Explanatory Unification in Experimental Philosophy

Hindriks, Frank

Published in:
Review of Philosophy and Psychology

DOI:
10.1007/s13164-018-0397-0

Rev.Phil.Psych. (2019) 10:219–242
https://doi.org/10.1007/s13164-018-0397-0

Explanatory Unification in Experimental Philosophy: Let's Keep It Real

Frank Hindriks[1]

Published online: 21 April 2018

Abstract Experimental philosophers have discovered a large number of asymmetries in our intuitions about philosophically significant notions. Often those intuitions turned out to be sensitive to normative factors. Whereas optimists have insisted on a unified explanation of these findings, pessimists have argued that it is impossible to formulate a single factor explanation. I defend the intermediate position according to which unification is possible to some extent, but should be pursued within limits. The key issue that I address is how it is possible to set such limits in a way that is true to the phenomena.

In the past decade or so, experimental philosophers have made a large number of surprising findings about philosophically significant notions such as acting intentionally, deciding, being free, and causing. As it turns out, intuitions about most of these notions are influenced by normative factors (Knobe 2003, 2010a; Phillips et al. 2011). I refer to those for which this holds as 'the Normativity Findings'.[1] Many Normativity Findings pose a challenge to existing analyses of the relevant notions, as they were previously thought to be independent of normative factors. Joshua Knobe (2010b; Pettit and Knobe 2009) has argued that a unified explanation can and should be provided of all Normativity Findings. In this paper, I investigate whether Knobe's optimism about explanatory unification is warranted. I ask whether it is possible and desirable to identify a single factor that explains all the Normativity Finding at once.[2]

[1] As is briefly discussed in section 3.1, the fact that these findings reveal that attributions of notions such as acting intentionally are sensitive to normative factors does not as such entail that such factors play an essential role in these notions. They do not do so in the accounts that Adams and Steadman (2004a, b) and Machery (2008) defend.

[2] I focus on the constructive strand within experimental philosophy that uses empirical methods to contribute to conceptual analysis. The critical strand uses empirical methods to debunk the project of conceptual analysis (Weinberg et al. 2001; Machery et al. 2004). See Nadelhoffer (2007), Alexander (2012), and Knobe (2016) for similar classifications of experimental philosophy.

✉ Frank Hindriks
f.a.hindriks@rug.nl

[1] Department of Ethics, Social and Political Philosophy, University of Groningen, Groningen, The Netherlands

Explanatory unification plays an important role in the debate about the Normativity Findings. Phelan and Sarkissian (2009) are among the pessimists who doubt that a single-factor or silver bullet explanation is to be had.[3] They maintain that it 'is time to abandon the dream of parsimony' (Ibid.: 179). Mark Alfano, James Beebe, and Brian Robinson, in turn, flout this advice and propose an explanation by which they endeavor 'to keep the dream of parsimony alive' (2012: 284). So should we side with the pessimists or with the optimists? This question is important, because the extent to which a theory unifies plays an important role in the evaluation of rival theories. Knobe, for instance, dismisses a number of explanations by claiming that they leave us 'with a mystery as to why the impact of moral judgment is so pervasive' (2010b: 655).

I maintain that unification is desirable to the extent that a number of empirical findings constitute one phenomenon, and that unification is possible to the extent that one factor explains that phenomenon. This may sound puzzling for the following reason. Whether different findings constitute one phenomenon ultimately depends on whether there is in fact one factor that explains it. So it seems that explanans and explanandum are interdependent in a way that makes it impossible to delineate the one without the other. This poses what I call 'the Identification Problem'. A solution requires a well-argued characterization of the empirical findings and thereby of the explananda. I argue that, independently of any particular hypothesis about the underlying causes, we often have reason to adopt a certain view about which phenomena should be grouped together as requiring a unified explanation. Fitting characterizations of the explananda can in turn bear on what is the correct explanans.

In section 1 I discuss how the Identification Problem can be solved. Section 2 introduces the Normativity Findings. Sections 3 and 4 illustrate how the Identification Problem can be solved for these findings by first carefully characterizing the explanandum, and then assessing how this bears on what is the best explanans. This endeavor serves to adjudicate the controversy between optimists and pessimists. I defend an intermediate position that is realistic and – in a sense to be explained – realist.

1 Explanatory Unification and the Identification Problem

1.1 Two Theories of Explanatory Unification

An explanation unifies by forging connections between phenomena that were previously thought to be unrelated. The more phenomena it explains, the more the explanation unifies. Philip Kitcher (1989) regards explanations as arguments. On his view, to unify different phenomena is to account for them in terms of one and the same argumentative pattern, i.e. to apply the same argumentative structure to distinct sets of premises. Kitcher regards explanatory unification as valuable because it economizes on cognitive resources. In Jim Woodward's terms, Kitcher takes the goal of explanatory unification to be to derive 'as much from as few patterns of inference as possible' (2011: 52). As this is a pragmatic goal, I refer to Kitcher's account of explanatory unification as 'the pragmatic conception'.

[3] I take the term 'silver bullet' explanation from Alfano et al. (2012: 283).

Uskali Mäki (2001: 504) argues that unification should instead be seen as a matter of factual discovery. On such an ontological conception, unification is 'a matter of turning apparent diversities into real unities' (ibid.: 502).[4] Unification occurs when two or more phenomena are successfully explained in terms of one theory or even one causal factor (Morrison 1990; Mäki 2001). The degree to which a theory unifies increases with its explanatory scope, or the number of phenomena that it explains, as well as when the number of explanatory factors decrease (Mäki 2001: 493). The extent to which unification is appropriate, according to the ontological conception, depends 'on whatever unity there is in the world' (ibid.: 502 and 503).

The pragmatic conception promotes maximal cognitive efficiency (Mäki 2001: 493–94). Woodward (2011) points out that the account entails that only the most unifying account is explanatory, an implication he regards as implausible. Even on a weaker interpretation of Kitcher – the more an argument unifies the higher the extent to which it explains – it is problematic: due to its one-sided focus on cognitive efficiency, the pragmatic conception licenses inflated claims of unification. It may be that, keeping empirical adequacy fixed, two theories of different findings is to be preferred to one theory that covers all of them. The ontological conception sets clear limits to the extent to which an explanation should unify: an explanation should include only those phenomena that are indeed caused by the same explanatory factor. It may well be, however, that pragmatic considerations such as the ones Kitcher invokes would count in favor for an explanation that includes more phenomena – too many according to the ontological conception. This is important because, as indicated in the introduction, the optimists – in particular Knobe (2010b) and Alfano et al. (2012) – strive for maximal unification. The case study below reveals that this is not always a good idea.

So how exactly does an explanation of a Normativity Finding work on the ontological conception? In the experiments under consideration, experimental philosophers conduct surveys in which participants read a vignette and answer questions about the scenario they have read. Different versions of the scenario are used in order to determine whether they elicit different answers. The answers that participants in these experiments give or the attributions they make are taken to reveal something about the concepts that people employ or about the way they apply them (the judgments they form are often called 'intuitions'[5]). This is why it can be of philosophical interest to explain an asymmetry in intuitions about different versions of a scenario. The answers are elicited by questions, which are the triggering causes of the attributions or intuitions; the concepts that generate these answers can be seen as the structuring causes (Dretske 1988). In this way, an explanation of a Normativity Finding can be seen as a causal explanation. On the ontological conception, the extent to which there are genuine causal relations between concepts and attributions in particular contexts sets limits to the degree to which an explanation should unify.

[4] There are of course many other accounts of explanatory unification. A seminal contribution is Friedman (1974). More recent accounts include Bartelborth (2002) and Myrvold (2003). The basic point I want to make, however, is how explanatory unification can connect to ontological considerations. I believe that this is best made in terms of Mäki (2001), who puts them center stage, in contrast to Kitcher (1989), who ignores them. The optimists seem to side with Kitcher's pragmatic conception, whereas I defend the ontological conception.

[5] I follow the common practice of referring to such folk judgments as 'intuitions', leaving open how this term is best interpreted. See Bealer (1998), Lewis (1983) & Williamson (2007), Kauppinen (2007), Ludwig (2007), and Sosa (1998) for five different views of the nature of intuitions.

