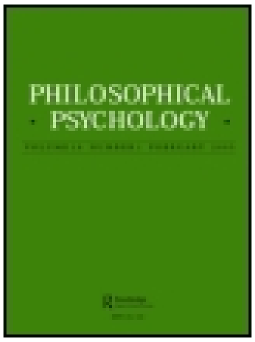

PHILOSOPHICAL
PSYCHOLOGY

Philosophical Psychology

ISSN: 0951-5089 (Print) 1465-394X (Online) Journal homepage: http://www.tandfonline.com/loi/cphp20

Objective and epistemic gradability: Is the new angle on the Knobe effect empirically grounded?

Tomasz Zyglewicz & Bartosz Maćkiewicz

To cite this article: Tomasz Zyglewicz & Bartosz Maćkiewicz (2018): Objective and epistemic gradability: Is the new angle on the Knobe effect empirically grounded?, Philosophical Psychology, DOI: 10.1080/09515089.2018.1540776

To link to this article: https://doi.org/10.1080/09515089.2018.1540776

Published online: 07 Nov 2018.

PHILOSOPHICAL PSYCHOLOGY
https://doi.org/10.1080/09515089.2018.1540776

Routledge
Taylor & Francis Group

Objective and epistemic gradability: Is the new angle on the Knobe effect empirically grounded?

Tomasz Zyglewicz and Bartosz Maćkiewicz

Institute of Philosophy, Warsaw University, Warsaw, Poland

ABSTRACT

According to the New Angle, any explanation of the Knobe effect must be gradable and asymmetric. It has been argued that only Hindriks' approach meets both criteria. First, we argue that Holton's hypothesis also meets the criteria. Second, we show that the authors are not justified in taking the criteria to be empirically justified. We have failed to replicate the asymmetry result in two experiments. Moreover, gradability can be objective or epistemic. We show that the New Angle presupposes objective gradability. In our experiments, the patterns of responses to questions about epistemic and objective gradability are the same, irrespective of whether the feature is objectively gradable (e.g., blameworthiness) or not (e.g., intentionality). Our results thus question the extent to which the New Angle is empirically grounded. Moreover, they raise doubt whether the answers to questions about epistemic and objective gradability can be taken at face value at all.

Abbreviations: NRH - normative reasons hypothesis; NVH - norm violation hypothesis; DQ - degree question; DAQ - degree of agreement question

ARTICLE HISTORY
Received 19 December 2017
Accepted 20 June 2018

KEYWORDS
Epistemic/objective gradability; experimental philosophy; intentionality; Knobe effect; new angle; normative reasons hypothesis

1. Introduction

In their recent paper, Hindriks, Douven, and Singmann (2016) suggest a new approach to the Knobe effect. They reran Knobe's (2003) original scenario and applied subtler statistical means, the logistic regression, to the data. As a result of this analysis, they formulated two desiderata to be met by any plausible explanation of the Knobe effect: *gradability* and *asymmetry*. Moreover, they argued that, among the prominent accounts of the Knobe effect, only Hindriks's Normative Reasons Hypothesis meets both of these criteria. We question their conclusions. First, we argue that not only Hindriks's approach meets the two criteria. We take Holton's Norm Violation Hypothesis (Holton, 2010) as an example and argue that, contrary to what the authors claim, there are very good reasons to think that

CONTACT Tomasz Zyglewicz ✉ tzyglewicz@gradcenter.cuny.edu Philosophy Program, Graduate Center, City University of New York, 365 Fifth Avenue, New York, NY 10016, USA

norm violation is gradable (section 2). Second, we show that there are problems with the authors' argument that the two criteria are empirically grounded. As far as the asymmetry criterion is concerned, we have failed to replicate the asymmetry results in two studies (sections 4 and 5). Our results (also analyzed by means of logistic regression) do not show that "the degree of responsibility ascribed correlates with the chance with which intentionality is attributed in the harm condition, but not in the help condition" (Hindriks et al., 2016, p. 211). As far as the gradability criterion is concerned, we show that there is a serious problem with it (section 3). The authors have altogether neglected the possibility that gradability need not be objective in character (in that it reflects the extent to which a feature pertains to an object) but rather that it could be merely epistemic (in that it reflects the extent to which a person agrees that an object has a certain feature). The authors' argument crucially relies on the assumption that the data they have gathered reflect objective gradability. We call this assumption into question in two experiments (sections 5 and 6).

2. The new angle on the knobe effect

Hindriks and colleagues' experiment was basically a replication of Knobe's (2003) chairman scenario with two small modifications. First, their question about intentionality was: "Did the chairman intentionally bring about the effect on the environment?" They departed from the original question by using a neutral formulation, devoid of morally loaded verbs such as "harm" and "help." Second, they asked a single blame/praise question across the scenarios: "In your opinion, how blameworthy or praiseworthy is the chairman, given that his decision affected the environment?" The answers comprised a seven-point pseudo-Likert scale: *very blameworthy, blameworthy, somewhat blameworthy, neither blameworthy nor praiseworthy, somewhat praiseworthy, praiseworthy, very praiseworthy*.

Despite the differences, their experiment confirmed the existence of two asymmetries already present in Knobe's (2003) study: the Knobe effect, which is the asymmetry in intentionality ascriptions between the harm and the help condition, and what Hindriks (2008) calls "the Praise-Blame Asymmetry," the asymmetry between the amount of blame/praise ascribed in the respective conditions. Like Knobe, Hindriks and colleagues reported a general correlation between intentionality ascriptions and the amount of blame/praise ascribed.

The central novelty in their data analysis is that they investigated the correlation between intentionality and blame/praise ascriptions for each of the conditions separately. Moreover, as the answers to the intentionality question generate a binary variable, they decided that logistic regression would be the most suitable method for this purpose (Hindriks et al., 2016,

p. 209). The upshot of their analysis is the formulation of the following two theses:

> **HARM**. The more blame that participants attribute in the harm condition, the higher the chance that they take the indifferent agent to have acted intentionally.
>
> **HELP**. The chance that participants attribute intentionality to the indifferent agent in the help condition is low, irrespective of the praise or blame they assign. (p. 211)

According to Hindriks and colleagues, the lesson to be drawn from the data is that any plausible explanation of the Knobe effect must meet the following two desiderata:

> **Gradability**. The property it invokes must be *gradable* to account for the HARM thesis.
>
> **Asymmetry**. It must explain why there is a correlation between intentionality ascriptions and blameworthiness (HARM), whereas there is no correlation between intentionality ascriptions and praiseworthiness (HELP).

