Three-and-a-half folk concepts of intentional action

by Alessandro Lanteri (alelanteri@gmail.com)

POLIS, Faculty of Political Science, Università del Piemonte Orientale - Alessandria (Italy)

forthcoming on: *Philosophical Studies*

this is an unedited draft. please refer to and quote from the published version

Abstract (170 words): Fiery Cushman and Alfred Mele recently proposed a 'two-and-a-half rules' theory of folk intentionality. They suggested that laypersons attribute intentionality employing: one rule based on desire, one based on belief, and another principle based on moral judgment, which may either reflect a folk concept (and so count as a third rule) or a bias (and so not count as a rule proper) and which they provisionally count as 'half a rule'. In this article, I discuss some cases in which an agent is judged as having neither belief nor desire to bring about an action, and yet laypersons find the agent's action to be intentional. Many lay responses apparently follow a rule, but many other seem biased. The contribution of this study is two-fold: by addressing actions performed without desire or belief, it expands Mele and Cushman's account; it also helps discriminate between a two-rules and a three-rules theory. As a conclusion, I argue in favor of a three-and-a-half concepts theory.

Keywords: Action Theory; Experimental Philosophy; Folk Concepts; Knobe Effect; Intentionality

This article appears on the pages of a reputable academic journal, thanks to the colleagues and the referee who helped improve it, and ultimately thanks to the editor who approved of its publication. Its author can then record a new publication in his cv. Now imagine that, soon after the article appears in print, the author's department faces a major budget cut. The dean must now make the difficult choice of firing one of the two experimental philosophers – a colleague and the author – because there are not enough funds for the continued employment of both. Imagine that the two are equally desirable faculty members under every respect. However, thanks to this very article, the publication record of the author is now superior to that of the colleague. Hence the dean resolves to retain the author and fire the colleague.[1]

Did the journal editor bring about the firing of the colleague? He arguably did. But did he do so *intentionally*? One would be tempted to answer negatively: the editor did not anticipate the firing, so he could not have intended to bring it about. According to many philosophers, however, it is possible to perform an action intentionally even when not intending to perform it. For example, it can be argued that a bomber intentionally kills several children in a nursery – even if he does *not* want to – if he drops a bomb on the neighboring arms factory, which he anticipates will also blow up the nursery. Sometimes, therefore, a justified belief that something will happen, though as an undesired side-effect to a deliberate action, is enough for establishing that the side-effect had been brought about intentionally, especially when it is morally bad. Unlike the bomber, however, the editor in the earlier example did not anticipate what would happen to the laid-off colleague. Does this make a difference?

The traditional philosophical approach to this question would be to argue in support of the intuition that the lacking of relevant beliefs about the outcome of an action implies that such action could not have been performed intentionally, and to rebuff the arguments against such intuition. Instead, I will address this issue with the methods of experimental philosophy.[2] In a nutshell, these methods amount to presenting laypersons with a vignette (for example, one describing a situation like that of editor), and then asking whether they think the side-effect has been brought about intentionally. A set of lay responses to carefully designed scenarios lets us infer the principles from which such answers can plausibly follow, and the ensemble of these principles roughly constitutes a folk theory.

The result, of course, is not a philosophical but a lay response to the question. My goal in this article, therefore, is not to establish an account of intentionality that is superior to those proposed by other philosophers, but to contribute to an understanding of the way in which philosophically unsophisticated people attribute intentionality. Such contribution is nonetheless of value to a philosophical discussion. For one, the philosophical treatment of intentional action "runs the risk of having nothing more than a philosophical fiction as its subject matter" (Mele 2001, p. 27) when it remains indifferent to the folk concept. Furthermore, a better understanding of the actual judgments of laypersons can help philosophers develop better normative theories.

---

[1] At the time of writing, no other experimental philosopher was employed at the institution where the author was employed.

[2] For an introduction, see: Nadelhoffer and Nahmias (2007), Knobe and Nichols (2008).

Folk theories of intentionality have been a privileged topic of investigation in experimental philosophy (e.g., Adams and Steadman 2004; Knobe 2003, 2004, 2006; Knobe and Mendlow 2004; Machery 2008; Malle 2006; McCann 2005; Mele 2001, 2003; Nadelhoffer 2004, 2006a,b; Young et al. 2006). Among these theories, a prominent account is that proposed by Fiery Cushman and Alfred Mele (2008, Mele and Cushman 2007) and labeled 'two-and-a-half folk concepts of intentionality' theory, which is a compromise between a two-concepts theory (one desire-based and one belief-based) and a three-concepts theory (one based on moral judgment in addition to the previous two). Each such concept is embodied in a rule for the attribution of intentionality to some action.[3]

In this article, I discuss some cases in which an agent is judged as having neither belief nor desire to bring about an action, and yet laypersons find such action to be intentional. Therefore, the contribution of this study is two-fold: by addressing actions performed without desire or belief, it expands Mele and Cushman's account and it helps discriminate between a two-rules and a three-rules theory. As a conclusion, I argue in favor of a three-and-a-half rules theory.

