels and brackets omitted):

> **Negative conditions (i.e., very bad and somewhat bad):**
>
> The president of a riverside company is approached by one of his advisors, who says: "We could change the waste management system for a more efficient one. This will surely save the company large amounts of money. There would also be [severe / minor] adverse effects on the environment.

[20] Pre-registration link: https://aspredicted.org/blind.php?x=9sq6m4

The river will be [severely polluted / slightly soiled] and the animals living in and near the river will [die / be a bit disturbed]".
The president responds: "I do not care at all about the environment. All I care about is saving the company large amounts of money. Let's change the waste management system."
They implement the new waste management system. The river is [severely polluted/ slightly soiled] and the animals living in and near the river [die/are a bit disturbed].

Neutral condition:

The president of a riverside company is approached by one of his advisors, who says: "We could change the waste management system for a more efficient one. This will surely save the company large amounts of money. There would also be no effects on the environment. There will be no effects on the river or the animals living in and near the river".
The president responds: "I do not care at all about the environment. All I care about is saving the company large amounts of money. Let's change the waste management system."
They implement the new waste management system. There are no effects on the river or the animals living in and near the river.

Positive conditions (i.e., very good and somewhat good):

The president of a riverside company is approached by one of his advisors, who says: "We could change the waste management system for a more efficient one. This will surely save the company large amounts of money. There would also be [very / minor] positive effects on the environment. The river will be [much / a bit] cleaner and the animals living in and near the river will benefit [strongly from this/ from this somewhat]."
The president responds: "I do not care at all about the environment. All I care about is saving the company large amounts of money. Let's change the waste management system."
They implement the new waste management system. The river is [much/a bit] cleaner and the animals living in and near the river benefit [strongly from this / from this somewhat].

Participants were then asked to rate their agreement or disagreement on a 7-point Likert scale (1= 'completely disagree' to 7= 'completely agree') regarding the following statements:

"The president of the company intentionally [harmed the environment/kept the effects on the environment neutral/helped the environment]."

> "The president of the company knew that he would [harm the environment/keep the effects on the environment neutral/help the environment]."

Participants were also asked how much blame and praise the president deserved for his action (1 = 'a lot of blame', 7 = 'a lot of praise'). Finally, they had to rate the goodness or badness of the outcome they were presented with (1 = 'very bad', 7 = 'very good').

4.3. Results

Consistent with the severity-effect findings by K&B (2017), we found a positive correlation between intentionality ascriptions and perceived outcome badness (r =.416, p < .001): the worse participants perceived the outcome to be, the more likely they were to agree with the claim that it about intentionally. On the positive side of the outcome-spectrum, there was a weak negative correlation between intentionality ascriptions and perceived outcome goodness (r= -.192, p = .013): the more desirable participants perceived the outcome to be, the less likely they were to agree with the claim that the president of the company brought it about intentionally. As regards knowledge ascriptions, we also found a positive correlation with perceived outcome badness (r = .478, p < .001), but no correlation with perceived outcome goodness (r = .130, p = .096). The results are shown in Figure 5 and Figure 6.

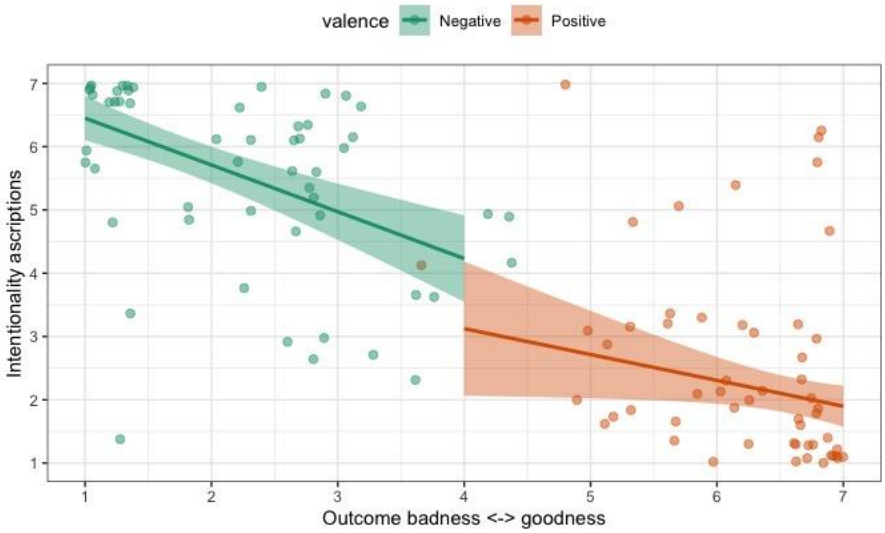

Figure 5. Relation between intentionality ascriptions and outcome badness/goodness