Even though these criteria do not seem to be too strict at first glance, Hindriks and colleagues believe that they suffice to show that in fact none of the proposed accounts of the Knobe effect meet both except for Hindriks' Normative Reasons Hypothesis (NRH).

According to NRH, the Knobe effect is to be explained in terms of the agent's *indifference toward normative reasons*. A normative reason is "a consideration which counts in favor or against an action and which one should take into account when considering it, irrespective of the question whether one does so" (Hindriks, 2008, p. 633). It is to be distinguished from motivating reasons, which are considerations that the agent actually does take into account, that actually do move the agent to act. In Knobe's scenarios, the prospect of increasing profits is both a normative and a motivating reason. It is a normative reason because we generally accept that a chairman should be concerned about the financial results of his company. It is a motivating reason because the chairman explicitly justifies his decision by admitting that this is something he cares about. By contrast, the prospect of harming (helping) the environment is a normative reason but is not a motivating reason in Knobe's scenarios. The prospect of harming (helping) the environment is generally recognized as something that one should take into account. However, in Knobe's stories, it is not a consideration that motivates the chairman, who claims to be indifferent to it ("I don't care about the environment at all").

Hindriks believes that the agent's indifference toward normative reasons is the factor that "blocks praise, but does not block blame" (Hindriks, 2008, p. 633). This suffices to ensure that the asymmetry criterion is satisfied. The more problematic part is whether NRH is capable of accommodating the gradability criterion. After all, one may doubt whether the agent's indifference is gradable. Hindriks and colleagues believe, however, that NRH does satisfy the gradability requirement. In fact, they point to two features invoked in the explanation that are gradable: the agent's indifference and the normative reasons themselves. First, they argue that, contrary to appearances, indifference is gradable. They notice that we do speak of people being "indifferent to some issue to some degree" (Hindriks et al., 2016, p. 216). In addition, they claim that indifference is a propositional attitude and propositional attitudes are gradable. Second, they claim that gradability is also associated with normative reasons (how much an agent should care about the rule of etiquette, the provision of criminal code, and a moral rule is a matter of degree).

Hindriks and colleagues thus provide an explanation of the HARM and the HELP theses in terms of NRH:

> **NRH*-HARM**. The larger the discrepancy between how much the agent should care about the harmful side effect and how much she actually cares about it, the higher the chance that people attribute intentionality.
> **NRH*-HELP**. The chance someone attributes intentionality to the agent is independent of how much she should care about the beneficial side effect. (p. 217)

The authors believe that no rival explanation of the Knobe effect admits of a reformulation along these lines. In other words, they claim that all other accounts fail to accommodate either the gradability involved in the HARM thesis or the asymmetry between the HARM and the HELP theses.[1] The upshot of their paper is that, despite the lack of direct empirical evidence in its favor (or even a clear idea how to provide it), NRH is the only plausible explanation of the Knobe effect (pp. 217–218).

Before proceeding, let us confront an objection one might put forward.[2] One might think that the asymmetry thesis has already been falsified because it has been shown that high intentionality attributions still occur even when there is no blame involved (Cova, 2017; Knobe, 2006; Knobe & Mendlow, 2004; Wright & Bengson, 2009). Knobe's *Sales Vignette* (Knobe & Mendlow, 2004) depicts a situation analogous to the well-known original story, except for the fact that the main effect of implementing a new program is an increase of sales in Massachusetts and the side effect is a decrease of sales in New Jersey. Almost all of the participants confronted with this vignette say that the CEO is neither blameworthy nor praiseworthy (80%) and yet they still say that she decreased sales in New Jersey intentionally (75%).

Two points should be raised, however. First, although the language employed is often misleading, Hindriks and colleagues do not claim that there is a causal link between attributions of blame and intentionality. In NRH, the main explanatory work is done by the notion of degree of indifference. People take the chairman to have harmed the environment intentionally because he is indifferent.[3] Similarly, people take the chairman to be blameworthy because of his indifference.[4] Thus, the correlation between blame ratings and intentionality attribution observed in Knobe's scenarios is to be explained by a common cause – the degree of indifference toward the bad outcome of an agent's action.

Second, Hindriks actually does consider the sales case (Hindriks, 2011, 2018). In his most recent work (Hindriks, 2018), he argues that the case differs from the rest of Knobe-like scenarios with respect to the attitude of the protagonist. In contrast to environmental vignette, the CEO in the sales case can be described as having a pro-attitude toward the side-effect (decreasing sales), which explains why she brought it about intentionally.

One can, of course, engage in a discussion with Hindriks' position (it is not clear, for example, that in the sales case the CEO wants to decrease sales in New Jersey). However, the charge that his position has already been falsified is premature. We will later (sections 4 and 5) present the results of two experiments which failed to reproduce the asymmetry result.

3. Gradable norm violation

Richard Holton's (2010) Norm Violation Hypothesis (NVH) is founded on the observation that there is an asymmetry between intentionally violating a norm and intentionally conforming to it. To intentionally violate a norm it suffices that one knowingly violates it, whereas to intentionally conform to a norm one needs to be counterfactually guided by it (Holton, 2010, p. 418). Because the asymmetry in the intentionality ascription is to be explained in terms of different requirements involved in the agent's intentionally conforming to (or violating) a norm, NVH is capable of accounting for the asymmetry between the HARM and the HELP theses. However, Hindriks and colleagues claim that NVH does not satisfy the gradability criterion. They claim that "it is difficult to see what more or less violating a norm would mean in this context" (Hindriks et al., 2016, p. 215).

However, this claim is questionable. Not only is it relatively easy to see that violating a norm can be gradable, but the fact that it is gradable is well entrenched in our practices. First, legislatures across the world have already developed vocabulary to trace the degrees of seriousness of norm violations. It is common to distinguish, for example, insignificant, minor, major, or gross violations of a legal norm. In fact, the distinction is important, for instance, in penal codes, where different penalties are associated with different degrees of norm violation. This already suffices to establish that there is a sense in which norm violation can be conceived of as a gradable property. Moreover, norms have different scopes. Some norms (e.g., "do not press the red button under any circumstances") have a relatively narrow scope and can be violated by relatively few action types (either pressing the button or refraining from doing so). Norms with a wider scope, on the other hand, may be violated by very many different types of actions, which is conducive to the introduction of gradation of how seriously the norm is violated. For example, the norm "thou shalt not harm thy neighbor" can be violated by an overfriendly punch on an arm, by carelessly stomping on a foot, by cutting off a hand, and so on. There is a clear sense in which the latter are more severe violations than the former.