1. Types of actions and folk theories of their intentionality

The starting point of any discussion of experimental philosophy and of folk theories of intentionality ought to be the 'chairman experiment' vignette study by Joshua Knobe (2003, p. 191), which arguably brought the experimental philosophy movement under the spotlight. He gave half of the participants in his experiment the following vignette:

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

The other half received instead a vignette in which "harm" was replaced with a different side-effect: "help." He then asked the participants how much blame (praise) did the chairman deserve for his decision and whether the harmful (beneficial) effects to the environment had been brought about intentionally. In the first treatment, 82% of the respondents said that the chairman brought about the harm intentionally and deserved a lot of blame for this (the mean value for blame was 4.8 on a 0 to 6 scale). In the other treatment, only 23% of the respondents found that the chairman intentionally helped the environment. Moreover, he did not deserve praise (the mean praise was 1.4 on the same scale). This asymmetry in intentionality (and that in moral judgment) is now commonly referred to as the 'Knobe effect' or as the 'side-effect effect.'

In the wake of this experiment, the issue of folk attributions of intentionality to various types of actions has come to the forefront of experimental philosophy. Mele and Cushman recently started sorting the responses laypersons have given to the question whether an agent did ϕ intentionally in a large number of vignette studies. Consistent with earlier findings, they show that the responses sometimes depend on the moral worth – positive, negative, or neutral – of

[3] In what follows I will therefore employ "concept(s)" and "rule(s)" interchangeably.

ϕ and that the actions with morally loaded, and negative outcomes are especially likely to be judged intentional. Mele and Cushman found that the responses also depend on whether ϕ is a deliberate end or a means to an end (*seek-type action*), or a foreseen side-effect to a seek-type action (*anticipated side-effect action*, or *ASEA*). The former obtains when the agent has a 'desire' to ϕ, the latter when the agent has a justified 'belief' that she will ϕ – that is, she is "*very confident*" she will ϕ (Cushman and Mele 2008, p. 173, emphasis in the original). In a nutshell, laypersons usually say that ϕ is performed intentionally when its agent has a desire to ϕ, a belief that she will ϕ (often limited to the cases when ϕ is morally bad), or both a desire and a belief (see below).

Beside seek-type actions and ASEAs, Mele and Cushman's taxonomy includes actions for which the agent neither sought nor anticipated ϕ (*0-type actions*). If ϕ is not sought, then it will come about as a side-effect. When such side-effect is not anticipated, the agent of ϕ also lacks a justified belief. For example, the journal editor's bringing about the firing of one experimental philosopher in the short case above counts as a 0-type action.

Mele and Cushman report a single example of 0-type actions: a hunter shoots at a deer and hits the target, but then his bullet passes through the deer to hit a bird watcher who was standing behind it. The hunter did neither seek to shoot the bird watcher nor did he see this coming, though "he certainly should have known" (Mele and Cushman 2007, p. 195). Although the outcome is morally bad, only 19% of the respondents said that the hunter shot the bird watcher intentionally.[4]

If non-sought and unanticipated actions are deemed not intentional even in a morally bad outcome scenario, then one may be led to thinking that they would be deemed not intentional also in a scenario with a morally neutral or morally good outcome. It is fair to point that this is not Mele and Cushman's explicit suggestion. Indeed, their treatment of 0-type actions remains marginal and they do not clearly locate this type of actions in their theory. Instead, the studies below investigate an action of the 0-type.

2. A modified chairman experiment

Knobe's chairman experiment has two vignettes: in one, the chairman launches a new business program that is anticipated to and in fact does result in a bad outcome (harming the environment); in the other, the program is anticipated to and in fact does result in a good outcome (helping the environment). The chairman is confident that his decision will affect the environment, but in either case he does not "care at all" (Knobe 2003, p. 191).[5] He just desires to secure the profits he anticipates from the new program. Hence, the harm/help to the environment is an ASEA.