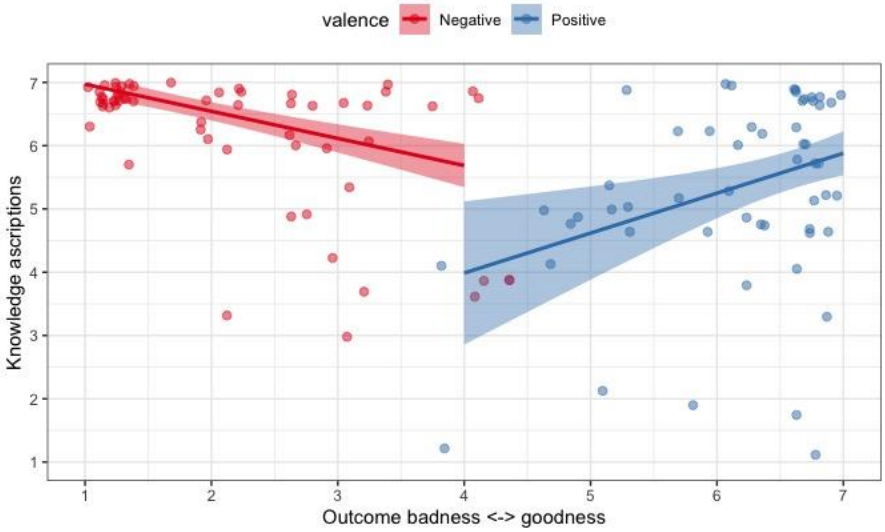

Figure 6. Relation between knowledge ascriptions and outcome badness/goodness

With regards to blame and praise, Experiment 1 revealed a strong positive correlation of perceived blameworthiness on intentionality (r = .741, p < .001) and knowledge ascriptions (r = .509, p < .001), and a (much weaker) positive correlation of perceived praiseworthiness on intentionality (r = .385, p <.001) and knowledge ascriptions (r = .205, p = .011). The more blameworthy or praiseworthy participants perceived the agent to be, the more likely they were to agree with a proposed ascription of an action relevant mental state (intentionality or knowledge). The former results are shown in Figure 7 and Figure 8.

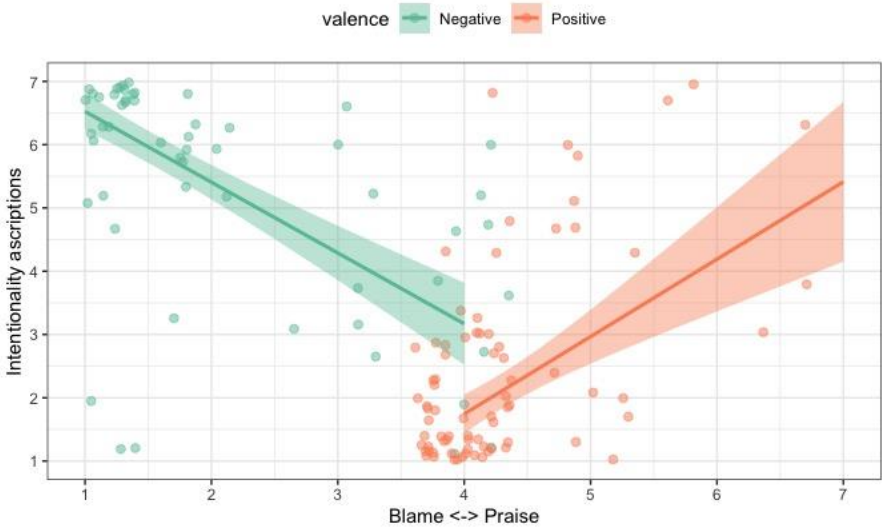

Figure 7. Relation between intentionality ascriptions and blame/praise

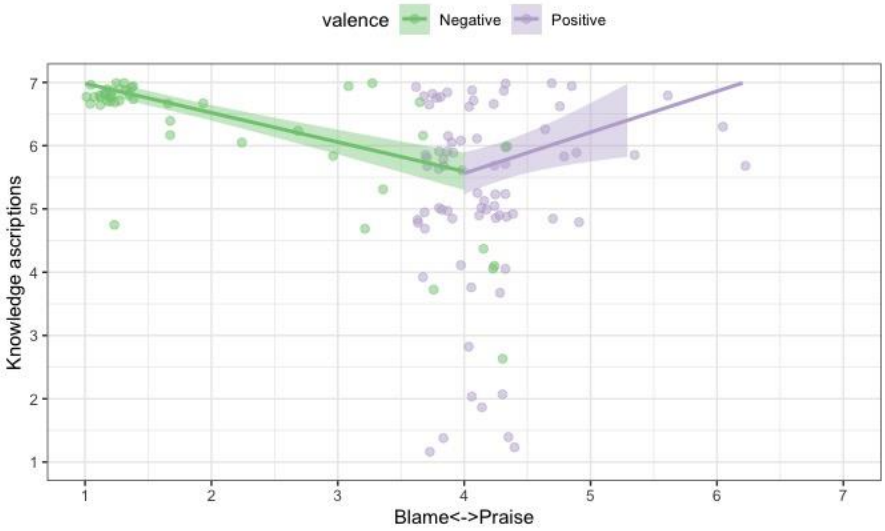

Figure 8. Relation between knowledge ascriptions and blame/praise

4.4. Discussion

The data suggests a number of implications. *First*, a strictly bivalent view, according to which intentionality (and knowledge) is ascribed in virtue of negative valence is at best incomplete, if not false: A somewhat crude conceptual tool, it cannot explain the difference across similarly valenced observations, and it risks mischaracterizing observations that are not at the far ends of the spectrum.[21] *Second*, the negative part of the spectrum for all four pairs of variables (i.e., intentionality and knowledge ascriptions paired with perceived outcome badness or blame) is consistent with the graded models proposed above. This is unsurprising, because the latter, based on previous findings by Kneer & Bourgeois-Gironde, are identical in this regard: An increase in severity of outcome (or blame) is expected to correlate positively with the attribution of intentionality (or knowledge). What stands out, however, is the fact that the correlation between intentionality and blame (r = .741, p = < .001) is nearly *twice* as pronounced as the one between intentionality and severity of negative outcome (r = .416, p = < .001).[22]

Let's turn, *third*, to the positive part of the spectrum – that is, a part of the conceptual space that has been near-universally neglected so far. The *Sloppy-V* view (**Fig. 4**) seems out of the question for two pairs of variables: Intentionality and knowledge ascriptions paired with perceived outcome goodness. This is to

[21] Though things work well enough for the particular scenario at hand, the data suggests that some outcomes which are deemed just a little good or a little bad might not give rise to clear verdicts (about intentionality or knowledge) as the bivalent view would imply.