Second, it has been argued that the gradability of norm violation underlies the very possibility of resolving conflicts of norms.[5] Typically, the problem is how to justify violating a higher-level norm while conforming to a lower-level norm. The gradability of norm violation allows a solution to the problem, for one can argue that the violation of a higher-level norm in order to conform to a lower-level norm is justified when the violation of the latter would be far more severe. For instance, most people take the norm "help your family" (F) to be more important than the norm "help strangers" (S). Nonetheless, most of us would think it appropriate to violate (F) by not helping one's mother cook the dinner to conform to (S) by calling the police upon seeing a stranger being assaulted across the street. This seems morally acceptable precisely because the violation of the less important norm (S) (not calling the police) would be more severe than the violation of more important norm (F) (not helping mother in cooking the dinner).

One could object that such a response to Hindriks and colleagues misses their point. They may grant that norm violation is gradable in general but still argue that it is not gradable in the particular context of Knobe's chairman scenario: the chairman either violates the norm of not harming the environment (by starting the program) or not (by not starting the program).[6] We believe, however, that the general point applies in this particular scenario as well. It seems quite unquestionable that spilling toxic waste into a river is a more severe violation of the norm than dumping organic waste into the river, which is a more severe violation than failure to use energy-efficient lightbulbs, and so on. Because the chairman scenario does not specify what sort of harm to the environment was involved, it is perfectly conceivable that different participants picture different degrees of norm-violating consequences and thereby assess the extent to which chairman's action violates the salient norm differently.

If Holton's NVH is supplemented with a plausible extension,[7] according to which a participant's perceived degree of norm violation is reflected in the amount of blame she ascribes to the chairman, NVH can be reformulated to fit both of the criteria posited by Hindriks and colleagues:

> **NVH*-Harm**. The greater the perceived violation of a relevant norm, the higher the chance that people attribute intentionality.
> **NVH*-Help**. The chance someone attributes intentionality to the agent is independent of the extent to which he or she conforms to a relevant norm.

Holton's NVH* is thus as plausible an explanation of the Knobe effect in light of Hindriks and colleagues' standards as is Hindriks' NRH because it meets both of their criteria.[8] The authors' conclusion that NRH is the only plausible account of the Knobe effect is thus premature.

4. Two kinds of gradability

Let us now turn to the question of whether the data presented by Hindriks and colleagues actually do justify their methodological conclusions, that is, the inclusion of the criteria of gradability and asymmetry. We will start by showing that there is a problem of whether the data the authors rely on justify the gradability criterion.

The authors assume that the gradability exhibited in Knobe-like scenarios reflects the gradable character of a certain actual feature, namely blameworthiness. They do not consider the possibility that the gradability they discovered might not be objective (associated with the fact that an object exhibits a property to a certain degree) but rather epistemic (it may reflect the degree to which the subjects are confident that a certain property is present).

There are clearly properties that we think of as gradable: height, stiffness, kindness, intelligence, charm, precision, and so on. In these cases, we think that objects may have these properties in different degrees. It is clear that we also think that the confidence with which we make judgments is gradable. In these cases, what admits of degrees is not a property that an object exhibits but rather the confidence with which we make certain claims. Let us speak of "objective" gradability in the first case and of "epistemic" gradability in the second case. We should thus distinguish the question concerning the degree to which a certain object exhibits a certain property (let us call it the degree question or DQ, for short) from the question concerning the degree to which a person agrees that a certain object exhibits a certain property (let us call it the degree of agreement question or DAQ, for short).

Moreover, it can be argued that the answers to these two types of questions need not be the same. In the following example, a person exhibits a high degree of confidence that an object has a certain property but does not attribute a high degree of that property to the object. Bob is not colorblind but the description of visual experiences is not really his strongest point. Emma has told him to buy some butter in the supermarket. However, there is only one stick left and Bob calls Emma to ask whether he should buy it.

Emma: Has it gone bad?
Bob: How can I tell?
E.: Is it yellow?
B.: Yes, certainly.
E.: How yellow is it?
B.: I'm not really quite sure.

Bob is certain that the butter is yellow, so the epistemic gradability exhibited in the answer to the degree of agreement question is very high. However, the objective gradability exhibited in the answer to the degree question is not high. Bob is not sure what the degree of yellowness of the butter is. In fact, in this case we might even doubt that Bob considers yellowness to be gradable.

Now, Hindriks and colleagues make an implicit, yet nontrivial, assumption that the gradability that comes into play in Knobe-like scenarios is objective. In fact, it is crucial to their argument that the gradability not be epistemic. If the gradable patterns of responses could be understood in terms of epistemic gradability, then they could not discard rival accounts of the Knobe effect on the basis of their not evoking a gradable property. If the gradable pattern of responses in the harm condition reflected epistemic gradability (the degree of respondents' confidence rather than the degree of a feature), then *any* explanation of the Knobe effect whatsoever could appeal to epistemic gradability. People can exhibit different degrees of confidence even with respect to features that are not (objectively) gradable. In other words, to use gradability as a criterion for the exclusion of rival hypotheses, they would have to show that the gradability exhibited in their data is not epistemic.

5. Experiment 1

In our experiments, we will show that the two criteria introduced by Hindriks and colleagues are not empirically justified. We begin by presenting the results of an experiment, which fails to replicate the asymmetry result.

The main aim of the experiment was to establish whether the Knobe effect is present in Polish.[9] However, because the ratings of blame/praise were also collected, it was possible to conduct a statistical analysis analogous to that of Hindriks and colleagues.

5.1. Method

The participants were recruited through social media as well as by means of various academic mailing lists. The sample consisted of 1074 native Polish speakers (679 females; average age: 21.07). The participants were not paid.

5.2. Materials and procedures

The original Knobe vignettes were translated into Polish by our research team. The participants were asked whether the CEO harmed or helped the environment "intentionally."