1.2 The Identification Problem

The ontological conception faces a problem. Consider Knobe's description of the famous Knobe Effect (which is described in more detail in section 2): 'people's moral judgments are somehow impacting their intuitions about intentional action' (2010a: 315–329) Subsequently, he uses judgments about whether the outcome is morally good or bad to explain the asymmetry he finds in people's judgments about intentional action. So the way he describes the empirical finding features both the explanandum and the explanans. This is not without problems. The reason for this is that, as is discussed extensively in section 3, empirical findings can often be described in different ways. For instance, the agent in the vignettes Knobe uses (which is also described in section 2.1) is indifferent with respect to a side effect of his action. For all we know, this should play a role in a proper explanation of the attribution asymmetry. But given Knobe's description of the empirical finding, there is no reason to think that it should.

This makes the description of an empirical finding in cases such as the ones at hand a delicate issue. If that description features both the explanandum and the explanans, it is difficult to evaluate the quality of the explanation. It may even seem that there is no way to identify the explanandum that is independent of the identification of the explanans. After all, a proper description of the phenomenon to be explained requires knowledge of both. In the cases at hand, it seems that someone who does not know what the explanans is cannot properly describe the empirical finding. I call this 'the Identification Problem'.

It also poses a problem for imposing limits on the extent to which an adequate explanation unifies. The argument for doing so is that some of the phenomena are not successfully explained by the factor she invokes for explaining the others. However, if the empirical finding is described as having a broad scope, such an objection might seem to miss its target. After all, the explanandum that is described in terms that entail a wide scope can (presumably) be derived from the explanans. This suggests that ontic unity is useless as an adequacy condition for explanatory unification, and that it is perfectly all right to strive for maximal cognitive efficiency and thereby for a maximally unified explanation.

The Identification Problem can be solved only if there are means other than knowledge of the explanans by which the explanandum can be identified. Scientific practice suggests that there are indeed more provisional means for delineating the explanandum. Scientists use indicators to arrive at a provisional characterization of the explanandum. Such a provisional delineation of the explanandum can be used to arrive at a reasonable estimate of the extent to which unification is desirable. Furthermore, as is illustrated below, scientists bring background knowledge to bear on the topic they investigate. This background knowledge consists of a combination of previous empirical findings and theoretical commitments. These factors explain why there are many situations in which the Identification Problem does not arise.

An empirical finding is a novel fact when it does not fit sufficiently well with existing theoretical commitments and prevailing theories do not predict it. Novel facts warrant reformulating the theory in such a way that it explains the re-described explanandum (Hitchcock and Sober 2004). Scientists can also check, independently of their theoretical commitments, whether a recently discovered phenomenon is similar to existing findings. The extent to which two findings are similar depends on whether

the experimental setups by which they are uncovered resemble each other sufficiently, and whether the findings share important features. When they are sufficiently similar, this is reason to regard both as part of one and the same explanandum. Both novelty and similarity can contribute to a proper description of an empirical finding. They set in motion a process of mutually adjusting explanandum and explanans. Thus, similarity and novelty form the engine of theoretical innovation. Ideally, this process results in a degree of unification that is true to reality.

How exactly this works will become clear when I illustrate what role novelty and similarity can play with respect to the Normativity Findings in sections 3 and 4. They reveal how the process of mutual adjustment works, and how it can be justified. I return to the Identification Problem in section 4.3 to further clarify how it can be solved. The basic point I have been making here is that the description of an empirical finding can be a delicate matter, as it can feature both the alleged explanandum and the alleged explanans. Before I can apply similarity and novelty as criteria for identifying the explanandum and the explanans, I introduce in section 2 the Normativity Findings that are most important to my argument, those concerning intentional action (in section 3.2 I introduce the other Normativity Findings that bear on the conclusions I draw, those concerning advocating, favoring, believing, and knowing). The key point of the paper is a methodological one: the way in which the explanandum is described is crucial for assessing the plausibility of an explanation. This insight is used to adjudicate the controversy between optimists and pessimists concerning the extent to which a unified explanation of these findings is called for.

2 Empirical Findings

The Knobe Effect Consider ENVIRONMENT, the vignette that Knobe used in his first studies concerning intentional action. The help [harm] version of this scenario reads[6]:

> The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, and [but] it will also help [harm] the environment.' The chairman of the board answered, 'I don't care at all about helping [harming] the environment. I just want to make as much profit as I can. Let's start the new program.' They started the new program. Sure enough, the environment was helped [harmed]. (Knobe 2003: 191).

Knobe asked the participants in his experiment whether the chairman intentionally helped [harmed] the environment. He found that only a minority (23%) say that the chairman helps the environment intentionally, whereas most people (82%) say that he harms it intentionally.[7] This asymmetry has become known as 'the Knobe Effect' (Machery 2008). As Knobe (2010a) discusses, it has been replicated using other vignettes featuring a lieutenant, or a terrorist, as well as with different subjects,

[6] For space considerations, ENVIRONMENT is one of two vignettes that I present in full. The references provided below allow the reader to find the other vignettes.

[7] All asymmetries presented in this paper are statistically significant at the 5% level or higher.

including children as young as four years old, people who speak Hindi, and by people with an affective deficit. The Knobe Effect has generated a lot of research because it took many people by surprise. Apparently, intuitions concerning intentionality are sensitive to a factor of which many philosophers were hitherto unaware.

The Knobe Effect is often described in terms of the claim that the moral valence of a side effect (good/bad) influences the attribution of intentional action (no/yes). In section 3 I argue that this is far from a neutral description of the Knobe Effect. This is important because it bears directly on what is seen as a successful explanation of the Knobe Effect (section 4).

Knobe also asked how much praise the agent deserved for what he did in the help condition, and how much blame in the harm condition. People turned out to attribute little praise, and a lot of blame (a Likert scale was used ranging from 0 to 6; $M = 1.4$ in the help condition, $M = 4.8$ in the harm condition). As is discussed below, this asymmetry concerning attributions of moral responsibility plays an important role in the explanation of the Knobe Effect.

IA-Aesthetics As it turns out, asymmetric intentionality attributions such as those involved in the Knobe Effect are not unique to the moral domain. Knobe (2004a) discovers a similar asymmetry pertaining to aesthetics. In the MOVIE vignette he uses, the CEO of a movie studio decides to implement a strategy that increases profits while expressing indifference with respect to the fact that it will make the movie artistically better [worse]. 18% of the participants attribute intentionality to the CEO in the better condition, and 54% in the worse condition.

IA-Prudence Knobe and Mendlow (2004) argue that prudential side effects also give rise to asymmetric intentionality attributions. Their SALES vignette features a president of a computer corporation who decides to implement a program that increases sales in Massachusetts, but decreases sales in New Jersey. As it turns out, people say that the president decreases sales in New Jersey intentionally (75%). They do not, however, blame her for doing so (90%). Interestingly, Knobe and Mendlow (2004) fail to establish an asymmetry, as they do not consider cases in which the effect is prudentially good. It is important to discuss this finding because it has been taken to challenge the claim that what unites the findings under discussion is the role of normativity (Machery 2008: 182).

IA-Perspective Intentionality attributions have turned out to be sensitive to the evaluative perspective of the agent that features in the vignette rather than to that of the participant in the experiment. As this finding plays a pivotal role in the argument below, I present the NAZI vignette in full:

> In Nazi Germany, there was a law called the "racial identification law." The purpose of the law was to help identify people of certain races so that they could be rounded up and sent to concentration camps. Shortly after this law was passed, the CEO of a small corporation decided to make certain organizational changes.
>
> The vice-president of the corporation said: "By making those changes, you'll definitely be increasing our profits. But you'll also be fulfilling [violating] the

> requirements of the racial identification law." The CEO said: "Look, I know that I'll be violating the requirements of the law, but I don't care one bit about that. All I care about is making as much profit as I can. Let's make those organizational changes!"
>
> As soon as the CEO gave this order, the corporation began making the organizational changes. (Knobe 2007: 105)

Fulfilling [violating] the requirements of a racial identification law is a side effect of making the organizational changes. Intentionality with respect to this side effect is attributed in the violate condition (81%) and not, or at least substantially less so, in the fulfill condition (30%). People attribute some blame to the CEO in both conditions ($M = -0.9$ in the violate condition and $M = -1.7$ in the fulfill condition; the difference between the two conditions is *not* significant).[8]

Presumably the racial identification law was meant to facilitate deporting and killing Jews. As this was an atrocity, contributing to it by fulfilling its requirements is questionable, to say the least. Knobe concludes that 'what we have here is a case in which subjects consciously believe that violating the requirements is actually a good thing and nonetheless end up concluding that the agent acted intentionally' (Ibid.: 102). Thus, when described in terms of moral valence, the finding is the reverse from the Knobe Effect. The CEO may well have had a different perspective on the case. It is typically regarded as bad to violate a norm. The CEO might think that, in principle, he has a duty to abide by the laws of the country he lives in. This difference in perspective seems relevant to why the way in which intentionality is attributed differs from other cases such as the Knobe Effect.