[22] The correlation between knowledge and blame (r = .509, p = < .001) is also more pronounced than that between knowledge and severity of outcome (r = .478, p = < .001). However, the difference in correlation coefficients is much smaller than the difference between the pairs (i) intentionality and blame and (ii) intentionality and badness.

say that the propensity to ascribe intentionality or knowledge does not simply depend on the *degree* to which an outcome is viewed in a positive *or* negative light. The other two models, i.e., the *tilted-graded* (**Fig. 3**) and the *semi-graded* (**Fig. 4**), hold more promise: Perceived outcome goodness is negatively correlated, weakly, with intentionality (r = -.192, p = .013) and not correlated at all with knowledge (r = .130, p = .096). For the remaining two pairs of variables in the positive part of the spectrum —i.e., intentionality and knowledge ascriptions paired with praise—, the *Sloppy-V* (**Fig. 4**) view comes back into the picture: Praise is positively correlated with intentionality (r = .385) and knowledge (r = .205). Put differently, the propensity to ascribe intentionality or knowledge does seems to depend on the *degree* to which an agent is viewed as blameworthy or praiseworthy.

So far, the implications of Experiment with regards to the negative part of the spectrum seem clear: Consistent with the graded models (**Figures 1 to 4**), the propensity to ascribe an action-relevant mental state (intentionality or knowledge) depends on the *degree* to which an outcome is viewed as bad or an agent is viewed as blameworthy. However, the implications for the positive part of the spectrum are not so clear: While the propensity to ascribe action-relevant mental states (again, intentionality or knowledge) does not depend on the *degree* to which an outcome is viewed as good, it does seem to depend (to a minor extent) on the *degree* to which an agent is viewed as praiseworthy. The bivalent views out of the picture, we are still left with the question of which of the graded models presented in Section 2 offers the best fit for the relation between outcomes (and blame or praise) and intentionality (and knowledge) ascriptions. To provide a definitive answer to the former question and to establish the external validity of the results of Experiment 1, we conducted an Experiment 2. In the latter, instead of the WASTE scenario, participants were presented with one of three different scenarios, where an agent's actions produce either *very bad*, *somewhat bad*, *neutral*, *somewhat good*, or *very good* side effects. Importantly, the datasets of all participants of Experiment 2 are analyzed together.

5. Experiment 2

5.1. Participants

497 participants were recruited on Amazon Mechanical Turk to complete a Qualtrics survey. We restricted the participants' IP address to the US. As preregistered,[23] non-native English speakers, participants who failed an attention check or a comprehension check, and those who responded under 10 seconds

[23] Pre-registration link: https://aspredicted.org/blind.php?x=7fb9pq

were excluded. The final sample of participants consisted of N = 400 (age M=41years, SD=12 years, range 21–79), of whom 228 were female.

5.2. Materials and procedure

In a between-subjects design, participants were randomly assigned to one out of three scenarios —DAM, MALL, AND PUBLIC IMAGE— (see **Appendix** for the full scenarios) either in the *very bad*, *somewhat bad*, *neutral*, *somewhat good*, or *very good* condition. That is, participants were randomly assigned to one out of fifteen conditions (3 types of scenarios x 5 outcome conditions). After reading the vignette, they were asked to rate their agreement or disagreement with a proposed ascription of intentionality and knowledge, on a 7–point Likert scale ranging from (1) 'completely disagree' to (7) 'completely agree'. As in Experiment 1, participants were asked to rate (also on a 7–point Likert scale) the extent to which they deemed the agent to deserve blame or praise for their actions, and the goodness or badness of the outcome they were presented with.

5.3. Results

For this second experiment we combined the datasets of all participants regardless of the scenario they were presented to. The results were similar to those of Experiment 1. There was a positive correlation between intentionality ascriptions and perceived outcome badness (r = .496, p <.001), and a weak negative correlation between intentionality ascriptions and perceived outcome goodness (r = -.158, p =.019). The worse participants perceived the outcome to be, the more likely they were to agree with the claim that the agent brought it about intentionally. Conversely, the more desirable participants perceived the outcome to be, the less likely they were to agree with the claim that the agent brought it about intentionally. As regards knowledge ascriptions, we found that they were positively correlated with perceived outcome badness (r = .363, p <.001), but not correlated at all with perceived outcome goodness (r = .085, p <.210). That is, the worse participants perceived the outcome to be, the more likely they were to agree with the claim that the agent brought it about knowingly. However, no matter how desirable they perceived the outcome to be, they were not more or less likely to agree that the agent brought it about knowingly. The results are shown in Figure 9 and Figure 10.

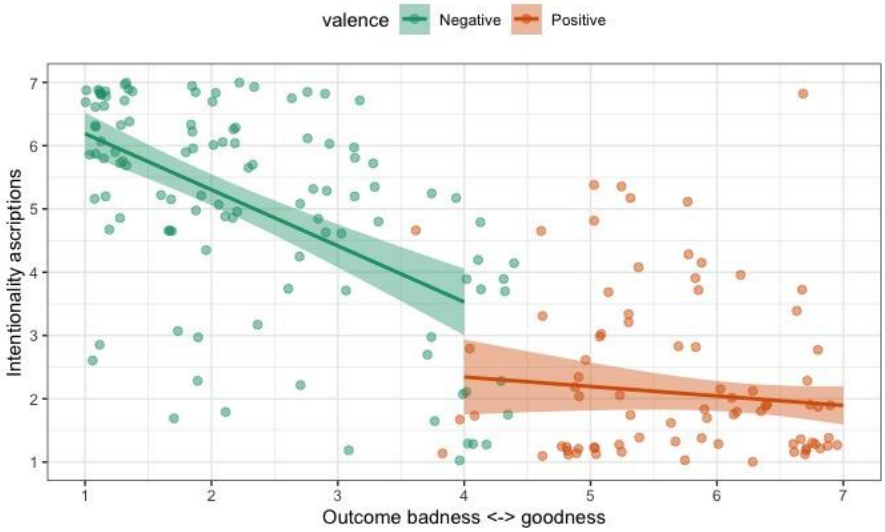

Figure 9. Relation between intentionality ascriptions and outcome badness/goodness