However, while there are many adverbs that can be thought of as attributing various elements of intentionality in Polish, there is no single adverb that has the same meaning as the English "intentionally." We have investigated five common adverbs that are used. There are two related adverbs *celowo* and *specjalnie*, which can be thought of as translations of the English *purposefully* (however, while ϕ-ing *specjalnie* implies that the end was to ϕ, ϕ-ing *celowo* can be used when ϕ-ing is a means to a different end). There is an adverb *umyślnie* (literally translated: in a mindful fashion), which is perhaps closest to the English *intentionally* in that it is used not only in cases where one intends to do something but also in cases where one does something without intending to do it but while foreseeing it. The problem is that unlike the English *intentionally*, the Polish *umyślnie* is applicable exclusively in negative contexts. When the adverb is used in a positive context (e.g., in any kind of helping scenario), it immediately calls to mind a negative interpretation. Another adverb phrase *z rozmysłem* corresponds rather closely to the English *deliberately* (it literally means: with deliberation). The final very commonly used adverb *świadomie* corresponds to the English *knowingly* (it literally means: consciously, with awareness).

Our study had a between-subject 2 (harm vs. help) × 6 design: in five conditions we applied the five different intentionality adverbs and in one condition we did not apply any adverb: the participants were asked whether the chairman harmed/helped the environment. In addition, we checked the usage of two verbs: *chcieć* (want) and *zamierzać* (intend), that is, we asked whether the chairman wanted/intended to harm/help the environment. (It is noteworthy that there is no problem with the translation of these verbs into Polish.).

On a separate screen, with the vignette still visible, the participants were then asked to assess how blameworthy or praiseworthy the CEO was. In the harm condition, they could choose between the following answers:

(1) Blameworthy
(2) Rather blameworthy
(3) Hard to say
(4) Rather not blameworthy
(5) Not blameworthy

In the help condition the answers were:

(1) Praiseworthy
(2) Rather praiseworthy
(3) Hard to say
(4) Rather not praiseworthy
(5) Not praiseworthy

The answers were coded as numbers from 1 to 5 (the numbers were not visible to the respondents).

5.3. Results and discussion

In all groups, there is a statistically significant difference between the harm and the help condition (see Table 1).

For the purposes of running logistic regression, we have merged the results from all groups in order to achieve greater statistical power. We conducted the same statistical analysis as Hindriks and colleagues. The authors' argument for the new kind of asymmetry requires that the data meet the following two conditions:

Table 1. The positive responses in experiment I (in Polish).

Intentionality phrase	Harm	Help	χ2(1)	Φ (effect size)	P-value
No adverb	93.5% (58)	77.8% (56)	5.34	0.22	0.02
Purposefully 1 (*celowo*)	51.4% (36)	11% (8)	25.61	0.44	< 0.001
Purposefully 2 (*specjalnie*)	51.5% (35)	5.2% (4)	37.01	0.52	< 0.001
Intentionally$_{neg}$ (*umyślnie*)	80.3% (57)	19.7% (13)	47.85	0.60	< 0.001
Deliberately (*z rozmysłem*)	83.9% (47)	6.3% (4)	69.73	0.78	< 0.001
Knowingly (*świadomie*)	89.6% (60)	51.5 (34)	21.41	0.42	< 0.001
Intended (*zamierzał*)	38.6% (27)	4.5% (3)	20.94	0.41	< 0.001
Wanted (*chciał*)	35.6% (21)	2.9% (2)	20.56	0.41	< 0.001

(1) There should be an interaction between the effect of the experimental condition (moral valence) on the ascription of intentionality and the effect of the praise/blame ascription on the ascription of intentionality.
(2) Post-hoc trend analysis should reveal a statistically significant relationship between blame ascription and intentionality attribution in the harm condition, whereas in the help condition there should be no statistically significant relationship between the praise ascription and the intentionality attribution.

The procedure was to run regression analysis with the two variables of interest together with their interaction – the experimental condition and the assessment of praise or blame. The blame/praise attribution *does predict* the intentionality ascription. Statistical tests reveal that the relationship is significant for the experimental condition (β = 2.35, $p < 0.001$) as well as for the blame/praise ascription (β = −0.78, $p < 0.001$). Their interaction also reached statistical significance (β = 0.61, p = 0.005).

Post-hoc trend analysis (for each experimental condition separately) reveals that there is a statistically significant correlation of the blame or praise assessment and the intentionality ascription in both the harm and the help conditions. We found that the ascription of blame/praise predicts intentionality judgments not only in the harm condition (β = −0.78, $p < 0.001$), but also in the help condition (β = −0.165, p = 0.018). Our data thus do not confirm the conclusions drawn by Hindriks and colleagues that there is an asymmetry between the harm condition, in which the intentionality attribution *is* predicted by the assessment of blame, and the help condition, in which the intentionality attribution *is not* predicted by the assessment of praise.

The data from this experiment show that even though both praise and blame are correlated with the intentionality attribution, the strength of this correlation varies. Still, the results straightforwardly undermine the conclusions drawn by Hindriks and colleagues. Given the data, the HELP thesis simply does not obtain.

One possible problem with the presented experiment is that it was not run in English. We have also noted the problems with translating *intentionally*. One could thus argue that our failure to find the asymmetry cannot undermine the conclusions drawn by Hindriks and colleagues. The remaining experiments were conducted in English.

6. Experiment 2

We have run an experiment to see whether there is a difference between objective and epistemic gradability in the cases at hand. We also wanted to see whether the results of Hindriks and colleagues' experiment would be replicated, that is, whether the asymmetry between the correlations would be observed. This time we followed the experimental design of the study by Hindriks and colleagues more closely to ensure that the studies were comparable.

The question the authors asked was: "In your opinion, how blameworthy or praiseworthy is the chairman, given that his decision affected the environment?" This is what we have called the degree question (DQ), which asks about the objective gradability of blameworthiness or praiseworthiness. We reran the experiment and added a question about epistemic gradability (the degree of agreement question, DAQ): "To what extent do you agree that the chairman is blameworthy or praiseworthy, given that his decision affected the environment?"

The aim of our experiment was to see whether the patterns of responses to the two questions would be different enough to support the assumption that the data presented by Hindriks and colleagues reflect objective rather than epistemic gradability. This would be the case, for example, if DAQ generated a different pattern of responses than that generated by DQ. It would also be the case if one could discover an order effect which might suggest that, for instance, objective gradability is more basic than epistemic gradability. For example, if responses to DAQ but not to DQ turned out to be sensitive to order, one could argue that DQ is a primary or a "more obvious" reading of the question. This would provide support for the conclusions drawn by Hindriks and colleagues.