In a modified version of the chairman experiment (Lanteri 2009), the chairman does not care about (hence, he does not seek), but neither does he foresee the consequences of his decision (hence, he has no justified belief). So, the

[4] Mele and Cushman (2007, p. 186) asked "did the hunter intentionally shoot the bird watcher?" and elicited responses on a 1-to-7 scale (where 1 is a strong no and 7 is a strong yes). They counted as 'yes' and 'no' respectively the responses ≥ 5 and ≤ 3,

[5] See Beebe and Buckwalter (2010) for a dissenting view.

outcome is a 0-type action. Also this study had two treatments – HARM and HELP – and the setting was the same for both:

> *An executive of a company goes to the Chief Executive Officer and tells him: 'We are thinking of launching a new product, which will increase profits. This product requires a new technique and we do not know what consequences this will have on the environment.' The CEO replies: 'I do not care at all about the consequences. I just want to make profits. We shall launch the new product immediately.' The new product is launched and, as predicted, profits increase.*

The difference was in the finale. The participants were randomly assigned to either the HARM or the HELP treatment, which read as follows:

> [HARM] *However, the new technique turns out to be polluting and it harms the environment.*
>
> [HELP] *Moreover, the new technique turns out to be ecological and it helps the environment.*

The respondents (N=52) were then asked whether the CEO intentionally harmed/helped the environment and whether he deserved praise or blame.[6]

What responses should one expect? As mentioned, Mele and Cushman are not very explicit about 0-type actions. From their hunter study, one could form the expectation that unanticipated actions are considered not intentional, regardless of a bad outcome. Plausible though this may sound, the very point of experimental philosophy is to not mistake plausible for truthful.

Here, both scenarios present an identical decision (launch a new product), based on the same information (there will be some unknown consequence on the environment), following from the same intention (make profits without regard for the environment), and disregarding any side-effect. What eventually occurs is found out only after the decision and, depending on the experimental treatment, it is described as either good or bad. Quite unexpectedly, such difference does significantly affect the responses and this study reveals two asymmetries akin to the Knobe effect.

79% of the respondents found the chairman blameworthy in HARM, and 43% said he deserved blame also in HELP. There is also an asymmetry in the attribution of intentionality: 58% responded that the environmental harm was intentional, whereas only 7% found help intentional. The difference is both large and puzzling.

In Knobe's original study more respondents stated that the action was intentional (remind: 82% in HARM and 23% in HELP) than in the modified chairman experiment.[7] Knobe's vignettes described a decision made by a chairman who was told in advance about the relevant side-effect. In the modified experiment, the decision-maker is told that the side-effect is unknown. So, the advance information about the side-effect may account for the difference between Knobe's study and mine.

The asymmetries between the pairs of treatments in each experiment, however, remain to be explained.

---

[6] The participants were asked: *According to you, the CEO harmed/helped the environment… a) intentionally, b) not intentionally, c) neither*; and then: *According to you, for his decision, the CEO should be… a) praised, b) blamed, c) neither*.

[7] This figure is also different from that of the respondents who said that the chairman did not intentionally harm the environment in Mele and Cushman's (2008, p. 176) experiment based on the original vignette (i.e., 5% when the question was asked at the beginning of the experiment and 27% when the question was asked towards the end of the experiment). At any rate, direct comparisons between these studies should be avoided, because they differ under several respects, and so the differences may be attributable to a plurality of causes.

2.1. Knowledge and intentionality

Before attending to the explanation of these asymmetries, it is important to establish that the respondents agree that the CEO did not have a belief that the environment would be harmed.

Like most commentators, I have assumed that the chairman in Knobe's experiment does know the side-effect. However, he is merely told that the environment would be helped/harmed but "no explicit statement of the chairman's knowledge is given" (Beebe and Buckwalter 2010).[8] To find out whether laypersons think he knows, James Beebe and Wesley Buckwalter (2010) employed the original chairman vignettes and asked: *Did the chairman know that the new program would help/harm the environment?* The respondents in HARM were more likely to say that the chairman did know the side-effect than those in HELP. Hence, even attributions of epistemic states seem to be affected by the moral valence of side-effects, a phenomenon Beebe ad Buckwalter label 'epistemic side-effect effect.'

The epistemic side-effect effect invites some caution before taking for granted that the CEO in the modified chairman experiment does not know either side-effect will come about. Perhaps laypersons believe that the CEO does know, at least in HARM, even though the scenario reads: "This product requires a new technique and we do not know what consequences this will have on the environment." If such were the case, the CEO would be regarded as having a belief and so the environmental harm in the vignette would be an ASEA and not a 0-type action.

Sandra Pellizzoni, Vittorio Girotto, and Luca Surian (2010) ran a vignette study similar to the modified chairman experiment, which addresses some of the concerns raised by the epistemic side-effect effect. In their experiment, the chairman does not know the outcome in advance because he is kept in the dark by the vice president, who instead knows the outcome. The relevant sentence in the vignette reads: "The vice-president knew that the program would have harmed the environment but he did not tell it to the chairman" (Pellizzoni et al. 2010, p. 203). Asked whether the chairman had harmed the environment intentionally, 50% of the respondents said that he did.[9]

Since participants are explicitly told that the knowledge about the harmful side-effect is available and that it is withheld from the chairman, it is plausible to assume that these respondents regard the chairman as having neither desire nor belief regarding the harmful side-effect. It is also possible, however, that the respondents reasoned that "any reasonable person could have anticipated this negative side effect" (Pellizzoni et al. 2010, p. 205) and so also the chairman must have foreseen it.