I refer to these four findings as 'Intentional Action Asymmetries' (IAAs). In sum:

1. The Knobe Effect: it concerns a difference in moral valence
2. IA-Aesthetics: it introduces a difference in aesthetic valence
3. IA-Prudence: it pertains to a prudential side effect
4. IA-Perspective: it concerns a switch of perspective; the moral valence of the agent differs form that of the attributor.[9]

As mentioned earlier, the Normativity Findings extend beyond the notion of intentional action. I will introduce asymmetries concerning notions such as advocating, desiring, believing, and knowing when they become relevant below in section 3.2.

[8] In the fulfil condition, the CEO is blameworthy because he is indifferent (attitude) with respect to a harmful effect (outcome). In the violate condition, there is no harmful effect. The only reason for blaming the CEO is that he flouts a reason he thinks he has (attitude).

[9] This list of IAAs is far from complete. Perhaps the most conspicuous absence concerns findings suggesting that there are a number of different conceptions of intentional action. According to this semantic diversity approach, only a minority employs an asymmetric notion of intentional action (Nichols and Ulatowski 2007; Cushman and Mele 2008; Cokely and Feltz 2009, and Pinillos et al. 2011). I do not consider this issue here, because it does not affect my argument. All that matters for my argument is that there is at least a substantial minority that employs the asymmetric notion.

3 Solving the Identification Problem: the Explanandum

3.1 Similarities among IA-Asymmetries

Given this enumeration of the core Normativity Findings, the question is how the explanandum should be described. Should it accommodate all of them or only some? In this section, I illustrate how similarity can be used as a criterion for systematizing the data, in particular for delineating the explanandum. As discussed in section 1.2, whether two findings are similar depends on whether the experimental setups by which they are uncovered resemble each other sufficiently, and whether the findings share important features. This criterion can be used to arrive at a proper description of the empirical finding and thereby of the explanandum. This is the first part of the solution to the Identification Problem. As similarity will serve to exclude some of the Normativity Findings as not being part of the explanandum, it is also conducive to achieving the appropriate degree of explanatory unification.

As the first Normativity Finding, the Knobe Effect, has been regarded as a novel fact, I should also comment on novelty here. Whether a finding is novel depends on existing theoretical commitments. The point to appreciate is that, when the Knobe Effect was discovered, virtually all theories of intentional action ruled out that the normative significance of an effect bears on intentionality. This explains why many commentators expressed their surprise or puzzlement.[10] Hence, it is natural to regard the Knobe Effect as a novel fact.

As the terms 'harm' and 'help' feature in the vignettes used for establishing it, the Knobe Effect – can be described as follows (I abstract from the fact that the finding concerns an attribution *tendency*), with 'DEF' for 'description of empirical finding':

> [DEF1] People attribute intentionality when a side effect is harmful, and not when it is helpful.

Strikingly, Knobe describes the finding in terms of moral valence of the side effect, i.e. whether it is good or bad, rather than in terms of harmful or helpful. And he proposed an explanation that is intimately related to this description (Knobe 2003, 2004b, 2006):

> *BAD1.* The moral valence of a side effect (good/bad) explains the intentionality attributions (no/yes).

This is what I call 'the badness model'. As Knobe refined it later, I refer to this account as *BAD1.* It is what has been called 'a competence account', because it takes the Knobe Effect to reflect the competencies that people have regarding a concept (Nado 2008). The thing to note is that both the explanans and the explanandum are contained in Knobe's description of the empirical finding. This makes it abundantly clear that the description of the explanandum matters.

[10] See Adams and Steadman (2004a), Guglielmo et al. (2009), Knobe (2003, 2010a), McCann (2005), Nadelhoffer (2004a), Pellizoni et al. (2010), and Wright and Bengson (2009). Harman (1976) is one of a few who considered intentionality attributions related to merely foreseen harm and regarded them as justified and, apparently, as unsurprising.

Many philosophers and psychologists assume that the goals of folk psychology do not include moral evaluation, but are restricted to explanation and prediction.[11] Because of this, they object to explaining intentionality attributions in terms of moral valence. Some have proposed to explain them instead in terms of moral responsibility attributions. As mentioned in section 2.1, people do not praise the chairman in the help condition of ENVIRONMENT, but they do blame the chairman in the harm condition. According to what I call 'the blame model' (*BLAME*), this explains the Knobe Effect:

> *BLAME*. The moral responsibility asymmetry (praise/blame) explains the intentionality attributions (no/yes).

It is commonly assumed that intentionality attributions provide input for responsibility attributions. *BLAME* reverses the order of explanation. Proponents of *BLAME* assume that people are mistaken when they let their responsibility judgments affect their intentionality judgments. Because of this, *BLAME* is what I call 'a bias account': it explains an empirical finding in terms of a bias that affects how people apply the notion at issue.[12] Note that both *BAD1* and *BLAME* are what I call 'attributor explanations' in that they explain the Knobe Effect in terms of moral judgments of the attributor rather than the agent.

Despite appearances, [DEF1] is far from a neutral description of the Knobe Effect. It ignores the fact that the protagonist in ENVIRONMENT is indifferent. Theories of intentionality typically assume that it requires either a pro or a con attitude. Given this theoretical commitment, it is surprising that intentionality is ascribed to someone who is indifferent with respect to the outcome at issue. This is important because it bears on the question of how similar other findings are to the Knobe Effect, which in turn bears on the plausibility of rival explanations. In light of this, I propose to move beyond [DEF1] and re-describe the empirical finding as follows[13]:

> [DEF2] People attribute intentionality to an agent who expresses indifference with respect to a side effect of his intended action when it is harmful, and not when it is helpful.

The agent's indifference does not play a role in *BAD1* or *BLAME*. Hence, these explanations fail to capture an important and surprising aspect of the phenomenon.

IA-Aesthetics is identical to the Knobe Effect in all respects except for the fact that the side effect has aesthetic rather than moral significance. As it is rather similar to the Knobe Effect, it warrants a re-description of the phenomenon:

> [DEF3] People attribute intentionality to an indifferent agent when the valence of the side effect is negative, and not when it is positive.

[11] Duff (1990) and Gallagher (2005) maintain that evaluation is also a goal of folk psychology.

[12] Nado (2008) contrasts competence accounts to 'performance accounts', because on such accounts the attributions are performance errors. I take the term 'bias account' to be more informative, as this term makes it immediately apparent that the responses are misattributions. Alicke (2008), Malle and Nelson (2003), McCann (2005), Mele (2001), Nadelhoffer (2004a, 2004b, 2005, 2006), Nado (2008) have proposed bias accounts.

[13] See Phelan and Sarkissian (2009) for a number of studies concerning the agent's attitude. In some the agent is indifferent, in others he cares about the side effect at issue.

The upshot is that the valence of an effect need not solely be moral.

IA-Prudence differs from the Knobe Effect and from IA-Aesthetics in that it concerns a prudential effect rather than a normative effect. As a consequence, people do not attribute blame to the agent who brings about the prudentially bad effect. Knobe and Mendlow (2004) formulated SALES as a test of *BLAME*. They argue that it refutes *BLAME*, as intentionality is attributed in spite of the fact that no blame is attributed. Hence, blame cannot be the cause of the surprising intentionality attributions involved in the Knobe Effect. If they are right, the empirical finding – and thereby the explanandum – should be re-described once more, perhaps without any reference to normative factors.

There are a number of problems with this conclusion. Strikingly, Knobe and Mendlow (2004) test only the version of SALES in which the effect is bad, not that in which the effect is good. This leaves us without experimental evidence for an asymmetry, which means that the similarity of IA-Prudence and the other IAAs mentioned has not been established. Furthermore, the protagonist in SALES is not indifferent with respect to the side effect of her action. She carefully weighs the pros and cons and decides to implement the program only because the gains in Massachusetts will be greater than those in New Jersey. This means that also this experiment does not provide support for the similarity claim.