6.1. Method

The participants were recruited by Amazon MTurk (small compensation was provided) from native English speakers. Our sample consisted of 306 subjects (174 females; average age: 31.5). All our studies were conducted online using our own instance of the LimeSurvey service.

6.2. Materials and procedures

Our first experiment was a slightly modified replication of the experiment conducted by Hindriks and colleagues. Subjects were presented with the original Knobe vignettes and asked about the intentionality of the action and the degree of blame or praise ascribed to the chairman. Our formulation of the first two questions and the scale followed Hindriks and colleagues. The first (yes/no) question was "Did the chairman bring about the effect on the environment intentionally?" The second question was the degree question (DQ): "In your opinion, how blameworthy or praiseworthy is the chairman, given that his decision affected the environment?" to which the following answers were possible:

- Very blameworthy
- Blameworthy
- Somewhat blameworthy
- Neither blameworthy nor praiseworthy
- Somewhat praiseworthy
- Praiseworthy
- Very praiseworthy

For the purposes of statistical analysis, answers were coded as numbers from 1 to 7. In addition to these two questions, we asked a degree of agreement question (DAQ): "To what extent do you agree that the chairman is blameworthy or praiseworthy, given that his decision affected the environment?" We used a pseudo-Likert scale again, with the following response options:

- I strongly agree that the chairman is blameworthy.
- I agree that the chairman is blameworthy.
- I somewhat agree that the chairman is blameworthy.
- I neither agree that the chairman is blameworthy nor that he is praiseworthy.
- I somewhat agree that the chairman is praiseworthy.
- I agree that the chairman is praiseworthy.
- I strongly agree that the chairman is praiseworthy.

The answers were also coded as numbers from 1 to 7. Each question, accompanied by the vignette, was shown on a separate screen and there was no possibility of going back to previous questions. Each participant was assigned to one experimental condition – harm or help – and to one of two groups that differed in the order in which the second and the third question were asked.

6.3. Results and discussion

The original Knobe effect was successfully replicated. Eighty-five percent of the participants ascribe intentionality to the chairman's bringing about the effect on the environment in the harm condition, whereas in the help condition only 25% agree that he did it intentionally (χ^2 = 108.73, p < 0.001, φ = *0.6*). With regard to the degree question DQ, participants say that the chairman is blameworthy in the harm condition (M = 1.56, SD = 0.87), but in the help condition the answers are very close to the mid-point (M = 3.69, SD = 1.25). The results for the degree of agreement question DAQ are not significantly different (harm condition: U = 12,177, p = 0.99; help condition: U = 11,038, p = 0.77). Again, in the harm condition participants overall strongly agree that the chairman is blameworthy (M = 1.6, SD = 0.99) but in the help condition they do not have a very clear opinion on this issue (M = 3.66, SD = 1.16).

6.3.1. Relationship between the assessment of blame or praise and the intentionality ascription

Again we have conducted the same statistical analysis as Hindriks and colleagues. Statistical tests reveal a statistically significant effect of the experimental condition ($\beta = -2.45$, $p < 0.001$) and of the blame/praise attribution ($\beta = -0.52$, $p = 0.02$). Thus far the results are not very different from those obtained by Hindriks and colleagues. The first important difference is that there is no interaction between these two variables ($p = 0.53$), which means that condition (1) concerning the interaction effect between two variables (see section 4) is not met.

When it comes to post-hoc analysis (condition [2]), once again the ascription of blame/praise predicts the intentionality judgments in both experimental conditions analyzed separately (the harm condition: $\beta = -0.51$, $p = 0.02$; the help condition: $\beta = -0.36$, $p = 0.04$). The data of our experiment in English also undermine the conclusions drawn by Hindriks and colleagues that there is an asymmetry between the harm condition and the help condition.

The more general point about the results is related to the fact that effect size in the help condition was quite small. The problem with small effect sizes is that they are not easily detectable when small sample sizes are used. We ran a simulation study to investigate how large a sample is required to detect a correlational pattern in the help condition. We assumed that the distribution of responses in the population reflects the distribution in our sample from Experiment 2. We used the bootstrap approach (e.g., Kleinman & Huang, 2016) to estimate the statistical power of logistic regression for various sample sizes. For N = 150 (the sample size used by Hindriks and colleagues for each experimental condition) the statistical power was only 0.6. It seems that Hindriks' study, as well as ours, was somewhat underpowered if the obtained data acceptably estimate the true distribution of responses in the population. To achieve the power of 0.8 (which is considered to be a reasonable level), one would require approximately 250 participants for each condition. Note that Experiment 1 had many more participants but the groups were not homogenous with respect to the adverb used to express the intentionality of an action.

One final remark must be made with regard to the issue of the methodology of experimental research and the replicability of studies in experimental philosophy. One of the striking features of Hindriks and colleagues' paper is that the whole line of argumentation is based on only one study. The first lesson to draw from our discussion is that making very general and strong claims requires robust empirical support, which can be achieved by conducting several studies tackling the research problem from different perspectives. The second lesson is that these studies should be replicable. Our results from experiments 1 and 2 show that Hindriks and colleagues' findings are not sufficiently robust. It is worthwhile to point out that the problem of replicability is recognized by the community of experimental philosophers, as indicated by existence of The XPhi Replicability Project (https://sites.google.com/site/thexphireplicabilityproject/).

6.3.2. *Relationship between the degree question (DQ) and the degree of agreement question (DAQ)*

The answers to our two gradability questions are strongly correlated. In all subgroups combined, the correlation is very high ($r = 0.88$, $p < 0.001$). The correlation computed separately for each subgroup is slightly lower but still very high ($r \approx 0.75$, statistically significant for $\alpha = 0.05$ across all subgroups). See Figure 1. It thus seems that the experimental results do not distinguish between objective and epistemic gradability.