---

[8] Knobe's (2003) original paper features a second scenario that mirrors the chairman scenario and in which a lieutenant sends a squad of soldiers to take control of a hill. Doing so, he disregards whether they move into/out of the enemy's line of fire and hence whether some of the soldiers will be killed/rescued; he only cares about controlling the hill. The lieutenant explicitly declares: "Look, *I know* that they'll be in/taken out of the line of fire, and *I know* that some of them will be killed/would have been killed otherwise" (Knobe 2003, p. 192-3, emphasis added). The intentionality asymmetry (i.e., 77% in HARM and 30% in HELP) is similar to that in the chairman scenario. Beebe and Buckwalter (2010, p.) acknowledge that "because of the structural similarity of the corporate and military cases, one might have expected that subjects would automatically assume the chairman knew his company's new program would have the stated consequences." Surprisingly, they find this not to be the case.

[9] Note that the chairman in this vignette does not say that he doesn't care at all about the environment (Pelizzoni et al. 2010, p. 203), as in the other vignettes discussed here. This difference may be responsible for (lower degrees of blame and hence) lower attributions of intentionality.

Another study by Pellizzoni, Michael Siegal, and Surian (2009) offers more direct evidence. The participants were preschoolers and the vignette featured a story in which Andy brought a frog over to Janine's house, ignoring whether she liked or disliked frogs, and not caring about whether Janine would be happy or upset. In alternative scenarios, Janine turned out to be happy, because she happened to like frogs as well, or upset, because she disliked them. The experimental technique is such that the researchers read the scenarios to individual children, employed some drawings to which they pointed while reading, and asked children two control questions: *Does Andy know if Janine will be happy or upset to see the frog?* and *Does Andy care if Janine will be happy or upset?* This way the children were asked about intentionality only after they responded that Andy did not know and did not care about how Janine would react. Across some variants of the Andy/Janine scenario, just below half (47.7%) of the agents who had no foreknowledge and whose actions had negative outcomes, were judged as having acted intentionally.

Although both these experiments suggest that about half of the respondents consider a negative side-effect intentional even in the absence of relevant knowledge, two concerns may still be raised. In the Andy/Janine experiment, the scenario, the respondents, and the methodology are so different from the original and the modified chairman experiments that direct comparisons would be unreliable. In the chairman experiment, instead, the respondents were not asked about the chairman's knowledge or belief, so the epistemic side-effect effect cannot be ruled out. A more direct investigation of the absence of belief is thus in order.

2.2. Intentionality without belief

As noted above, belief does not require certainty, but at least some confidence.[10] In order to investigate this condition, some respondents (N = 35) were presented with the HARM treatment of the modified chairman vignette, and then asked:

> * *According to you, did the CEO intentionally harm the environment?*
> * *According to you, was the CEO confident that the environment would be harmed?*

Participants could answer either YES or NO to each question. As in the previous experiment, the harm was found intentional by most respondents (71.4%). According to the majority (65.7%) of the respondents, however, the CEO was not confident that the environment would be harmed. Overall, about half (51.4%) of the participants responded that, though the CEO was not confident harm would occur, the harm was intentional.[11]

However, a recurring issue in experimental philosophy, and indeed one of the central themes of inquiry of the movement, is that lay concepts are often a bundle of distinct concepts with different meanings to different people and in

---

[10] The reader could see Nadelhoffer (2006) and Mele and Cushman (2007, p. 194) for a discussion of the influence of various degrees of confidence on lay judgments of intentionality.

[11] In a trial version of this experiment, some respondents (N = 12) were asked whether they thought the CEO was 'very confident', instead of simply 'confident', that harm would occur. 66.7% said that the harm was intentional, 91.7% said that the CEO was not very confident, and those who said both that the harm was intentional and the CEO not very confident were 58.3%. (One respondent did not answer the intentionality question and said that the CEO was not very confident.) In another trial (N = 11), the question asked was whether the CEO 'anticipated' the harm. The same number of respondents (72.7%) said that the harm was intentional and that the CEO did not anticipate the harm. 63.6% said both that the harm was intentional and that the CEO did not anticipate the harm.

different circumstances. For example, folk attributions of knowledge appear to be affected by order effects, even in morally neutral scenarios (Swain et al. 2008). The very word 'intentionality' investigated in the present article is but one example of how laypersons endorse different conceptualizations of the same word, as referring to belief, desire, or both (see also below). One should therefore be hesitant to assume that the folk concept of 'confidence' is identical to that employed by Mele and Cushman.