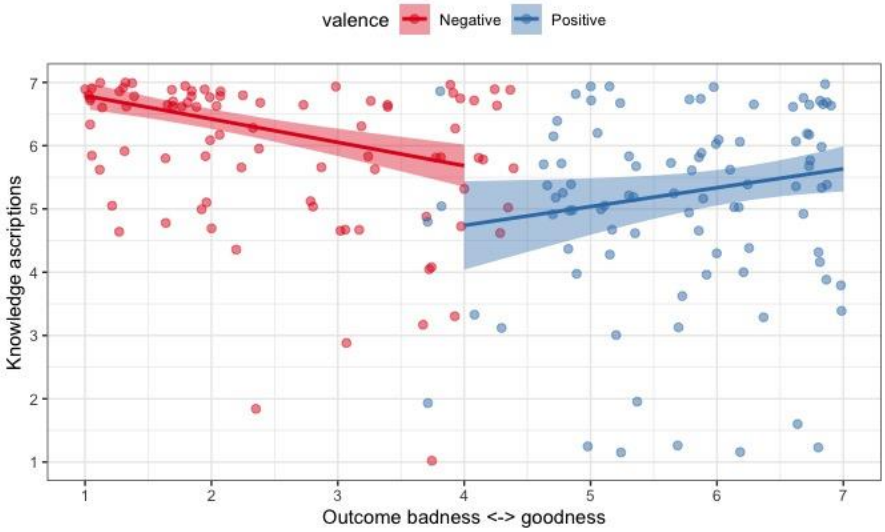

Figure 10. Relation between knowledge ascriptions and outcome badness/goodness

Experiment 2 also revealed, as did Experiment 1, that blame ascriptions were positively correlated with intentionality (r = .654, p <.001) and knowledge (r = .403, p < .001). The more blameworthy participants perceived the agent to be, the more likely they were to agree with a proposed ascription of intentionality or knowledge. Praise ascriptions were also positively correlated with intentionality ascriptions (r = .250, p <.001) as well as knowledge, although not significantly (r = .111, p = .116). That is, the more praiseworthy participants perceived the agent to be, the more likely they were to agree with a proposed ascription of intentionality but not knowledge. The former results are shown in Figure 11 and Figure 12.

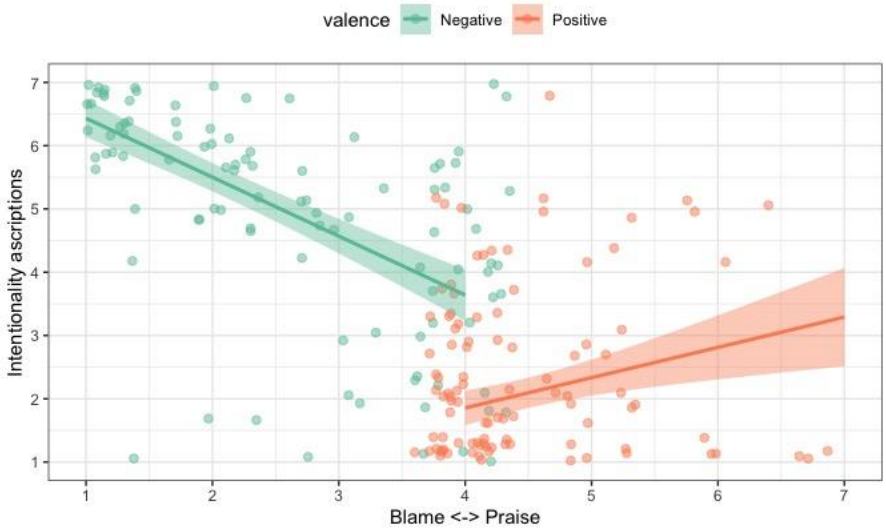

Figure 11. Relation between intentionality ascriptions and blame/praise

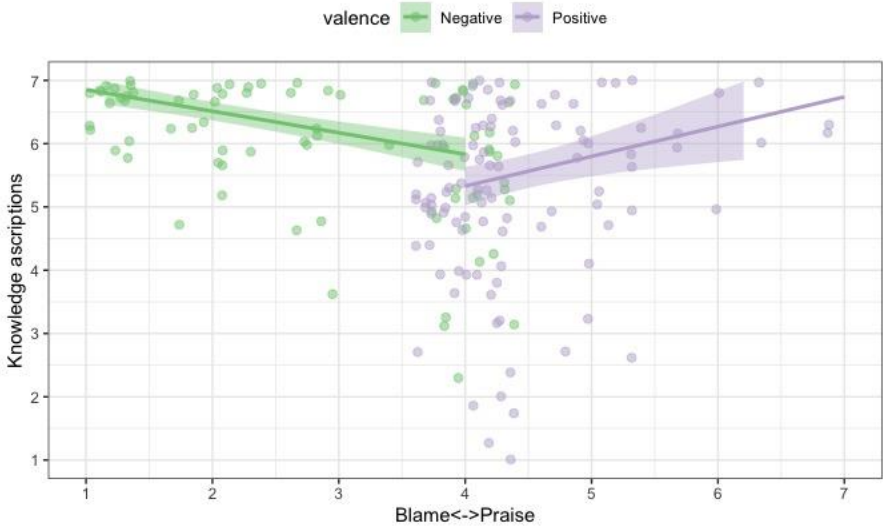

Figure 12. Relation between knowledge ascriptions and blame/praise

5.4. Discussion

Consistent with Experiment 1, the findings of Experiment 2 challenge strictly bivalent views and provide support for graded views. The negative part of the spectrum for all four pairs of variables (i.e., intentionality and knowledge ascriptions paired with perceived outcome badness or blame) is consistent with the graded models proposed in Section 2: As expected, an increase in severity of outcome (or blame) correlates positively with the attribution of intentionality or knowledge. Interestingly, and similar to what found in Experiment 1, the

correlation between intentionality and blame (r = .654) is much more pronounced than that between intentionality and severity of outcome (r = .496).[24]