Consider how people might respond in the untested good condition. Given that the agent is not indifferent, he might have a pro attitude with respect to the effect, which would mean that it is an intended effect that is brought about intentionally. This would mean that there is no asymmetry. Also on the basis of a related experiment that Machery (2008) conducted, Jennifer Wright and John Bengson conclude that 'there is no Knobe Effect here to be explained (2009, 41). As there is substantial reason to doubt the similarity between IA-Prudence and other IAAs, IA-Prudence should not be included in the description of the empirical finding.

IA-Perspective concerns the question whether attributions depend on the perspective of the agent or the attributor. The way people respond to the NAZI-vignette suggests that they prioritize the perspective of the agent. The difference in perspectives can be captured as follows:

(Participants) Fulfilling [violating] the legal requirements is bad [good].

(CEO) Fulfilling [violating] the legal requirements is good [bad].

Participants attribute intentionality when the outcome is bad as conceived of from the agent's perspective.

Philosophers of action take the notion of intentional action to concern the agent's perspective or frame of mind (Bratman 1987).[14] IA-Perspective confirms this. Hence, the extent to which it can be regarded, as a novel fact is limited. It is, however, worthwhile to make this theoretical commitment explicit and re-describe the explanandum as follows:

[DEF4] People attribute intentionality to an indifferent agent when the valence of the side effect as the agent perceives it is negative, and not when it is positive.

[14] The priority of the attributor's perspective is also apparent from findings in different epistemic conditions (Pellizoni et al. 2010).

IA-Perspective is mainly significant because it conflicts with theories that have been proposed to explain other findings. *BAD1* and *BLAME* cannot explain it (see Table 2 below for an overview of which explanation explains which findings).

The upshot is that description matters. More specifically, the way in which new findings are described affects the scope of the explanandum. I have argued that the explanandum should encompass all the findings discussed so far except for IA-Prudence. As it turns out, it is possible to draw sensible conclusions about the description of the explanandum without depending heavily on particular explanations. All this implies that the process of delineating the explanandum must be done with great care.

3.2 Beyond the IA-Asymmetries

The Normativity Findings extend beyond the IA-Asymmetries to other folk psychological notions.[15] Some findings concern epistemic notions such as believing and knowing. These are 'Epistemic Asymmetries' (Beebe and Buckwalter 2010; Alfano et al. 2012; Beebe and Jensen 2012, and Beebe 2013). They are structurally similar to the IAAs in that by and large people take a notion such as knowledge to apply in the harm condition, but not in the help condition.[16] The Epistemic Asymmetries contradict the widely accepted theoretical commitment that notions such as belief and knowledge should be sensitive to cognitive factors only. Thus, these asymmetries constitute novel facts. Because of this, [DEF4] should be revised by adding these epistemic notions:

> [DEF5] People attribute intentionality, belief and knowledge to an indifferent agent when the normative valence of the side effect as the agent perceives it is negative, and not when it is positive.

A large number of asymmetries pertain to notions that, just as intentional action, concern motivation – including for instance advocating, deciding, desiring, favoring, intending, opposing, and wanting (Knobe, Cushman, and Sinnott-Armstrong 2008, Pettit and Knobe 2009, Guglielmo and Malle 2010). In light of this, I refer to them as 'Conative Asymmetries' (I use this label for conative asymmetries other than those concerning intentional action). What is striking about these findings is that the asymmetry between the help and harm conditions is less pronounced as compared to IAAs and Epistemic Asymmetries. Rather than using a forced-choice format, participants rated the applicability of the conative notions in ENVIRONMENT on a Likert scale, usually ranging from 1 ('disagree') to 7 ('agree'). Consider

[15] A number of notions that have little or nothing to do with the theory of mind, have also turned out to be sensitive to normative considerations. These findings include causing, being free, happiness, innate, love, natural, and valuing (Knobe and Fraser 2008; Phillips and Knobe 2009; Knobe 2010b). They differ from the other Normativity Findings in that they are sensitive to the attributor's perspective rather than that of the agent. This provides substantial reason to doubt that they warrant a re-description of the explanandum. Furthermore, as I have argued elsewhere, these asymmetries are not new Hindriks 2014.

[16] Using a seven-point scale ranging from −3 (strongly disagree) to 3 (strongly agree), Beebe and Buckwalter 2010 find $M = 0.91$ in the help condition and $M = 2.25$ in the harm condition; using a scale from 1 to 7, Beebe and Jensen 2012 find $M = 6.37$ in the harm condition and $M = 3.35$ in the help condition. Using a − 3 to 3 scale, Beebe (2013) finds $M = 0.27$ in the help condition and $M = 1.27$ in the harm condition. Finally, using a forced-choice format, he finds that less than half of the subjects attribute belief: 40% in the harm condition, 19% in the help condition.

favoring. The respective means for the help and harm conditions are $M = 2.6$ and $M = 3.8$ (Pettit and Knobe 2009: 593). In this case, both means are in fact below the midpoint of the scale, which means that "the average participant" refrains from attributing the notion in either condition. People remain neutral about whether the notion applies in the harm condition. Rather than no/yes asymmetries, these findings are plausibly seen as no/neutral asymmetries.[17]

Should the explanandum be re-described so as to encompass the Conative and Epistemic Asymmetries? Pettit and Knobe (2009) defend an affirmative answer arguing that, as the differences between the two means are statistically significant, they call for an explanation. It is not obvious, however, that these no/neutral asymmetries should be seen as novel facts. In the vignette that Pettit and Knobe used for favoring and advocating the assistant manager of a popular coffee franchise discusses a new procedure for preparing and serving coffee with the employees. The vignette continues as follows:

> The assistant manager spoke forcefully in favor of adopting the new procedure, saying: I know that this new procedure will mean less [more] work for the employees, which will make them very happy [unhappy]. But that is not what we should be concerned about. The new procedure will increase profits, and that should be our goal. (Ibid., 592)

In this WORK vignette, the assistant manager claims that, irrespective of the effect it has on the employees, the new procedure should be evaluated in terms of profits only. In contrast to ENVIRONMENT, the protagonist of this story does not claim that she does not care about the side effect. All she claims is that the procedure *should* be supported irrespective of whether one likes or dislikes the side effect.

But way make a point of the fact that a decrease in workload is irrelevant? Why doesn't she present this as a welcome bonus? A natural interpretation of this is that she does not in fact care about the happiness of the employees. This means that she neither favors nor advocates the side effect (no). In contrast, it makes sense to explicitly state that for practical purposes an increase in workload is irrelevant. She might regret this side effect. If she does, the assistant manager defends the procedure in spite of it. This suggests that we do not have enough information to determine what her private attitude towards the effect is. Perhaps she dislikes it. But she might even like it. As we cannot be sure, it is natural to remain neutral about this. And this is exactly what we see (neutral). This suggests that these Conative Asymmetries are not novel, and do not need a special explanation (Hindriks 2014). The upshot is that they do not call for a re-description of the explanandum.

The formulation of an explanandum is only the first step in solving the Identification Problem, that of using similarity for the purpose of properly describing the empirical finding and thereby delineating the explanandum. In order to assuage worries about circularity, I need to discuss the second step, that of using novelty for the purpose of circumscribing the explanans. In the preceding discussion I have already mentioned novelty – the fact that a finding is not predicted by existing theories – a few times.

[17] Knobe confirmed in personal communication that in the cases of advocating and favoring the difference between rating in the harm condition and the midpoint of the scale is not statistically significant.

These discussions already suggest that similarity provides a basis for re-describing the explanandum that is sufficiently independent of the role that novelty plays to allay worries about circularity. In order to argue for this more fully, however, I go on to discuss the process of re-description of the explanans in some detail, and pit rival theories against each other.

4 Solving the Identification Problem: the Explanans

4.1 Valence and Attitudes: Beyond *BAD1*

Both *BAD1* and *BLAME* offer what I call 'attributor explanations' of the Normativity Findings: they explain these findings in terms of the attributor's moral judgments. Knobe has introduced a new version of *BAD1* that also invokes the attitudes of the agent. According to what I call '*BAD2*', people attribute notions such as intentionality by comparing the attitude an agent is taken to have towards the effect to a standard attitude. The standard changes when moral judgments are involved (Pettit and Knobe 2009: 596). As a more positive attitude is required when the effect is beneficial, the standard moves upwards in such cases. When the effect is harmful, a less positive attitude suffices. The upshot is that the attitudes agents have toward 'different outcomes end up getting compared to different defaults' (ibid.: 596):

> *BAD2.* The moral valence of a side effect (good/bad) explains the default (higher/lower) with respect to which the agent's attitudes are to be compared. Intentionality (no/yes) is attributed only if the default is met (no/yes).