In order to minimize the reliance of the findings on the notion of confidence, another treatment was conducted (N = 39), in which the participants were given the HARM scenario of the modified chairman vignette, and then asked:

** According to you, did the CEO intentionally harm the environment?*

** According to you, did the CEO think that the environment would be harmed?*

As in the previous treatment, participants answered YES or NO to each question. Also this time, most respondents (66.7%) found the harm intentional and said that the CEO did not think the environment would be harmed (64.1%). About half of the respondents (51.3%) answered both that the harm was intentional and that the CEO didn't think it would occur.[12] The seeming robustness to different predicates for the elicited answers may be regarded as a confirmation that many subjects indeed consider the CEO as lacking belief.

Admittedly, in these few experiments cited as evidence that attributions of intentionality occur even in the absence of (attributions of) belief, the responses that support such view are not a strong majority. Yet, if the number of respondents that reveal a common pattern of responses is not insignificant (perhaps around 20% and above), that pattern of responses can be dismissed as biased or flawed only provisionally. When one in five respondents show some consistency in their responses, the conclusion that the pattern of the remaining four is 'the folk theory' may be hasty. Perhaps there are several folk theories, which may be more or less prevalent among a particular pool of respondents and for a specific vignette. So, the responses of a substantial minority of respondents probably deserve additional scrutiny. Hence, for the sake of my argument, it would have sufficed to show that a non-negligible number of laypersons treat the CEO in the modified chairman experiment as lacking a justified belief and yet find the harm intentional. That such is the view of the majority of respondents – though admittedly by a tiny margin – strengthens the point that 0-type actions deserve additional scrutiny.

3. Folk concept(s) of intentionality

The puzzle in the Knobe effect is that lay attributions of intentionality to an ASEA are influenced by the moral valence of such side-effect. This undermines our intuitive conviction that most people do (and should) first assess the intentionality of an action and only afterwards express their moral judgment (Mele and Sverdlik 1996).

Several explanations have been proposed for this asymmetry, which can be traced to two main strands. Some

[12] The difference in the responses to the first question (i.e., did the CEO intentionally harm the environment?) in the 'confident' and in the 'think' treatments is not significant: χ^2 (1, N=74) = 0.1953, p = 0.70. Also the difference in the responses to the second questions (i.e., was the CEO confident / did the CEO think that the environment would be harmed?) is not significant: χ^2 (1, N=74) = 0.0210, p = 0.90.

accounts propose that the folk concept of intentionality invokes evaluative notions (Knobe 2006), perhaps in a fashion that is reminiscent of the 'single phenomenon view' (Turner 2004, Hindriks 2008).[13] According to the supporters of these accounts, which may be collectively referred to as a *competence view*, the responses to the chairman vignettes correctly reflect the underlying folk concept.

According to the rival *performance view*, the folk concept of intentionality is free from evaluative notions, but then some factor (usually blame) interferes with the judgment. For instance, in a pragmatic interpretation of the Knobe effect, the lay use of intentionality may carry implicatures of blameworthiness (Adams and Steadman 2004); alternatively, the folk may have dissonant moral beliefs, so that their intuition suggests that some doing is unintentional, but some other intuition suggests that all blameworthy doings are intentional, and so the respondents correct their answers in order to resolve the inconsistency either consciously (Mele 2001) or unconsciously (Nado 2008); another account suggests that negative affect toward the agent, provoked by the blameworthiness of his act, can bias judgments of intentionality (Malle and Nelson 2003, Nadelhoffer 2004, Lanteri et al. 2008). The upshot of all these performance accounts is that most laypersons fail to consistently employ their concept of intentionality.[14]

With few exceptions, the accounts just mentioned share the goal of uncovering *the* principle on which folk intentionality is based. That there be only one such principle, however, is unwarranted. In a path breaking paper, Shaun Nichols and Joe Ulatowski (2007) suggest that the Knobe effect is the result of 'interpretative diversity,' or the fact that laypersons conceive of intentionality in a plurality of ways. They arranged a study in which subjects responded to both the HELP and the HARM scenarios of the original chairman experiment, instead of only one. They observed that only about a third of the participants revealed the Knobe effect and responded that harm was intentional, but help wasn't. About a third responded that harm and help were both intentional, and about a third that neither harm nor help was intentional. When asked to explain their answers, the participants who judged the side-effects intentional argued that the chairman knew in advance what would happen, while those who judged them not intentional explained that the chairman had no desire to bring about the side-effects. Nichols and Ulatowski thus conclude that intentionality can be interpreted in two different ways, as dependent on belief or desire.

Cushman and Mele (2008) reckon that this is a promising path to explore, but consider the standard chairman experiment insufficient for the task. The chairman has some belief about the side-effect of his decision and does not have a desire to bring it about. A complete investigation should map several scenarios (they employ a total of sixteen vignettes) in which the agent alternatively has or lacks the belief, the desire, or both to bring about some effect, which in turn is either morally loaded or neutral.[15] After extensive experimentation, they conclude that folk judgments of

---

[13] The single phenomenon view of intentionality is associated with Bratman (1987). Turner (2004) and Hindriks (2008) invoked it as an explanation of the Knobe Effect.