On the positive part of the spectrum, replicating Experiment 1, the data suggests that the *tilted-graded* (Fig. 3) and the *sloppy-V* (Fig. 5) models best capture the relation of intentionality ascriptions on perceived outcome goodness and praise ascriptions, respectively. Participants' propensity to ascribe intentionality does not depend on the degree to which an outcome is perceived to be bad *or* good. However, it does seem to depend on the degree to which the agent is seen as blameworthy *or* praiseworthy (although to a minor extent). The more praise (or blame) participants attribute to the agent, the more likely they are to say that the agent brought about an outcome intentionally. When it comes to the remaining two pairs of variables —i.e., knowledge ascriptions paired with perceived outcome goodness and praise— the *semi-graded* model (Fig. 2) seems the best fit. Participants' propensity to ascribe knowledge depends on the degree to which an outcome or an agent is viewed in a negative light —as does the propensity to ascribe intentionality—, but not on the degree to which an outcome or an agent is viewed in a positive light. This is where the only difference between Experiment 1 and 2 arises: The positive correlation between knowledge ascriptions and perceived praiseworthiness, significant in Experiment 1, was non-significant in Experiment 2. For the rest, Experiment 2 replicates the findings of Experiment 1.

To sum up: The negative side of the spectrum is consistent with all of the graded models presented in Section 2 and with K&B's findings. The propensity to ascribe intentionality or knowledge depends on the *degree* to which an outcome or an agent is viewed in a negative light. Furthermore, the degree to which an agent is seen as blameworthy warrants a greater the increase in people's willingness to attribute intentionality than the degree to which an outcome is perceived to be bad. On the positive part of the spectrum, things are a bit more complicated: The relation of intentionality ascriptions on perceived outcome goodness is consistent with the *tilted-graded* model (Fig. 3), whereas the relation of intentionality ascriptions on praise is consistent with the *sloppy-V* model (Fig.5). As if that weren't enough, the relation of knowledge (on the one hand) on perceived outcome goodness and praiseworthiness (on the other hand), once we account for many scenarios, is not consistent with any of the former two models but with the *semi-graded* model (Fig. 2).

[24] The correlation between knowledge and blame (r = .403, p = < .001) is also more pronounced than that between knowledge and severity of outcome (r = .363, p = < .001). However, the difference in correlation coefficients is much smaller than the difference between the pairs (i) intentionality and blame and (ii) intentionality and badness.

6. General Discussion

Experiments 1 and 2 revealed that the relation between intentionality ascriptions and outcomes is of graded nature and that outcome valence plays a central role in defining the strength of such relation. Whereas the degree to which an outcome is perceived as harmful warrants a *considerable increase* in the propensity to ascribe intentionality —hence a steep curve gradient—, the degree to which an outcome is perceived as helpful only warrants a *minor decrease* in people's propensity to judge it unintentional —hence a less steep or flatter curve gradient—. The relation between intentionality ascriptions and outcomes, our experiments thus reveal, is best represented by the *tilted-graded* view (Figure 3).

Binary accounts of the Knobe effect face pressure from our findings. Knobe's own conceptual account (as well as the conceptual diversity accounts summarized in **Table 1**), which rely on moral valence (bad or good), cannot explain a difference in the propensity to ascribe intentionality between outcomes that are similarly valenced. Neither can norm-based accounts explain a difference in the propensity to ascribe intentionality between outcomes that have the same normative status (norm-violating or norm-conforming). Knobe's conceptual account and norm-based accounts can only explain an asymmetry in intentionality ascriptions between cases contrasting morally good v. bad or norm-violating v. norm-conforming outcomes (or actions). The former accounts (or any other account that conceives of the Knobe effect in binary terms)[25], as they stand, are thus inappropriate to explain our empirical findings. This, however, does not mean that they cannot be amended. One could say, forwarding a conceptual account, that the concept of intentionality and its application is determined (partially) by the outcome's *degree* of moral badness (or goodness), and not simply its moral valence. Or, if one were to forward a normative account, that the propensity to ascribe intentionality is driven by a gradual difference in features pertaining to norms or norm-violation. It could be that, for instance, people are more willing to say that an outcome is intentionally brought about if it violates a norm to a greater extent (e.g., because it is more harmful to the interest protected by the norm). Providing such amendments, however, is not the purpose of this paper.

By contrast, the findings of our experiments provide partial support to the *belief-attribution heuristic* account.[26] On the negative part of the spectrum, we found

[25] The trade-off hypothesis (Machery, 2008), for instance, is also inappropriate to explain the findings of our experiments. A trade-off/no trade-off dichotomy is still too crude of a conceptual tool that cannot explain a difference in people's propensity to ascribe intentionality between cases that can be conceptualized similarly — i.e., as a trade-off or not.

[26] A main challenge to this account (noted in Section 3.4) regards the central role given by the belief-attribution heuristic account to the action's normative status. This, however, we leave aside with the purpose of further discussing the plausibility of a modified version of the aforementioned account.

that people's propensity to ascribe both intentionality and knowledge was positively correlated (to a similar extent) with outcome badness. This could be interpreted as follows: the more severe the outcome of an action the more likely people interpret the agent as having reflected (and formed beliefs) about the outcome, which in turn translate into a greater propensity to attribute belief-entailing mental states (intentionality and knowledge). As Alfano, Beebe & Robinson put it:

> …a greater degree of reflection is in general rationally required of agents whose actions will bring about a harm or violate a salient norm, and that this greater degree of reflection leads attributors to ascribe higher degrees of belief to these agents." (p. 275, 2012)

As regards the positive part of the spectrum, proponents of this account could argue that the absence of a relation between perceived outcome goodness and knowledge (but not intentionality) ascriptions is due to the fact that people do not engage in greater degrees of reflection and belief about potentially good outcomes, no matter how desirable they can be, because they do not involve any practical costs.[27] Problematically, however, whereas increasingly helpful outcomes do not significantly increase or decrease people's propensity to ascribe knowledge, they do significantly (though weakly) decrease the propensity to ascribe intentionality. This difference in the shape of the relation between outcomes (on the one hand) and intentionality and knowledge ascriptions (on the other) is problematic for the belief-attribution heuristic account, since its proponents argue for a unified account of all the existing data concerning the Knobe effect. We then wonder: What feature of the belief-attribution heuristic explains the difference in the relation of outcomes on (i) intentionality and (ii) knowledge ascriptions?