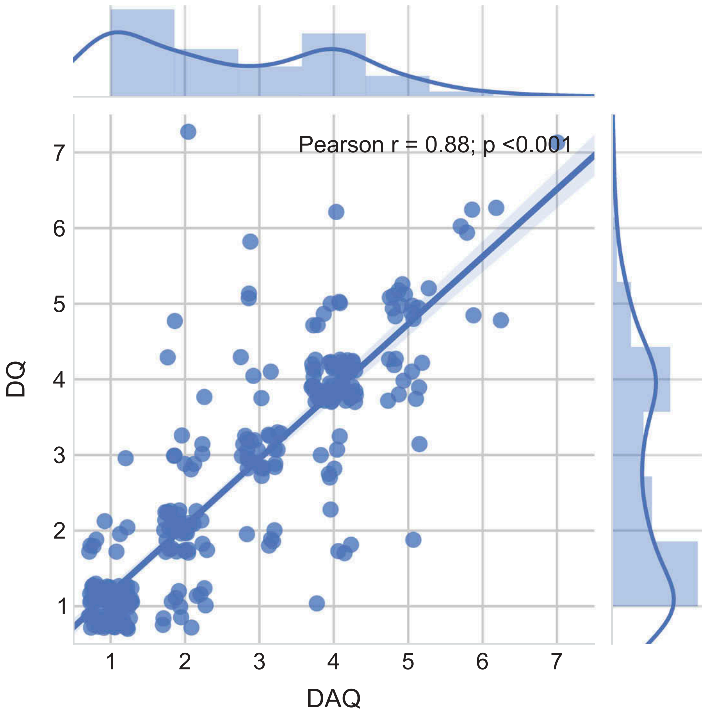

Figure 1. The correlation between the answers to DQ and DAQ questions about blameworthiness/praiseworthiness (Experiment 2) in all groups combined. Small jitter was added to improve readability.

Statistical tests do not reveal any kind of order effect between two questions ($p > 0.05$). We have run Mann–Whitney U test (due to non-normal distribution of answers). We have also run the t-test, but it has shown an even greater insignificance (for $\alpha = 0.05$).

6.4. Summary

The main finding of Hindriks and colleagues was a new kind of asymmetry expressed by the HARM and HELP theses. They claim that the attribution of intentionality is predicted by the amount of blame ascribed to the agent in the harm condition but that the attribution of intentionality is not predicted by the amount of praise ascribed in the help condition.

Our results show that there is a statistically significant relation between praise/blame and intentionality regardless of the moral valence of the side effect. There is thus no asymmetry of the kind postulated by the authors between the effects of the amount of praise or blame ascribed to the agent and the intentionality attribution in the harm and help scenarios. First, there is no interaction between the praise/blame ascription and the experimental condition. The first premise of the authors' argument for asymmetry is thus false. Second, post-hoc analysis reveals that in both conditions there is a statistically significant relationship between the intentionality attribution and the amount of blame/praise ascribed to the agent. Their second premise is thus also false. Of course, one could still argue that there is an asymmetry between the harm and the help condition in this respect. We have only shown that it cannot be demonstrated in the way that Hindriks and colleagues have proposed.

The results further indicate that there are no major differences in how the objective gradability and epistemic gradability questions are answered. Hindriks and colleagues assume that the gradability results express the objective gradability of blameworthiness rather than the epistemic gradability of the subjects' agreement with an attribution of blameworthiness. As we argued, this assumption is crucial in their argument for the exclusion of competing accounts of the Knobe effect. Our data show that one cannot decide whether the question they used really does accomplish it. The strong correlation between the answers to both questions casts doubt on the possibility of experimentally measuring one of them.

One may object, however, that the two judgments (about the objective gradability and the epistemic gradability) may very well coincide in some cases. When people attribute a high degree of blameworthiness, they may also agree to a high degree with the attribution of blameworthiness, and vice versa. The objection is well taken. It may be that there simply *is* a correlation between the answers to the two questions in the case at hand. While this is a possible interpretation of the results, another possible interpretation is that people do not actually distinguish between one and the other.[10] As we pointed out, however, the mere fact that it is *possible* to interpret the data as data about objective gradability is insufficient for the authors' purposes. Their whole argument for the exclusion of alternative accounts of the Knobe effect rests on the claim that their results reveal objective gradability. We have run another experiment to show that the contrary interpretation of the data is not as far-fetched as it might at first appear.

7. Experiment 3

We have argued that Hindriks and colleagues are not entitled to the conclusions they draw from the data they have gathered because they have no way of distinguishing the objective gradability presupposed by their conclusions from epistemic gradability. In Experiment 2, we asked questions about the objective and the epistemic gradability of a feature, which is in fact objectively gradable (blameworthiness). In Experiment 3, we ask questions about the objective and epistemic gradability of a feature, which is not objectively gradable (intentionality).

The intentionality of action is recognized as a nongradable property both in the traditional philosophical literature (e.g., Anscombe, 1957; Davidson, 1980; Ginet, 1990; Goldman, 1970; Mele, 1992; Mele & Moser, 1994; Wilson, 1989) as well as in experimental studies. It has usually been tested by means of a yes/no question (e.g., Hindriks et al., 2016; Knobe, 2003) or by means of an epistemic gradability measure (e.g., Sripada & Konrath, 2011; Tobia, 2014). We have run an experiment where we decided to ask the participants a degree of agreement question as well as a degree question about the intentionality of action. If the authors were right about the reliability of gradability data, people's responses to the degree of agreement question (DAQ) about the intentionality of the action could show a substantial distribution but the answers to the degree question (DQ) ought to be focused on two points (the attribution of intentionality and the attribution of unintentionality) and possibly a third one expressing the lack of attribution.

7.1. Methods

The participants were recruited by Clickworker (small compensation was provided) from native English speakers. Our sample consisted of 150 subjects (87 females, average age: 35).

7.2. Materials and procedures

We used the Knobe vignettes and asked two questions – the degree question (DQ) and the degree of agreement question (DAQ):

> (DQ) In your opinion, how intentional or unintentional is the chairman's action (of harming [helping] the environment)?
>
> (DAQ) To what extent do you agree that the chairman's action (harming [helping] the environment) is intentional or unintentional?

The answers were recorded on a seven-point pseudo-Likert scale similar to that employed by Hindriks and colleagues. For DQ, the possible answers were:

- Very intentional
- Intentional
- Somewhat intentional
- Neither intentional nor unintentional
- Somewhat unintentional
- Unintentional
- Very unintentional

For DAQ, the possible answers were:

- I strongly agree that the chairman's action is intentional.
- I agree that the chairman's action is intentional.
- I somewhat agree that the chairman's action is intentional.
- I neither agree that the chairman's action is intentional nor that it is unintentional.
- I somewhat agree that the chairman's action is unintentional.
- I agree that the chairman's action is unintentional.
- I strongly agree that the chairman's action is unintentional.