This account retains the core insight of *BAD1*: that the moral valence of a side effect (good/bad) explains the intentionality attributions (no/yes). The difference is that the agent's attitudes also play a role in it.[18]

Explanatory Unification Knobe believes that an explanation of the Normativity Findings should also account for the Conative Asymmetries. As these are no/neutral asymmetries, *BAD1* does not explain them. Knobe's motivation for introducing *BAD2* is that, because of the moving default, it does explain these findings.[19] I have argued that the Conative Asymmetries do not require a special explanation. If this is correct, the scope of *BAD2* is too wide.

In other respects, however, the scope of *BAD2* is too narrow. In spite of the fact that it invokes the agent's attitudes, *BAD2* cannot explain IA-Perspective. The reason for this is that it preserves the core of *BAD1*: it explains the asymmetries in terms of the judgments that the attributor forms about the moral valence of the side effect. The attributor's judgments determine the standard, and thereby whether a side effect is

[18] Hindriks, Douven and Singmann 2016 provides empirical evidence that counts against *BAD1*, *BAD2*, and *BLAME*. These theories predict that responsibility attributions and intentionality attributions correlate with one another. As it turns out, this holds for blame, but not for praise.

[19] Pettit and Knobe suggest that the scope of their explanation extends beyond these asymmetries: 'the data presented here suggests that that effect actually arises both in cases that do not involve pro-attitudes and in cases that do not involve side effects' (Ibid.: 602).

regarded as good or bad. However, in IA-Perspective the attributor's take on the effect differs from that of the agent. Hence, *BAD2* makes the wrong predictions. Belief and knowledge do not seem to come with expectations about whether the agent has some positive attitude towards the epistemic object. Hence, *BAD2* does not apply to the Epistemic Asymmetries.

Specifically, the attributor regards fulfilling the requirements of the racial identification law as bad. As a consequence, the standard moves downward, which means that a weaker attitude such as indifference meets the standard. Thus, *BAD2* predicts the attribution of intentionality in these cases, which is not what we see. Similarly, an agent who violates the law is regarded as doing something good by the attributor. This means that an attitude that is more positive than indifference is required for the attribution of intentionality. In fact, however, participants do attribute intentionality. The conclusion I arrive at is striking in the light of Knobe's optimism about explanatory unification. *BAD2* is too narrow in some respect and too broad in others. In other words, Knobe both promises more and delivers less than he should.

Theoretical Integration Where have things gone wrong? Knobe aims at a maximally unified explanation. If explanation were merely a matter of cognitive efficiency, the account that has the widest scope would by definition be the most attractive theory. When unification is constrained by ontology, however, other explanatory virtues become relevant. This is where existing theoretical commitments come in. Pettit and Knobe do not regard them as relevant, as is implied by the fact that they do not perceive it as a problem that their hypothesis 'does not make any claims about the other aspects of the relevant concepts' (2009: 594). Knobe (2016) goes so far as to claim that experimental philosophers should not use findings obtained by conceptual analysis at all. I believe that this is a mistake.

By integrating new findings within existing theories scientists can tighten the theoretical connections within a theory, and thereby increase the degree of theoretical integration. When a new explanation is embedded in an existing theory, mutatis mutandis, the explanatory power of the theory increases. This in turn means that it can be used for answering a wider range of questions about the phenomenon under investigation (Ylikoski and Kuorikoski 2010). Knobe's eagerness to increase the scope of his explanation comes at the expense of other explanatory virtues such as this one.

So which theoretical commitments might be relevant? Well-established theories of intentional action share the following three ideas (Anscombe 1957; Harman 1976; Bratman 1987, and Enç 2003):

(i) Intentional action is a frame of mind notion, and should as such be understood from the perspective of the agent.
(ii) Intentional action should be explicated in terms of the attitudes of the agent.
(iii) An agent brings about an effect intentionally only if it has a pro or a con attitude towards that effect.

The Normativity Findings present a challenge to (iii), but not to (i) and (ii). Note that these two commitments may well generalize to other notions such as favoring or advocating. As attributor judgments about moral valence still play a role in *BAD2*, it violates (i) and (ii).

Preserving existing theoretical commitments is not only important for enhancing the explanatory power of a theory. It is also crucial for appreciating the significance of empirical findings in general and, more specifically, for establishing whether they are novel. Consider the following two examples. IA-Perspective was surprising only because attributor explanations had abandoned the idea that intentional action is a frame of mind notion and, as such, concerns the agent's perspective (i). The attributions made in SALES can be explained in terms of con attitudes (ii and iii). This implies that IA-Prudence is not a novel fact. The upshot is that commitments (i) and (ii) have been abandoned without good reason.

4.2 Normative Factors: the *NORM* Model

The third perspective on the Normativity Findings is the normativity model *NORM* (see Table 1 for an overview of the explanations discussed in this paper).[20] The normativity model explains asymmetries in terms of the agent's response to a normative factor: a norm or a normative reason. As such it provides an agent explanation rather than an attributor explanation. In order to assess how *NORM* accounts compare to *BAD2*, I will discuss their scope (explanatory unification) as well as the extent to which they do justice to existing theoretical commitments (theoretical integration). The point is to illustrate how the quality of a theory can be appreciated when explanatory unification is not the only virtue that counts.

According to the *NORM* account that Richard Holton (2010) has proposed, the intentionality attributions are to be explained in terms of how the agent responds to a norm:

> *NORM*-violation. The agent's response to a norm (conform/violate) explains the intentionality attributions (no/yes).

As I discuss shortly, Holton takes this explanation to generalize. The core idea is that notions such as intentionality are attributed when the agent violates a norm, but not when he conforms to it.

Alfano et al. (2012) are also concerned with how the agent responds to a norm, but focus on the consequences it has for the way in which a rational agent forms beliefs. They propose that 'the expected payoffs for violating and conforming to a norm differ' (ibid.: 269). As violating a norm is more costly than conforming to it, it is more important to hold accurate beliefs about the former as compared to the latter. In fact, Alfano, Beebe, and Robinson claim that 'it is safe to forget about conforming to a norm but unsafe to forget about violating one' (Ibid.: 270). They go on to argue that 'it is rational to attribute beliefs to people who violate norms (and not to attribute beliefs to people who conform to them)' (ibid.: 269). As notions such as intentionality involve belief, the belief asymmetry explains the other asymmetries. In sum:

[20] Many other explanations have been proposed. Particularly important lines of research are the semantic diversity approach mentioned in note 9 and the conversational pragmatics approach defended by Adams and Steadman (2004a, b, for criticism see Alexander 2012: 54-57, Alfano et al. 2012: 278-79, Knobe 2010b: 324-26 and Machery 2008: 182). I confine my discussion to three families of explanations, because this suffices to defend an all-round approach to theory selection in experimental philosophy.

Table 1 Explanations

Explanations	Explanatory factor
BAD	Moral valence of the side effect (good/bad)
BAD2	Moral valence + the agent's attitude
BLAME	Responsibility attributions (no-praise/blame)
NORM-violation	Response to a norm (conform/violate)
NORM-cost	Costs of having a belief about a norm (conform/violate)
NORM-reason	Indifference about the side effect (beneficial/harmful)

> *NORM*-cost. Responding to a norm (conform/violate) comes with costs (low/high). They bear on the rationality of forming beliefs about it (no/yes), which explains the intentionality attributions (no/yes).

This cost explanation applies directly to belief (and knowledge), and indirectly to intentionality – and as discussed below to other conative notions.

The explanation that I have proposed focuses on the way in which agents make decisions and form intentions (Hindriks 2008, 2011). Agents act for reasons. When making a decision, an agent attributes some weight to certain reasons, but not to others. An agent who is indifferent about a certain issue attributes no weight to it. What an agent intends depends on the reasons he has for performing the action. Notions such as intentional action are sensitive more generally to how an agent responds to the reasons she has.