Blame-driven bias accounts, we think, are better suited to explain the findings of our study. On the negative part of the spectrum, we found that the relation of intentionality (and knowledge) ascriptions on (i) badness and (ii) blame was similar: both outcome badness and blame were positively correlated with intentionality ascriptions. More importantly, however, the degree of blame attributed to an agent warrants a greater increase in intentionality ascriptions (all $rs > 0.654$) than the degree to which an outcome is considered bad (all $rs < 0.496$)—it is not simply outcome severity; it is much rather the desire to blame the agent that drives intentionality ascriptions. This provides further support to Kneer & Bourgeois-Gironde's (2017) study, where the average intentionality ascriptions were almost identical to the blame attributions across the severity-

[27] In the words of Alfano, Beebe & Robinson, one does not say to oneself "Wait! I need to stop and think carefully about whether helping (…) is something that I should be doing." (2012, p. 269).

effect conditions. It might also explain why Prochownik, et. al (2020) only find a severity effect when presenting participants with Kneer & Bourgeois-Gironde's BEACHTOWN scenario: Testing scenarios where the conditions vary in degrees of outcome badness may not give rise to a difference in intentionality ascriptions if the agents in such conditions are perceived as equally blameworthy.

The findings on the positive part of the spectrum are also consistent with *blame-driven bias* accounts.[28] *First*, as stated above, proponents of this type of accounts have also argued that if an agent were perceived as deserving considerable praise, people would also be inclined to ascribe intentionality to the foreseen (yet undesired) outcomes of his actions (cf. Alicke, 2008; Alicke & Rose, 2010; Nadelhoffer, 2004b, 2006). The desire to praise an agent, as to blame her, can shape people's attributions of intentionality. Not surprisingly then, we found that the propensity to ascribe intentionality was positively correlated with the perceived praiseworthiness of the agent.[29] Also consistent with Nadelhoffer's (2004b) previous findings, the relation of praise on intentionality ascriptions is almost *half* as pronounced than that of blame. *Second*, the fact that the propensity to ascribe intentionality is negatively correlated with graded helpful outcomes but positively correlated with perceived praiseworthiness might seem counterintuitive, but it is only in appearance. As put by Knobe & Mendlow, there is a "distinction between praiseworthiness and blameworthiness (on the one hand) and goodness and badness (on the other)." (2004, p. 253), such that there are cases in which actions are good without being praiseworthy and vice-versa. Proponents of the *blame-driven bias* accounts could then argue that, because of the difference in nature between judgments of goodness (or badness) and praiseworthiness (or blameworthiness), they do not necessarily (although could) elicit the same influence in judgments of intentionality. If the agents that bring about increasingly helpful outcomes were also perceived as increasingly praiseworthy (or blameworthy), the argument would follow, the relation of intentionality ascriptions on perceived outcome goodness would be similar to that on praise.[30]

[28] The fact that praise is not significantly correlated with knowledge (in Experiment 2) but with intentionality, could be perceived as a challenge to blame-driven bias accounts. However, proponents of this account do not seem to strive for unifying all the data on the Knobe-effect (or the severity effect for that matter). There is no attempt of explaining the asymmetrical attributions of other mental states (e.g., knowledge or foresight) to agents bringing about helpful and harmful side effects (Alicke & Rose, 2010, e.g., say that their blame-driven account explains the Knobe effect on intentionality ascriptions but only *presumably* on causation and foresight attributions).

[29] It is worth noting that the ascriptions of intentionality of participants presented with the positive (and neutral) conditions are located mostly below the mid-point of the intentionality scale and around the mid-point of the blame/praise scale (see Figures 7 and 11). This could indicate that participants in those conditions were, on average, not propense to attribute intentionality.

[30] A different way to frame this is to simply argue that, on the praise part of the spectrum, the data points were those of participants who perceived the agent as praiseworthy but not necessarily all the participants who thought that outcome was good.

7. Conclusion

In our study, we found that there is (i) a positive correlation between the propensity to ascribe intentionality (or knowledge) to an outcome and the degree to which an outcome is viewed as bad, and (ii) a more pronounced positive correlation between the propensity to ascribe intentionality (or knowledge) and the degree to which an agent is perceived as blameworthy. We also found that (iii) the degree to which an outcome is viewed as good is negatively correlated (although weakly) with the propensity to ascribe intentionality, but that (iii) the degree to which an agent is viewed as praiseworthy is positively correlated with the propensity to ascribe intentionality. These findings, we argue, support the hypothesis that the relation between intentionality ascriptions and outcomes is of graded nature. This, we further argue, challenges binary accounts of the Knobe effect while, at the same time, provides support to gradable accounts such as the *belief-attribution heuristic* account and the blame-driven bias accounts. Blame-driven bias accounts, however, seem better suited to explain our findings than the belief-attribution heuristic account.