[14] This review was deliberately kept brief, and so incomplete. The competence/performance distinction is now conventional in the literature (e.g., Nichols and Ulatowski 2007, Nado 2008). See Feltz (2007) for a different classification and for a broader review.

[15] For the full text of the vignettes, see Cushman and Mele (2008, pp. 183ff.).

intentionality variously depend on desire and belief, and thus propose the following 'two-and-a-half' rules that laypersons seem to employ when attributing intentionality (pp. 182-3).

RULE D. An action is intentional if it is performed with desire, given the necessary background conditions.

As mentioned above, 'desire' is the agent's desire to perform the action either as an end or as a means to an end (p. 173).

RULE B. An action is intentional if it is performed with belief.

Again, 'belief' refers to the agent's justified belief that she will perform the action when she is "*very confident* that this is so" (p. 173). Only a minority of the people (roughly 20%) employ this rule, while many more employ the following instead:[16]

RULE MB. Belief is a sufficient condition for intentional action in and only in cases of morally bad actions.

It is not clear whether this latter counts as a (third) rule of its own, in a fashion akin to the competence accounts, or as a bias in the responses like one of the performance views. Cushman and Mele (p. 177) themselves struggle with the distinction, so they settle on a provisional middle-ground, whereby MB counts as half a rule – whence the label two-and-a-half rules theory, instead of a two or a three rules theory.

Instead, I suggest that *both* interpretations are correct and so I propose, as it were, a three-and-a-half rules theory.[17]

3.1. Three-and-a-half folk rules of intentionality

According to Cushman and Mele (2008, pp. 176-7), the asymmetry in the original chairman study can be explained by rule MB. Although the chairman lacks a desire to bring about either side-effect, he has beliefs that he will bring them about, at least in HARM. Laypersons thus respond that he brings about the harm intentionally, because environmental harm is morally bad. The side-effect in the HELP treatment, instead, is considered not intentional because it is not morally bad. Also in the modified chairman study presented above, there is an intentionality asymmetry in the responses. Yet, it cannot be explained by rule MB, because the chairman lacks belief also in HARM. So, Mele and Cushman's account cannot be directly employed to capture the findings of the modified vignette experiment.[18]

In the original experiment, the followers of rule MB may be consistently applying a concept of intentionality that includes a moral valence component. Perhaps they employ a theory akin to the single phenomenon view, according to which an action is performed intentionally if its agent disregards a normative reason against performing it. In the modified experiment, such a normative reason could be not to take environmental risks. Accordingly, many respondents (43%) blame the CEO even when the new production technique helps the environment. Yet, the majority of the

---

[16] Cushman and Mele (2008, pp. 182-3) recommend further testing of these rules and of the hypotheses about their prevalence among laypersons.

[17] A consistent investigation of whether MB should count as rule or as bias requires that there be a uniform opinion of what counts as moral bad. Clearly, such assumption can be granted only on occasion. Such uniformity will not be needed for my account below.

[18] In what follows I refer to the finding of Lanteri (2009), because they include both intentionality attributions and moral judgments.

respondents in HELP does *not* blame the CEO, which is against the predictions of the single phenomenon view. Moreover, the single phenomenon view cannot illuminate the moral judgment asymmetry between the two scenarios, since taking environmental risks is presumably blameworthy regardless of the valence of the outcome.[19] Though this competence interpretation of rule MB may explain some of the responses, it does not cover in full the observations in the modified chairman experiment.

Given that many laypersons do attribute intentionality even though belief and desire are absent, perhaps rule MB should be extended to non-belief cases. In other words, the sufficient condition for intentional action in cases of morally bad actions may not be belief, but moral bad itself.[20] More specifically, it could be the blameworthiness such moral bad entails. As noted above, in performance accounts blame is typically singled out as the element that distorts folk uses of the concept of intentionality. Such interpretation would support the two-rules theory of folk intentionality and this, as it were, 'extended rule MB' would not be listed as a rule proper, but as a bias.

If this explanation were sufficient to account for the evidence, we would observe a strong correspondence between the attributions of intentionality and the blame deserved by the chairman. However, we do not. While 79% of the respondents find the chairman blameworthy in the HARM scenario, only 58% find the harmful side-effect intentional. Though an extended rule MB might account for some of the responses, it is again not sufficient to fully address the new evidence.