Crucially, this holds not only for the reasons that actually motivate the agent (motivating reasons); but also for the reasons that should motivate him (normative reasons). A harmful side effect provides the agent with a normative reason not to perform the intended action. In contrast, a beneficial side effect constitutes a reason in favor of performing it. I have argued that the relevant notions apply if the agent flouts a reason that counts against the intended action, but not if it counts in favor of it:

> *NORM*-reason. Indifference with respect to a normatively significant side effect (beneficial/harmful) explains the intentionality attributions (no/yes).

Below I discuss whether this explanation could and should be taken to generalize to other notions.

Explanatory Unification So how does the *NORM* model compare to *BAD2*? The *NORM* model prioritizes the perspective of the agent. Because of this, it can explain IA-Perspective. The explanatory factors that the three accounts invoke can also influence the way in which people apply other conative notions. Hence, the *NORM* model can in principle be extended to the Conative Asymmetries.[21] *NORM*-cost has been designed

[21] The proponents of the *NORM*-violation account and the *NORM*-cost account actually take their accounts to apply to these notions. As not all of them require a special explanation, this is not necessarily a virtue. The *NORM*-reason account extends to them only insofar as they cannot be explained in terms of prior theoretical commitments.

to explain the Epistemic Asymmetries, and is the only *NORM* account that does so.[22] As it explains IA-Perspective, the scope of the *NORM* model is larger than that of *BAD2*. The extent to which *NORM*-cost unifies is substantially larger than *BAD2*, because it also accounts for the Epistemic Asymmetries.

Theoretical Integration A proper assessment of a theory looks beyond unification and considers how it fares with respect to other explanatory virtues. The fact that the *NORM* model has a wider scope does not as such establish that it is to be preferred to *BAD2*. So how does it fare with respect to theoretical integration? As it is an agent explanation rather than an attributor explanation, it preserves the idea that notions such as intentional action are frame of mind notions (i). The *NORM* model is consistent with the idea that attitudes should play a role in explicating notions such as intentional action (ii). The significance of this can hardly be overstated, as these are core commitment in our understanding of intentional action and related notions. In section 4.2, I argued that *BAD2* abandons (i) and (ii). Thus, the *NORM*-model outperforms *BAD2* both on empirical and on theoretical grounds (see Table 2).

The upshot is that explanatory unification is not the only explanatory virtue that matters for assessing the quality of a theory. The extent to which it preserves existing theoretical commitments is, or at least can be, important as well. Sensitivity to such commitments is important for providing a suitable description of the explanandum. In particular, the agent's indifference becomes salient as an important factor against the background of (i)-(iii). In this way, the discussion of theoretical integration contributes to the main point of this paper: the way in which the explanandum is described is crucial for assessing the plausibility of an explanation.

4.3 *NORM*: Violations vs Costs vs Reasons

A similar line of reasoning can be employed to adjudicate between the different *NORM*-accounts. This provides a further illustration of the methodological point just made. And it is of independent interest for the debate on how best to explain the Normativity Findings.[23]

Theoretical Integration I mentioned that all *NORM* accounts are consistent with (ii). But *NORM*-violation does not invoke them in explaining the asymmetries. Beliefs play a role in *NORM*-cost.[24] *NORM*-reason actively utilizes this commitment, as it focuses on the reasons for action that the agent has and how it respond to them. It takes the fact that the agent is indifferent with respect to the side effect to play a key role in the explanation. Because of his indifference, the agent flouts the reason that the beneficial

[22] Alfano et al. (2012) and Beebe (2013) criticize rival explanations of the Knobe Effect for being inconsistent with the belief asymmetry. As beliefs play no role in the other *NORM* accounts, this criticism is too rash.

[23] Even so, those who believe I have proved my point are welcome to skip to section 4.4. It is important that the main point of the paper does not depend on the discussion of this section, also because I will end up arguing that my own explanation, *NORM*-reason, performs best.

[24] The claim that belief is sensitive to costs is controversial and may well violate core commitments of epistemological theories. This need not be a problem if the Epistemic Asymmetries are regarded as biases. But then they should not be regarded as part of the same explanandum.

Table 2 Explanatory payoff

Explanations	Explanatory successes[a]	Explanatory failures	Explanatory virtues
BAD	- Knobe Effect - IA-Aesthetics	- IA-Perspective - Conative[b] - Epistemic Notions	- Unification: medium[c] - Integration: low - Depth: low
BAD2	- Knobe Effect - IA-Aesthetics - Conative Notions[b]	- IA-Perspective - Epistemic Notions	- Unification: medium[c] - Integration: medium - Depth: medium
BLAME	- Knobe Effect	- IA-Aesthetics - IA-Perspective - Conative Notions[b] - Epistemic Notions	- Unification: minimal - Integration: low - Depth: low
NORM-violation	- Knobe Effect - IA-Aesthetics - IA-Perspective - Conative Notions[b]	- Epistemic Notions	- Unification: high - Integration: low - Depth: low
NORM-cost	- Knobe Effect - IA-Aesthetics - IA-Perspective - Conative Notions[b] - Epistemic Notions		- Unification: maximal - Integration: low - Depth: low
NORM-reason	- Knobe Effect - IA-Aesthetics - IA-Perspective - Conative Notions[b]	- Epistemic Notions	- Unification: medium - Integration: high - Depth: high

[a] As no asymmetry has been established, I exclude IA-Prudence from this table

[b] The table indicates whether an account *can* explain these asymmetries, not whether it should

[c] When Knobe (2010b) claims that his explanation provides a maximal degree of unification, he does not consider IA-Perspective and the Epistemic Asymmetries

or harmful prospect constitutes. Even though it rejects (iii) – as all other explanations of the Normativity Findings do – it is a relatively conservative extension of existing theories of intentional action. As such, it departs the least from existing theories of intentional action. Hence, theoretical integration counts in favor of *NORM*-reason.[25]

Explanatory Depth In addition to theoretical integration, explanatory depth is an important virtue of a theory. The depth of an explanation increases when one of its presuppositions is itself explained, for instance by providing a mechanism (Ylikoski and Kuorikoski 2010). How do the *NORM* accounts fare with respect to explanatory depth? *NORM*-violation postulates an asymmetry between violating and conforming to a norm. Holton (2010) observes that knowingly violating a norm is done intentionally, whereas knowingly conforming to a norm can be done unintentionally. By way of diagnosis, he points out that intentionally conforming to a norm requires counterfactual guidance. Importantly, an indifferent agent is not counterfactually guided by the norm

[25] In Hindriks 2008 I argue that it fits well with Michael Bratman's (1987) planning theory of intentional action. Below I argue that it also preserves the common sense conception of the relation between intentional action and moral responsibility.

to which she happens to conform. The upshot is that knowingly violating a norm suffices for doing so intentionally, whereas knowingly conforming to a norm does not.

Although true, none of this is very informative. Counterfactual guidance is intimately related to intending (Bratman 1987). By invoking it, Holton in effect maintains that intentionally conforming to a norm requires the intention to do so, and in the help condition the agent clearly does not intend to help. In the harm condition, Holton assumes that knowingly violating a norm against harming implies intentionality. If his explanation is to have any depth, Holton should explain why this is the case. In fact, he does little more than use a contrast between intentionally violating a norm and unintentionally conforming to a norm to account for the Knobe Effect, which is itself an asymmetry in intentionality attributions. Little if any explanatory progress is made, because the explanandum and the explanans are intimately related to one another.

According to *NORM*-cost, the costs of false beliefs are higher for norm violations than norm conformations. It is not very clear, however, how this cost-asymmetry could lead to different belief attributions. An indifferent agent pays little attention to a norm. Hence, he ignores both the high costs of violating the norm as well as the low costs of conforming to it. After all, someone who takes little care to conform to a norm may well end up violating one. The notion that he chooses to ignore one of these costs makes sense only because, in the scenarios at issue, another agent tells him that he will conform to or violate a norm if he performs the intended action. This raises doubts about the explanation of the belief asymmetry. It seems to be an unexplained explainer, which detracts from the depth that its proponents claim for their account.