In any case, further inquiry into the relation between outcomes and action-relevant mental states is required. First, research on the relation between belief attributions and outcome goodness or badness seems necessary. This, rather than relying simply on the relation between intentionality or knowledge ascriptions (on the one hand) and outcome goodness or badness (on the other), is a direct test of the plausibility of the belief-attribution heuristic account. Second, it seems worthwhile testing for intentionality (and other mental states) ascriptions with scenarios where a praiseworthy agent brings about graded helpful and harmful outcomes—similar to what done by Nadelhoffer (2004b). This, we think, should provide us with further insights on the relation between outcomes, praise (and blame), and mental state ascriptions. Third, a different avenue for research, which could (also) potentially explain the findings of our study, has been set here: the relation between mental state ascriptions and gradable features of norms or norm violation. Finally, even when we find support for the *belief-attribution heuristic* and the *blame-driven bias* accounts, the findings here reported are not conclusive evidence in favor of any of the former. We thus recommend running further experiments and, more specifically, mediation analyses.

8. References

Adams, F., and Steadman, A. (2004a). Intentional action in ordinary language: Core concept or pragmatic understanding. *Analysis*, 64, 173-181.

Adams, F., and Steadman, A. (2004b). Intentional actions and moral considerations: Still pragmatic. *Analysis*, 64, 264-267.

Alfano, M., Beebe, J. R., & Robinson, B. (2012). The centrality of belief and reflection in Knobe-effect cases. *The Monist*, 95(2), 264–289.

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

Alicke, M. (2008). Blaming badly. *Journal of Cognition and Culture*, 8(1), 179–186.

Alicke, M., & Rose, D. (2010). Culpable control or moral concepts? *Behavioral and Brain Sciences,33*(4), 330-331.

Beebe, J.R (2013). A Knobe effect for belief ascriptions. Review of Philosophy and Psychology 4 (2): 235–258.

Beebe, J.R., and Buckwalter, M. (2010). The epistemic side-effect effect. *Mind & Languag*e, 25 (4), 474–498.

Beebe, J.R., and Jensen, M. (2012). Surprising connections between knowledge and action: The robustness of the epistemic side-effect effect. *Philosophical Psychology,* 25 (5), 689–715.

Cova, F., & Naar, H. (2012). Side-effect effect without side effects: The pervasive impact of moral considerations on judgments of intentionality. *Philosophical Psychology*, 25(6), 837–854.

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

Cushman, F., & Mele, A. (2008). Intentional action: Two and a half folk concepts? In J. Knobe, & S. Nichols (Eds.). Experimental philosophy (pp. 171–188). Oxford: Oxford University Press.

Egré, P. (2014). Intentional Action and the Semantics of Gradable Expressions (On the Knobe Effect). In *Causation in Grammatical Structures* (p. Causation in Grammatical Structures, Chapter 8). Oxford University Press.

Egré, P. & Cova, F. (2015). Moral asymmetries and the semantics of many. *Semantics and Pragmatics, 8*, 1-45.

Feltz, A. (2007). The Knobe effect: A brief overview. *Journal of Mind and Behavior*, 3(4), 265–277.

Hindriks, F. (2008). Intentional action and the praise-blame asymmetry. *The Philosophical Quarterly*, 58(233), 630–641.

Holton, R. (2010). Norms and the Knobe effect. *Analysis*, 70(3), 417–424.

Kneer, & Bourgeois-Gironde. (2017). Mens rea ascription, expertise and outcome effects: Professional judges surveyed. *Cognition*, 169, 139-146.

Kneer, M., & Machery, E. (2019). No luck for moral luck. *Cognition*, 182, 331–348.

Kneer, M., Hannikainen, I.R., Almeida, G., Aguiar, F., Bystranowski, P., Dranseika, V., Janik, B. M., Garcia Olier, J., Güver, L., Liefgreen, A., Tobia, K., Próchnicki, M., Rosas, A., Skoczén, I., Strohmaier, N. & Struchiner, N., (in preparation). Outcome effects on mental state ascriptions across cultures.

Knobe, J., & Burra, A. (2006). The folk concept of intention and intentional action: A cross-cultural study. *Journal of Culture and Cognition*, 6, 113–132.

Knobe, J., & Mendlow, G. (2004). The good, the bad, and the blameworthy: Understanding the role of evaluative reasoning in folk psychology. *Journal of Theoretical and Philosophical Psychology*, 24, 252–258.

Knobe, J. (2003a). Intentional Action and Side Effects in Ordinary Language. *Analysis*, 63(3), 190- 194.

Knobe, J. (2003b) Intentional action in folk psychology: An experimental investigation, Philosophical Psychology, 16:2, 309-324

Knobe, J. (2004a). Intention, intentional action and moral considerations. *Analysis*, 64(282), 181–187.

Knobe, J. (2006). The concept of intentional action: A case study in the uses of folk psychology. *Philosophical Studies*, 130(2), 203–231.

Knobe, J. (2007). Reason explanation in folk psychology. *Midwest Studies in Philosophy*, 31, 90–107.

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

Knobe, J., & Burra, A. (2006). The folk concepts of intention and intentional action: A cross-cultural study. *Journal of Cognition and Culture*, 6(1), 113–132.

Knobe, J., and Mendlow, G. (2004). The good, the bad and the blameworthy: understanding the role of evaluative reasoning in folk psychology. *Journal of Theoretical and Philosophical Psychology*, 24: 252-258

Lanteri, A. (2012). Three-and-a-half folk concepts of intentional action. *Philosophical Studies*, 158(1), 17–30.

Leslie, A., Knobe, J., & Cohen, A. (2006). Acting Intentionally and the Side-Effect Effect: Theory of Mind and Moral Judgment. *Psychological Science*, 17(5), 421-427.

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

Malle, B. (n.d). Graded representations of norm strength.

Malle, B. (2021). Moral Judgments. *Annual Review of Psychology*, 72:3, 1–26.

McCann, H. (2005). Intentional action and intending: Recent empirical studies. *Philosophical Psychology*, 18, 737–748.