Mele and Cushman suggest that about 80% of the laypersons employ rule MB. Although this figure should be further tested, it is a useful reference. In the modified chairman experiment, the CEO lacks desire to bring about the side-effect, so rule D does not apply and judgments of intentionality cannot be consequential from its application. He also lacks belief, so also rule B does not apply and cannot account for judgments of intentionality. As for rule MB, if it were a rule proper, and so not a bias, we would expect nobody to say that the CEO intentionally harmed the environment. Yet, many respondents judge the side-effect intentional. If, instead, (extended) rule MB reflected a bias of the kind suggested, we should observe that those who employ it find the harm intentional. We would thus expect roughly 80% of the respondents to say that the CEO had intentionally harmed the environment in the modified chairman study.[21] Yet, we observe a much lower figure (remind: 58%).

Both interpretations of rule MB are capable of explaining a portion of the evidence from the original and the modified chairman experiment, but not all. It would seem that a non-negligible number of participants give responses conforming to the competence-view, but also non-negligible is the number of participants whose responses conform to the performance-view. So, I propose to count rule MB as both a rule and a bias, or as 'one-rule-and-a-half.'

---

[19] For a more elaborate discussion of the single phenomenon view and the Knobe Effect, see Hindriks (2008) and my comment (Lanteri 2009).

[20] There is evidence (Knobe and Mendlow 2004, Sverdlik 2004, Phelan and Sarkissian 2008) that a bad outcome in itself does not explain asymmetric attributions of intentionality, but maybe a *morally* bad outcome does.

[21] This expectation rests on the assumption that judgments of blameworthiness operate as a reliable proxy for moral bad.

It is possible that the proportion of laypersons employing the biased extended rule MB be greater than those employing the belief-based standard rule MB. It is also likely that these proportions are subject to variations. For example, Cushman and Mele (2008, p. 176) found that the respondents who had previous experience responding to a series of (non-moral) vignettes correct their answers away from rule MB. Findings such as these seem to support the view that laypersons' responses are often susceptible to an array of influences, among which emotional variations play an especially prominent role. (In fact, this is arguably a central feature of *lay* judgments.)

Like Cushman and Mele's, mine, too, is a provisional suggestion. Admittedly, the figures obtained in the experiments may reflect an inherently imperfect methodology.[22] Even if this were not the case, they would benefit from validation from other studies.

4. Concluding remarks

In the experimental philosophical literature on folk accounts of intentionality, two broad types of explanations have been proposed for the observed responses of laypersons to vignette studies. According to the competence view, the responses correctly reflect the underlying folk concept. According to the performance view, the responses reflect a judgment bias. That only one of the two views is (and can be) correct seems a rather radical claim. That different respondents, respondents to different vignettes, or respondents to vignettes at different moments in time may or may not be influenced by some affective bias is a more modest, and to my mind more plausible, claim.

Cushman and Mele (2008) reckon that both views are plausible ways to describe the responses consistent with their rule MB. They therefore opt for a provisional compromise and count MB as half rule, with the implication that further research will tell whether the competence view or the performance view are correct, and so also whether MB should count either as a full rule or not as a rule. However, there is as of yet no basis to stipulate that only one of the two accounts can be correct. So, I hesitate to discard either account; though I cannot rule out the possibility that future evidence will force me to. In deciding between recording rule MB as a rule proper or as a bias, I am therefore inclined to saying that it can be alternatively either one, and in fact that it is both. Hence – provided that there exist no other folk rules that have so far escaped us, and that the two rules previously identified are indeed rules and not yet unknown biases – the counting should stop at three-and-a-half rules.[23]

---

[22] For critiques of experimental philosophy see: Sosa (2007), Kauppinen (2007), Liao (2008), and Cullen (2010). However imperfect, the methods of experimental philosophy will prove relatively more useful for investigating certain domains of philosophical inquiry: those in which the opinions of the folk matter the most, like moral and political philosophy (Mele 2003).

[23] This proposal may be more accurately regarded as a three-rules theory, with bias. I chose the 'three-and-a-half' label as a way of acknowledging the paternity of the original idea.

Acknowledgments

This article was written in part when the author was a Research Scholar at the TINT program, at the Faculty of Philosophy of the University of Helsinki (Finland). An earlier version of this paper has benefited from comments by Joshua Knobe and Alfred Mele. The author is grateful to an anonymous referee of *Philosophical Studies* for several insightful suggestions.

References

Adams, F., & Steadman, A. (2004). Intentional action in ordinary language: Core concept of pragmatic understanding. *Analysis,* 64, 173–181.

Beebe, J., & Buckwalter, W. (2010). The epistemic side-effect effect. *Mind and Language*, 25 (4): 474-498.

Bratman, M. (1987). *Intention, plans, and practical reason*. Cambridge, MA: Harvard University Press.

Cullen, S. (2010). Survey-driven romanticism. *Review of Philosophy and Psychology*, 1 (2): 275-296.