NORM-reason provides a deeper explanation than its rivals. The way in which an agent responds to reasons, and thereby what motivates him, is relevant to the attribution of moral responsibility. An agent is praiseworthy for bringing about a beneficial effect only if brings about that effect for the right reasons. Because of this, the motivation of an agent who is indifferent with respect to a beneficial effect does not support the attribution of praise. This in turn means that the attribution of praise is not at all facilitated by qualifying the action of an agent who ignores the reasons she has for bringing about an effect as intentional. It would only be confusing to claim that he brings it about intentionally, as this could easily be taken to imply that his motivation was morally adequate.

Things are different when it comes to harmful effects. Someone who is indifferent with respect to a harmful side effect thereby flouts a reason against performing the intended action. This suggests that his motivation is morally inadequate, and that he should have responded in a more appropriate manner. Note, however, that this does not settle whether he is to blame: he might be able to justify his action or present some excuse for performing it. Instead, it provides a defeasible reason for believing that he is blameworthy. As such, this information is clearly relevant to the attribution of blame. Because it concerns the motivation of the agent, it would be useful to mark this in terms of intentionality.

At this point, one might worry that the account in effect invokes responsibility attributions to explain intentionality attributions in roughly the way that *BLAME* does. This is not the case. According to *NORM*-reason, the notion of intentional action is asymmetric in a way that is informative to the attribution of moral responsibility.

However, intentionality as such does not settle whether or not the agent is praise- or blameworthy. In some cases, it is justified for an agent to attribute little or no weight to a harmful side effect. Think, for instance, of the classic trolley dilemma in which an agent saves five by steering the trolley onto a sidetrack thereby causing the death of one. Perhaps it would be odd or inappropriate to say 'I don't care about the person on the sidetrack'. Even so, for practical purposes his presence is of little relevance. So claiming that he caused the death of the person on the sidetrack intentionally does not as such entail that the agent is blameworthy.[26]

As it is an open question whether an agent who brought about a harmful effect intentionally is blameworthy for doing so, intentionality attributions are prior to responsibility attributions according to *NORM*-reason. Instead, notions such as intentional action provide information that is relevant to the attribution of moral responsibility, albeit partial and defeasible. For the reasons just explained, they mark the inadequacy of the agent's motivation (only) when this counts towards the attribution of moral responsibility. In order to appreciate this, it is useful to consider side-effects that are brought about unintentionally. This will be the case if the agent is indifferent with respect to a normatively insignificant side effect, for which the agent is not to blame at all. It will also be the case if the effect is harmful, but rather than being indifferent with respect to it, the agent is negligent. Indifference is a major flaw in the agent's motivation, which is why it warrants more blame than negligence. Because of this motivational defect, the action is marked as intentional. This is informative because it provides a defeasible reason to attribute a substantial amount of blame, more than for outcomes that are brought about unintentionally. The upshot is that this account preserves the common-sense conception of the relation between intentionality and moral responsibility (Hindriks 2008, 2011). Its rationale consists in the usefulness that the asymmetries have for marking the agent's motivation in a way that is conducive to the ascription of moral responsibility.

This analysis supports moderate optimism about the extent to which an explanation of the Normativity Findings can and should be unified. Although *NORM*-cost outperforms other agent explanations when it comes to scope, *NORM*-reason is more attractive overall: it is better integrated with related theories and has more theoretical depth. As the fourth column of Table 2 reveals, the degree of unification of this account is medium. Knobe criticized it for leaving us 'with a mystery as to why the impact of moral judgment is so pervasive' (2010b: 355). In response, I have argued that it is only unified to the extent to which this is warranted by the phenomena. Hence, the fact that its scope is narrower than Knobe takes *BAD2* to be is not a disadvantage but rather speaks in its favor.[27]

[26] The fact that the agent expresses indifference does not necessarily entail that she does not care. Indifference need not be interpreted in terms of strict *neutrality* – the absence of pro or con attitudes. An agent can also be indifferent with respect to an effect of a decision, because it she regards it as *irrelevant to the decision at hand*. This *irrelevance* interpretation is as such consistent with the agent caring about the effect. Even though she will attribute weight to it in other contexts, she believes that it does or should not make a difference to the decision at hand. An agent who is indifferent in this second sense may well regret the fact that her decision has a harmful effect. This interpretation suggests that an expression of indifference need not be regarded as odd or inappropriate.

[27] Knobe (2010b) claims that *BAD2* provides a maximal degree of unification, but his assessment does not include IA-Perspective and the Epistemic Asymmetries (as indicated under Table 2).

4.4 The Identification Problem Revisited

My argument in favor of the *NORM* model illustrates how the Identification Problem can be solved. Section 3 used similarity as a criterion to properly describe the empirical finding. As the descriptions contained the explanandum, this was the first step towards solving the Identification Problem. Section 4 considered novelty. This criterion was used to argue that the Conative Asymmetries should not be regarded as part of the explanandum. I also argued that theoretical integration is an important theoretical virtue. It acquires special significance in the light of IA-Perspective.

The descriptions of the empirical findings were often intimately related to the proposed explanatory factors. *BAD1*is basically a restatement of Knobe's description of the phenomenon. I should note that, due to the role of indifference, *NORM*-reason is rather similar to [DEF4]. It should now be clear that this is not a problem as such. A good explanation is a proper description of the causal relation between the explanans and the explanandum. The point, however, is that it is a delicate process to arrive at an adequate formulation of the empirical finding. The discussion of the Normativity Findings above reveals a process of mutual adjustment between the explanandum and the explanans. This process – and thereby solving the Identification Problem – is vital to identifying a proper explanation of the Normativity Findings.[28]

5 Conclusion

I have argued that unification is possible to the extent it is desirable. Unification is desirable to the extent that a number of empirical findings constitute one phenomenon, and unification is possible to the extent that one factor explains that phenomenon. These claims support a process of re-description of explanandum and explanans that is conducive to unification. As illustrated, this process provides a solution for the Identification Problem, the problem of identifying the explanandum in a way that is at least to some extent independent of knowledge of the explanans.

The position that I defend is not overly optimistic because it avoids unification for the sake of unification, or unification that is motivated only by a preference for cognitive efficiency. It is not overly pessimistic because it encapsulates a valuable heuristic for identifying the degree to which an explanation can be unified (medium; see Table 2). Instead, it is realistic. I have argued that unchecked optimism about explanatory unification is counterproductive, as unification is valuable only to the extent that it corresponds to underlying unity. In this sense, the alternative I have presented is realist: it is conducive to identifying the appropriate degree of unification, which is why it enables experimental philosophers to keep it real.

[28] The theory of bootstrap confirmation can be used for providing a formal solution to the Identification Problem (Glymour 1980; Douven and Meijs 2006).

References

Adams, F., and A. Steadman. 2004a. Intentional Action in Ordinary Language: Core Concept or Pragmatic Understanding. *Analysis* 64: 173–181.

Adams, F., and A. Steadman. 2004b. Intentional Actions and Moral Considerations: Still Pragmatic. *Analysis* 64: 264–267.

Alexander, J. 2012. *Experimental Philosophy: An Introduction*. Cambridge: Polity Press.

Alfano, M., J. Beebe, and B. Robinson. 2012. The Centrality of Belief and Reflection in Knobe-Effect Cases: A Unified Account of the Data. *The Monist* 95: 264–289.

Alicke, M.D. 2008. Blaming Badly. *Journal of Cognition and Culture* 8: 179–186.

Anscombe, G.E.M. 1957. *Intention*. Oxford: Basil Blackwell.

Bartelborth, T. 2002. Explanatory Unification. *Synthese* 130: 91–107.

Bealer, G. 1998. Intuition and the Automomy of Philosophy. In *Rethinking Intuition*, ed. M. DePaul and W. Ramsey, 201–240. Lanham: Rowman and Littlefield.

Beebe, J. 2013. A Knobe Effect for Belief Ascriptions. *Review of Philosophy and Psychology* 4: 235–258.

Beebe, J.R., and W. Buckwalter. 2010. The Epistemic Side effect Effect. *Mind & Language* 25: 474–498.

Beebe, J.R., and M. Jensen. 2012. Surprising Connections between Knowledge and Action: The Robustness of the Epistemic Side effect Effect. *Philosophical Psychology* 25: 689–715.

Bratman, M. 1987. *Intention, Plans, and Practical Reason*. Cambridge: Harvard University Press.

Cokely, E.T., and A. Feltz. 2009. Individual Differences, Judgment Biases, and Theory-of-Mind: Deconstructing the Intentional Action Side Effect Asymmetry. *Journal of Research in Personality* 43: 18–24.