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

Nadelhoffer, T. (2004b). On praise, side effects, and folk ascriptions of intentional action. *Journal of Theoretical and Philosophical Psychology*, 24, 196–213.

Nadelhoffer, T. (2005). Skill, luck, control, and intentional action. *Philosophical Psychology*, 18, 341–352.

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

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

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

Prochownik, K., Krebs, M., Wiegmann, A., & Horvath, J. (2020). CogSci 2020 Paper "Not as Bad as Painted? Legal Expertise, Intentionality Ascription, and Outcome Effects Revisited." Retrieved from osf.io/n9h2b

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Robbins, R., Shepard, J., & Rochat, P. (2017). Variations in judgments of intentional action and moral evaluation across eight cultures. *Cognition*, 164, 22–30.

Robinson, B., P. Stey, and M. Alfano. 2015. Reversing the side-effect effect: The power of salient norms. *Philosophical Studies* 172 (1): 177–206.

Tannenbaum, D., Ditto, P. H., and Pizarro, D. A. (2007) Different moral values produce different judgments of intentional action. Unpublished manuscript.

Tobia, K. (in preparation). Legal Concepts And Legal Expertise.

Uttich, K., & Lombrozo, T. (2010). Norms inform mental state ascriptions: A rational explanation for the side-effect effect. Cognition, 116(1), 87–100.

Wickham et al., (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686

Appendix. Vignettes for Experiment 2

1. **DAM**

Negative conditions (i.e., very bad and somewhat bad):

The mayor of a riverside town wants to build a new dam. One of his advisers approaches him and says: “Building a dam will surely increase the supply of electrical power to our town. However, there would be **[**severe / minor**]** adverse effects on the indigenous communities in the area. Their livelihoods will be [completely destroyed/ slightly affected]. They will be [displaced/ a bit disturbed]."

The mayor responds: “I do not care at all about the indigenous communities. All I want is to increase the supply of electrical power to our town. Let's build the dam.”

They build the dam. The livelihoods of the indigenous communities are [completely destroyed/slightly affected]. The communities are [displaced/a bit disturbed].

Neutral condition:

The mayor of a riverside town wants to build a new dam. One of his advisers approaches him and says: “Building a dam will surely increase the supply of electrical power to our town. There would be no effects on the indigenous communities in the area whatsoever."

The Mayor responds: “I do not care at all about the indigenous communities. All I want is to increase the supply of electrical power to our town. Let's build the dam.”

They build the dam. There are no effects on the indigenous communities.

Positive conditions (i.e., very good and somewhat good):

The mayor of a riverside town wants to build a new dam. One of his advisers approaches him and says: “Building a dam will surely increase the supply of electrical power to our town. There would also be [minor/very] positive effects on the indigenous communities in the area. Their livelihoods will improve [slightly/greatly]. They will benefit [somewhat/strongly] from this."

The mayor responds: “I do not care at all about the indigenous communities. All I want is to increase the supply of electrical power to our town. Let's build the dam.”

They build the dam. The livelihoods of the indigenous communities improve [slightly/greatly]. The communities benefit [somewhat/strongly] from this.

2. MALL

Negative conditions (i.e., very bad and somewhat bad):

The CEO of a construction company is planning on building a mall in a small town. The vice president of the company approaches him and says: “Building a mall will definitely increase our profits. However, there would be [severe/minor] adverse effects on the local businesses outside the mall. There will be a [drastic/small] decline in their customers. [They go bankrupt/Their profits will slightly decrease].”

The CEO responds: “I do not care at all about the businesses outside the mall. All I want is to make profit. Let's build the mall.”

They build the mall. The local businesses outside the mall report a [drastic/small] decline in customers. [They go bankrupt/Their profits slightly decrease].

Neutral condition:

The CEO of a construction company is planning on building a mall in a small town. The vice president of the company approaches him and says: “Building a mall will definitely increase our profits. There would be no effects on the local businesses outside the mall whatsoever.

The CEO responds: “I do not care at all about the businesses outside the mall. All I want is to make profit. Let's build the mall.”

They build the mall. There are no effects on the local businesses outside the mall.

Positive conditions (i.e., very good and somewhat good):

The CEO of a construction company is planning on building a mall in a small town. The vice president of the company approaches him and says: “Building a mall will definitely increase our profits. There would also be [minor/very] positive effects on the local businesses outside the mall. There will be a [small/considerable] increase in their customers. Their profits will [slightly/greatly] increase.”

The CEO responds: “I do not care at all about the businesses outside the mall. All I want is to make profit. Let’s build the mall.”

They build the mall. The local businesses outside the mall report a [small/considerable] increase in customers. Their profits increase [greatly/slightly].

3. PUBLIC IMAGE

Negative conditions (i.e., very bad and somewhat bad):

The administrator of a State agency is approached by his advisor, who says: “Implementing cutback measures will save the agency a considerable amount of money. However, there would be [severe/minor] adverse effects on the agency’s public image. It will become a [very/bit] negative.”

The administrator responds: “I do not care at all about the agency's public image. All I want is to save some money. Let's implement the cutback measures.”

They implement the cutback measures. The agency’s public image is a [very/bit] negative.

Neutral condition:

The administrator of a State agency is approached by his advisor, who says: “Implementing cutback measures will save the agency a considerable amount of money. There would be no effects on the agency’s public image whatsoever.”

The administrator responds: “I do not care at all about the agency's public

image. All I want is to save some money. Let's implement the cutback measures."

They implement the cutback measures. There are no effects on the agency's public image.

Positive conditions (i.e., very good and somewhat good):

The administrator of a State agency is approached by his advisor, who says: "Implementing cutback measures will save the agency a considerable amount of money. There would also be [minor/very] positive effects on the agency's public image. It will become [slightly/very] positive."

The administrator responds: "I do not care at all about the agency's public image. All I want is to save some money. Let's implement the cutback measures."

They implement the cutback measures. The agency's public image is [slightly/very] positive.