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

Feltz, A. (2007). The Knobe effect: A brief overview. *Journal of Mind and Behavior*, 28(4), 265-277.

Harman, G. (1976). Practical reasoning. *Review of Metaphisics*, 79, 431-463.

Hindriks, F. (2008). Intentional action and the praise-blame asymmetry. *Philosophical Quarterly*, 58(4), 630-641.

Kauppinen, A. (2007). The rise and fall of experimental philosophy. *Philosophical Explorations*, 10, 95-118.

Knobe, J. (2003). Intentional action and side-effects in ordinary language. *Analysis*, 63 (279), 190-193.

Knobe, J. (2004). Intention, Intentional action and moral consideration. *Analysis*, 64 (282), 181-187.

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

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

Knobe, J., & Nichols, S. (2008). An experimental philosophy manifesto. In: J. Knobe & S. Nichols (Eds.), *Experimental philosophy* (pp. 3-14). Oxford, UK: Oxford University Press.

Lanteri, A. (2009). Judgments of intentionality and moral worth: Experimental challenges to Hindriks. *Philosophical Quarterly*, 59 (237), 713-720.

Lanteri, A., Chelini, C., & Rizzello, S. (2008). An experimental investigation of emotions and reasoning in the trolley problem. *Journal of Business Ethics*, 83 (4), 789-804.

Liao, M. (2008). A defense of intuitions. *Philosophical Studies*, 140 (2), 247-262.

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

Malle, B. (2006). Intentionality, morality, and their relationship in human judgment. *Journal of Cognition and Culture*, 6 (1-2), 87–112.

Malle, B., & Nelson, S. (2003). Judging mens rea: The tension between folk concepts and legal concepts of intentionality. *Behavioral Sciences and the Law,* 21, 563-580.

McCann, H.J. (2005). Intentional action and intending: Recent empirical studies. *Philosophical Psychology*, 18 (6): 737-748.

Mele, A. (2001). Acting intentionally: Probing folk notions. In: B. Malle, L. Moses, & D. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 27-44). Cambridge, MA: MIT Press.

Mele, A. (2003). Intentional action: Controversies, data, and core hypotheses. *Philosophical Psychology*, 16, 325–340.

Mele, A., & Cushman, F. (2007). Intentional action, folk judgments, and stories: Sorting things out. *Midwest Studies in Philosophy*, 31, 184-201.

Mele, A., & Sverdlik, S. (1996). Intention, intentional action, and moral responsibility. *Philosophical Studies*, 82 (3), 265-287.

Nadelhoffer, T. (2004). Praise, side effects, and intentional action. *Journal of Theoretical and Philosophical Psychology*, 24 (2), 196-213.

Nadelhoffer, T. (2006a). Foresight, moral considerations, and intentional action. *Journal of Cognition and Culture*, 6 (1), 133-158.

Nadelhoffer, T. (2006b). Bad acts, blameworthy agents, and intentional actions: Some problems for jury impartiality. *Philosophical Explorations*, 9, 203-220.

Nadelhoffer, T., & Nahmias, E. (2007). The past and future of experimental philosophy. *Philosophical Explorations*, 10, 123-149.

Nado, J. (2008). Effects of moral cognition on judgments of intentionality. *British Journal for the Philosophy of Science*, 59 (4), 709-731.

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

Pellizzoni, S., Siegal M., & Surian, L. (2009). Foreknowledge, Caring, and the Side-Effect Effect in Young Children. *Developmental Psychology*, 45 (1): 289-295.

Pellizzoni, S., Girotto V., & Surian, L. (2010). Beliefs and moral valence affect intentionality attributions: The case of Side Effects. *Review of Philosophy and Psychology*, 1 (2): 201-209.

Phelan, M., & Sarkissian, H. (2008). The folk strikes back: Or, why you didn't do it intentionally, though it was bad and you knew it. *Philosophical Studies*, 138 (2), 291-298.

Sosa, E. (2007). Experimental philosophy and philosophical intuition. *Philosophical Studies*, 132 (1), 99-107.

Sverdlik, S. (2004). Intentionality and moral judgments in commonsense thought about action. *Journal of Theoretical and Philosophical Psychology*, 34, 224-236.

Swain, S., Alexander, J., & Weinberg, J. (2008). The instability of philosophical intuitions: Running hot and cold on Truetemp. *Philosophy and Phenomenological Research*, 76 (1): 138-155.

Turner, J. (2004). Folk intuitions, asymmetry, and intentional side effects. *Journal of Theoretical and Philosophical Psychology*, 24, 214-219.

Young, L., Cushman, F., Adolphs, R., Tranel, D., & Hauser, M. (2006). Does emotion mediate the effect of an action's moral status on its intentional status? Neuropsychological evidence. *Journal of Cognition and Culture*, 6 (1-2), 291-304.