Cushman, F., and A. Mele. 2008. Intentional Action: Two-and-a-Half Folk Concepts? In *Experimental Philosophy*, ed. J. Knobe and S. Nichols, 171–188. New York: Oxford University Press.

Douven, I., and W. Meijs. 2006. Bootstrap Confirmation Made Quantative. *Synthese* 149: 97–132.

Dretske, F. 1988. *Explaining Behavior: Reasons in a World of Causes*. Cambridge: MIT Press.

Duff, A. 1990. *Intention, Agency and Criminal Liability*. Oxford: Blackwell Press.

Enç, B. 2003. *How We Act: Causes, Reasons, and Intentions*. Oxford: Oxford University Press.

Friedman, F. 1974. Explanation and Scientific Understanding. *Journal of Philosophy* 71: 5–19.

Gallagher, S. 2005. *How the Body Shapes the Mind*. Oxford: Oxford University Press.

Glymour, C. 1980. *Theory and Evidence*. Princeton: Princeton University Press.

Guglielmo, S., and B.F. Malle. 2010. Can Unintended Side Effects Be Intentional? Resolving a Controversy Over Intentionality and Morality. *Personality and Social Psychology Bulletin* 36: 1635–1647.

Guglielmo, S., A.E. Monroe, and B.F. Malle. 2009. At the Heart of Morality Lies Folk Psychology. *Inquiry* 52: 449–466.

Harman, G. 1976. Practical Reasoning. *Review of Metaphysics* 29: 431–463.

Hindrikds, 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: 787–801.

Hindriks, F. 2014. 'Normativity in Action: How to Explain the Knobe Effect and Its Relatives'. *Mind & Language* 29: 51–72.

Hindriks, F., I. Douven, and H. Singmann. 2016. 'A New Angle on the Knobe Effect: Intentionality Correlates with Blame, Not with Praise'. *Mind & Language* 31(2): 204–220.

Hitchcock, C., and E. Sober. 2004. Prediction versus Accommodation and the Risk of Overfitting. *The British Journal for the Philosophy of Science* 55: 1–34.

Holton, R. 2010. Norms and the Knobe Effect. *Analysis* 70: 417–424.

Kauppinen, A. 2007. The Rise and Fall of Experimental Philosophy. *Philosophical Explorations* 10: 95–118.

Kitcher, P. 1989. Explanatory Unification. *Philosophy of Science* 48: 507–531.

Knobe, J. 2003. Intentional Action and Side Effects in Ordinary Language. *Analysis* 63: 190–194.

Knobe, J. 2004a. Intention, Intentional Action and Moral Considerations. *Analysis* 64: 181–187.

Knobe, J. 2004b. Folk Psychology and Folk Morality: Response to Critics. *Journal of Theoretical and Philosophical Psychology* 24: 270–279.
Knobe, J. 2006. The Concept of Intentional Action: A Case Study in the Uses of Folk Psychology. *Philosophical Studies* 130: 203–231.
Knobe, J. 2007. Reason Explanation in Folk Psychology. *Midwest Studies in Philosophy* 31: 90–106.
Knobe, J. 2010a. Person as Scientist, Person as Moralist. *Behavioral and Brain Sciences* 33: 315–329.
Knobe, J. 2010b. The Person as Moralist Account and Its Alternatives. *Behavioral and Brain Sciences* 33: 353–365.
Knobe, J. 2016. Experimental Philosophy is Cognitive Science. In *Companion Experimental Philosophy*, ed. W. Buckwalter and J. Sytsma: 37–52. New York:Blackwell.
Knobe, J., and B. Fraser. 2008. Causal Judgment and Moral Judgment: Two Experiments. In *Moral Psychology (volume 2)*, ed. W. Sinnott-Armstrong. Cambridge (MA): MIT Press.
Knobe, J., and G. Mendlow. 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.
Lewis, D. 1983. *Philosophical Papers (volume 1)*. New York: Oxford University Press.
Ludwig, K. 2007. The Epistemology of Thought Experiments: First Person versus Third Person Approaches. *Midwest Studies in Philosophy* 31: 128–159.
Machery, E. 2008. Folk Concept of Intentional Action: Philosophical and Experimental Issues. *Mind & Language* 23: 165–189.
Machery, E., R. Mallon, S. Nichols, and S. Stich. 2004. Semantics: Cross-Cultural Style. *Cognition* 92: B1–B12.
Mäki, U. 2001. Explanatory Unification: Double and Doubtful. *Philosophy of the Social Sciences* 31: 488–506.
Malle, B., and S. Nelson. 2003. Judging Mens Rea: The Tension between Folk Concepts and Legal Concepts of Intentionality. *Behavioral Sciences & the Law* 21: 563–580.
McCann, H.J. 2005. Intentional Action and Intending: Recent Empirical Studies. *Philosophical Psychology* 18: 737–748.
Mele, A. 2001. Acting Intentionally: Probing Folk Notions. In *Intentions and Intentionality: Foundations of Social Cognition*, ed. B. Malle, L. Moses, and D. Baldwin, 27–43. Cambridge: MIT Press.
Morrison, M. 1990. Unification, Realism and Inference. *British Journal for the Philosophy of Science* 41: 305–332.
Myrvold, W.C. 2003. A Bayesian Account of the Virtue of Unification. *Philosophy of Science* 70: 399–423.
Nadelhoffer, T. 2004a. Praise, Side Effects, and Intentional Action. *Journal of Theoretical and Philosophical Psychology* 24: 196–213.
Nadelhoffer, T. 2004b. Blame, Badness, and Intentional Action: A Reply to Knobe and Mendlow. *Journal of Theoretical and Philosophical Psychology* 24: 259–269.
Nadelhoffer, T. 2005. Skill, Luck, Control, and Intentional Action. *Philosophical Psychology* 18: 341–352.
Nadelhoffer, T. 2006. Desire, Foresight, Intentions, and Intentional Actions: Probing Folk Intuitions. *Journal of Cognition and Culture* 6: 133–157.
Nadelhoffer, T., and E. Nahmias. 2007. The past and future of experimental philosophy. *Philosophical Explorations* 10(2): 123–149.
Nado, J. 2008. Effects of Moral Cognition on Judgments of Intentionality. *British Journal for the Philosophy of Science* 59: 709–731.
Nichols, S., and J. Ulatowski. 2007. Intuitions and Individual Differences: the Knobe Effect Revisited. *Mind & Language* 22: 346–365.
Pellizoni, S., V. Girotto, and L. Surian. 2010. Beliefs and Moral Valence Affect Intentionality Attributions: the Case of Side Effects. *Review of Philosophy and Psychology* 1: 201–209.
Pettit, D., and J. Knobe. 2009. The Pervasive Impact of Moral Judgments. *Mind & Language* 24: 586–604.
Phelan, M., and H. Sarkissian. 2009. Is the 'Trade-off Hypothesis' Worth Trading For? *Mind & Language* 24: 164–180.
Phillips, J., and J. Knobe. 2009. Moral Judgments and Intuitions About Freedom. *Psychological Inquiry* 20: 30–36.
Phillips, J., L. Misenheimer, and J. Knobe. 2011. The Ordinary Concept of Happiness (and Others Like It). *Emotion Review* 3: 320–322.
Pinillos, N.A., N. Smith, G.S. Nair, P. Marchetto, and C. Mun. 2011. Philosophy's New Challenge: Experiments and Intentional Action. *Mind & Language* 26: 115–139.
Sosa, E. 1998. Minimal Intuition. In *Rethinking Intuition*, ed. M. DePaul and W. Ramsey, 257–270. Lanham: Rowman and Littlefield.
Weinberg, J., S. Nichols, and S. Stich. 2001. Normativity and Epistemic Intuitions. *Philosophical Topics* 29: 429–460.
Williamson, T. 2007. *The Philosophy of Philosophy*. Oxford: Blackwell Publishing.

Woodward, J. 2011. Scientific Explanation. In *Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta, Winter 2011 ed. URL = <http://plato.stanford.edu/archives/win2011/entries/scientific-explanation/>.

Wright, J.C., and J. Bengson. 2009. Asymmetries in Judgments of Responsibility and Intentional Action. *Mind & Language* 24: 24–50.

Ylikoski, P., and J. Kuorikoski. 2010. Dissecting Explanatory Power. *Philosophical Studies* 148: 201–219.