To check for order effects, we divided each experimental condition into two groups that differed in the order in which the DQ and DAQ questions were asked.

As in the previous experiment, we have coded the answers as numbers from 1 to 7 for the purposes of statistical analysis.

7.3. Results and discussion

The answers to the two questions asked are very similar across all conditions in which the moral valence of an outcome is the same. In the harm condition, participants tend to say that the chairman's action is intentional

(DQ: M = 2.13, SD = 1.28) and simultaneously they tend to agree with the claim that it is intentional (DAQ: M = 2.04, SD = 1.22). The same is true in the help condition. The participants tend to think that the chairman's action is somewhat unintentional (DQ: M = 5.2, SD = 1.94), and they tend to somewhat agree with the claim that the action is unintentional (DAQ: M = 5.05, SD = 2.08). The differences between the harm and help conditions are statistically significant (DQ: U = 2670, $p < 0.001$, Cohen's $d = 1.87$; DAQ: U = 3061, $p < 0.001$, Cohen's $d = 1.77$), but there is no difference in answers to the DQ and DAQ questions (HARM: U = 10,958, $p = 0.27$; HELP: U = 10,986, $p = 0.29$).

Our suspicion was that despite the fact that intentionality is not a gradable property, the results will not reveal it. Indeed, the answers to the degree question and to the degree of agreement question are highly correlated (r = 0.86, $p < 0.001$). The answers to the degree question are distributed along the scale just as are the answers to the degree of agreement question (see Figure 2).

It is also noteworthy that no order effects have been observed. After all, one could argue that people who answered the degree question first would interpret it charitably as a degree of agreement question, which would explain why there are no differences between the questions. However, one would expect

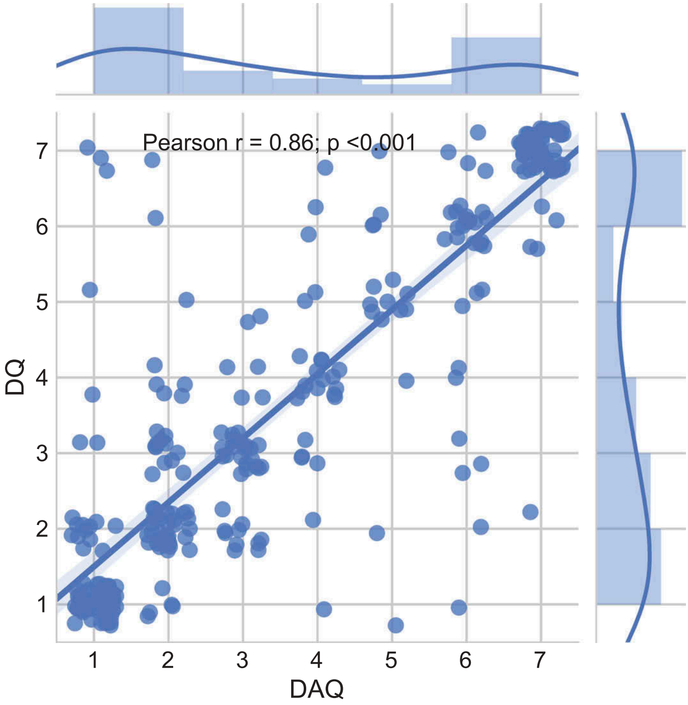

Figure 2. The correlation between answers to the DQ and DAQ questions about intentionality (Experiment 3). Small jitter was added to improve readability.

the answers given to the degree question by people who were confronted with the degree of agreement question first to differ. This was not the case.

The results thus support our skeptical contention that the mere use of a question that asks about objective gradability does not show that the answers gathered should be interpreted as concerning objective rather than epistemic gradability. We have shown that it is possible to obtain empirical data that might appear to reveal objective gradability even for a feature that is usually taken to be nongradable. The simplest way to account for such data is to think that subjects interpret the DQ question about intentionality as a DAQ question. Once again this casts doubt on whether we can be sure that the empirical data on which Hindriks and colleagues rely really do show the objectively gradable character of blame and praise attributions. Our experiments put the onus of proof on them.

8. Conclusion

In this paper, we have critically analyzed the New Angle on the Knobe effect put forward by Hindriks and colleagues (Hindriks et al., 2016). First, we have argued that there are good reasons to think that the two criteria of asymmetry and gradability do not single out only Hindriks' theory. In particular, Holton's Norm Violation Hypothesis meets not only the asymmetry criterion but also the gradability criterion.

We have further argued that the authors are not justified in drawing empirical support for the two criteria. As regards the asymmetry criterion, we did not replicate their results: we found a statistically significant correlation between the blame/praise and the intentionality ascription in both (harm and help) conditions in experiment 1 and 2.

As far as the gradability criterion is concerned, we have argued that the data available are insufficient to conclusively disambiguate between objective and epistemic gradability. We have shown that the authors need objective gradability to support their argument for the exclusion of rival accounts. Our experiments cast doubt on the natural thought that we can discriminate between objective and epistemic gradability by asking the degree question and the degree of agreement question, respectively. In Experiment 2, we have shown that the patterns of response to a question about epistemic and objective gradability are virtually the same. In Experiment 3, we have shown that the pattern of responses to questions about objective and epistemic gradability is the same even if the feature in question (intentionality in our case) is objectively nongradable. Our results thus show that we need to be careful in interpreting the responses to such questions at face value.

In view of the fact that Hindriks and colleagues need to appeal to objective (not epistemic) gradability to support the gradability criterion, and the fact that our studies show that standardly formulated questions do not appear to discriminate between these two types of gradability, we conclude that Hindriks and colleagues cannot claim to have shown that their data support the gradability criterion. It may, of course, be that such support could be found but our results put the burden of proof on them.

Our major aim was to undermine the conclusions proposed by Hindriks and colleagues. More generally, we suggested that one should not rely on such fine-grained distinctions as the one between objective and epistemic gradability too readily. Perhaps one can take our results to argue against experimental philosophy in general.[11] This would require much independent argument, however. We want to emphasize that it is possible to take our results constructively. In order to achieve reliable experimental results, we need to know the limits of the method. Our conclusions suggest that we cannot reliably discriminate between objective and epistemic gradability by means of asking degree questions and degree of agreement questions, respectively.

Notes

1. The hypotheses they evaluate include: Holton's Norm Violation Hypothesis, Knobe's Moral Valence Hypothesis, Blame Hypothesis, Sripada's (2010) Deep-Self Hypothesis, and the gradable versions of the last three. The only theory to fulfill both criteria is the gradable version of Knobe's Moral Valence Hypothesis. They reject it for another reason, namely the fact that it cannot account for cases where the agent's system of values is the reverse of that of participants (Hindriks et al., 2016, p. 212), as shown in the Nazi Germany study (Knobe, 2007).
2. We thank an anonymous reviewer for pressing this point.
3. "According to Hindriks (2008, 2011, 2014)) NRH, the indifference of the agent plays a central role in the explanation of the Knobe Effect. Due to his indifference, the agent fails to be motivated by an effect that he should care about. In other words, he ignores a normative reason"(p. 215).
4. "It seems plausible to say that, *ceteris paribus*, the less someone cares about a harmful side effect, the more she will be blamed. It also seems unobjectionable to say that, *ceteris paribus*, the worse the effect is, the more blameworthy the agent is ... Given these two claims, the idea that comes in sight is that the amount of blame people attribute depends on the extent to which they see a discrepancy between how much the agent should care and how much she actually cares" (p. 217).
5. For a similar point, see, for example, Brennan, Eriksson, Goodin, and Southwood (2013, p. 210): "Of course, not all norms are such as to admit degrees of violation. Most are, however. At the very least, the relation between competing norms will usually not be one of lexical ordering but rather something softer, allowing an agent to optimize his norm violations in such a way as to minimize their overall badness from his own perspective."
6. We would like to thank an anonymous reviewer for this objection.
7. Holton does not address the blame-praise asymmetry.

8. Moreover, inasmuch as duties are constituted by the norms, what has been said in defense of NVH applies also to the omissions account of the Knobe effect (Paprzycka, 2015, 2016).
9. The details of the experiment are discussed in Kuś and Maćkiewicz (2016).
10. Indeed, one may even raise doubt that two different things are measured. It is widely accepted in psychological research that a very high correlation coefficient indicates that not two but one construct is actually measured.
11. We would like to thank an anonymous reviewer for pressing us on this issue.

Acknowledgments

We are deeply indebted to Katarzyna Paprzycka-Hausman and Katarzyna Kuś for encouraging us to write this paper and their ceaseless support in the process of writing. We would also like to thank Adam Wierzbicki and Ignacy Szczeniowski for their valuable comments on the earlier drafts.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Narodowe Centrum Nauki [2013/11/B/HS1/03939].

Notes on contributors

Tomasz Zyglewicz is a graduate student of philosophy at the Graduate Center, City University of New York. His research interests include philosophy of language, metaphysics, and philosophy of law. He also holds a degree in law from the University of Warsaw.

Bartosz Maćkiewicz is a graduate student of philosophy at University of Warsaw. His work focuses on application of methods from corpus linguistics and psycholinguistics to research in experimental philosophy.

ORCID

Tomasz Zyglewicz http://orcid.org/0000-0002-5656-3561
Bartosz Maćkiewicz http://orcid.org/0000-0002-9460-5742

References

Anscombe, G. E. (1957). *Intention*. Oxford: Blackwell.

Brennan, G., Eriksson, L., Goodin, R. E., & Southwood, N. (2013). *Explaining norms*. Oxford: Oxford University Press.

Cova, F. (2017). Intentional action and the frame-of-mind argument: New experimental challenges to Hindriks. *Philosophical Explorations*, *20*(1), 35–53.

Davidson, D. (1980). *Essays on actions and events*. Oxford: Oxford University Press.

Ginet, C. (1990). *On action*. Cambridge: Cambridge University Press.

Goldman, A. I. (1970). *A theory of human action*. Princeton, NJ: Princeton University Press.

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

Hindriks, F. (2011). Control, intentional action, and moral responsibility. *Philosophical Psychology*, *24*(6), 787–801.

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

Hindriks, F. (2018). Explanatory unification in experimental philosophy: Let's keep it real. *Review of Philosophy and Psychology*, 1–24. https://doi.org/10.1007/s13164-018-0397-0

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

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

Kleinman, K., & Huang, S. S. (2016) Calculating power by bootstrap, with an application to cluster-randomized trials. *eGEMs*, *4*(1), 1-18.

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

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*, 90–106.

Knobe, J., & Mendlow, G. S. (2004). The good, the bad and the blameworthy: Understanding the role of evaluative reasoning in folk psychology. *Journal of Theoretical and Philosophical Psychology*, *24*(2), 252.

Kuś, K., & Maćkiewicz, B. (2016). Z rozmysłem, ale nie specjalnie. O językowej wrażliwości filozofii eksperymentalnej. *Filozofia Nauki*, *3*(95), 73–101.

Mele, A. R. (1992). *Springs of actions: Understanding intentional behavior*. New York: Oxford University Press.

Mele, A. R., & Moser, P. K. (1994). Intentional action. *Noûs*, *28*, 39–68.

Paprzycka, K. (2015). The omissions account of the Knobe effect and the asymmetry challenge. *Mind & Language*, *30*(5), 550–571.

Paprzycka, K. (2016). Intention, knowledge, and disregard for norms: The omissions account and Holton's account of the asymmetrical intentionality attributions. *Poznań Studies in the Philosophy of the Sciences and the Humanities*, *107*, 204–233.

Sripada, C. (2010). The deep self model and asymmetries in folk judgements about intentional action. *Philosophical Studies*, *151*, 159–176.

Sripada, C., & Konrath, S. (2011). Telling more than we can know about intentional action. *Mind and Language*, *26*, 353–380.

Tobia, K. P. (2014). Reflective intentions: Philosophical concepts of intentionality. *Proceedings of the Cognitive Science Society*, *36*(36). https://mindmodeling.org/cogsci2014/papers/790/paper790.pdf

Wilson, G. (1989). *The intentionality of human action*. Stanford, CA: Stanford University Press.

Wright, J. C., & Bengson, J. (2009). Asymmetries in judgments of responsibility and intentional action. *Mind and Language*, *24*, 24–50.