abstracta Linguagem, Mente & Ação ISSN 1807-9792 Volume 8, Number 1, pp. 15–29, 2014

Reactive Commitments: Reasoning Dialectically about Responsibility[1]

David Botting

davidbotting33@yahoo.co.uk

Abstract

Philosophy has recently been presented with, and started to take seriously, sociological studies in which our 'folk concepts' are elaborated. The most interesting concepts studied are moral concepts, and results have been achieved that seem to sharply contradict the speculation of philosophers and to threaten the very way in which moral philosophy has been done in the past. In this paper, I consider these results and then sketch a version of a reactive attitude theory that allows for a genuine sense in which our intuitions about responsibility may be incoherent in a certain sense but without making moral reasoning radically contextual.

1 The Problem

1.1 The Data

In several studies, scenarios were described to people and they were asked whether the agent in the scenario was responsible for his actions. Judgments have been shown to be asymmetrical around several axes:

A. The abstract versus the concrete (Nelkin 2007, 247–48).
   When given specific details, respondents are more likely to find a person responsible, but if questions are given in abstract form then respondents are less likely to find a person responsible.

B. Moral status asymmetries (Nelkin 2007, 248–50).
   i) Emotion asymmetry
      When the act performed is bad, we accept the presence of high emotions in the agent as an explanatory and mitigating factor. The bad act seems to us worse if done calmly, but the good act is not judged differently depending on whether it is done on impulse or deliberation.
   ii) Side-effect asymmetry[2]
      When a side-effect is unintended but foreseen, then people will usually say that the agent is responsible for it when it is bad, but not when it is good.
   iii) Severity

[1]The author would like to acknowledge that funding for this paper was received from the FCT Portugal under grant awards "Argumentation, Communication and Context", PTDC/FIL-FIL/110117/2009 and "Is moral reasoning essentially dialogical?" SFRH/BPD/77687/2011.

[2]In this position Nelkin has "Intention and act asymmetry" that "When the act performed is bad, then the intention to perform it is usually held blameworthy. When the act performed is good, the tendency to praise the intention to perform it is less strong"; I will not discuss this here. This explains the difference between my labelling and Nelkin's.

> We judge acts more harshly when they have more harmful consequences than when they do not even when the act itself is the same. This means that a drunk driver who happens not to hit anything is the beneficiary of moral luck.

This data poses poses a threat to the whole way moral philosophy has been done so far. How has moral philosophy been done so far? It is widely agreed that it proceeds by *the method of cases*. What does this mean? What, according to the method of cases, is the relationship between the theory and the data? In the next section I will outline two approaches to moral philosophy – roughly, one that focusses on normative issues and one that focusses on descriptive issues – and try to show what follows from the data on these approaches.

1.2 The Relation of the Theory to the Data

On the first approach, moral philosophy can be seen as aiming at a theory or analysis of moral concepts. The explication of a moral concept may or may not involve a decision procedure. For instance, it does not follow necessarily from a utilitarian theory of 'the good' that agents can, or even should, try to calculate the net utility of all the possible consequences of their action; we would think there is actually something morally defective in an agent who, when his wife was drowning, tried to decide whether or not to save her by hedonic calculus or by wondering whether universalization of his maxim leads to a contradiction in conception. The theory is an account of what it is for such decisions to be correct and not a description of how such decisions are reached or what decision procedures should be used, although it does provide a norm for those procedures to aim at. This being so, it is possible that the decision procedure that should be used is one that does not, in the particular case in question, result in a decision that does, for example, maximize expected utility; rule-utilitarianism, for instance, may be the correct decision procedure even in cases where its application leads to sub-optimific results. Hence, there are two senses in which our intuitions about cases may be correct: they may result in the best outcome, or they may issue from the best procedure. This is important because what the data is really capturing – at least when we reason about whether to hold an agent responsible – is our decision procedures or, equivalently, our criteria for applying the concept. In consequence, we should expect some discrepancy between the data and the theory; futher argumentation is required to relate the data to the theory of responsibility, and this will be shown to rest on certain assumptions that the data will show to be questionable. However, we should not give this fact more than its due; it is not unreasonable to suppose that at least the embryo of a correct theory is discoverable in the data, yet this too will turn out to be problematic.

A theory of a moral concept will predict when something falls under that concept, e. g., when a scenario is fully described the concept will tell you whether the actor in that scenario is morally responsible. How are we to test such a theory? We use the *method of cases*: we test the scenario against the moral judgments of ourselves and others. Since whether that actor is responsible is not something observable, the prediction to be tested is not whether the actor is responsible but whether judges presented with the scenario will make a particular moral judgment, e. g., to hold the actor responsible, and this will depend on and reflect their decision procedures (arguably, simple intuition may qualify as such a procedure), and it is these that have a direct relation to the data. So, to be able to test the theory and indirectly to make the data relate to the theory we have to make the assumption that, most of the time at least, the moral intuitions of human subjects asked to judge the scenario will be correct.

However, this assumption has the further presupposition that intuitions are coherent amongst themselves, that subjects are not adversely influenced by non-formal features of the

scenarios such as high affect, yet the data shows that scenarios that are formally identical elicit varying intuitions and one plausible explanation why this is is that these non-formal features are selecting different psychological processes/criteria for holding responsible/decision procedures, not only across different subjects but, what is more to the point, within the same subject across formally identical cases. A second possible explanation is that literally different concepts of responsibility are being selected, each with a single set of criteria. A third is that we have a single concept of responsibility but one that is highly sensitive to contextual features. In most cases the second and third possibilities will be indistinguishable.

It is this third possibility that seems to be the focus of debate, but I will show that this may be the result of a misunderstanding and a confusion between this and the first possibility, between a normative and a descriptive bias – criteria that are constitutive of a concept (its analysans) are not necessarily those of its application, e. g., we may attribute 'good' to an action on the basis that it follows a certain utilitarian rule without that action satisfying all the necessary conditions of goodness as the theory defines it, and it is not a mistake to do so. It cannot be assumed that such a procedure is the wrong one to use simply on the grounds that in this particular case its result is not the one the theory defines as "correct" – the theory of responsibility tells you what is the correct decision, not necessarily what is the correct decision procedure. These judgments/intuitions may be the correct ones to have.

Even so, the problem is that whatever the theory and whatever our definitions of correctness are, they are formal, and if intuitions do not depend on application of a single criterion but on different criteria depending on non-formal features of the individual case (and this is one possible explanation of the asymmetries in the data), then the relation between data and theory breaks down; the data cannot tell us anything at all about the theory *including, in particular, that we have different concepts of/a variantist concept of responsibility*. We *could* be selecting different concepts on the basis of non-formal features and this might explain the data, but the data itself does not and cannot show this. This point is important in what follows.

What is this "descriptive bias"? It is to take the second approach, in which the theorist sees his task less as providing a conceptual analysis and more as providing a philosophical clarification of the folk concept of responsibility. This can be seen as a descriptive, naturalizing approach, and is allied with the first possible explanation mentioned above.[3] The philosopher taking this approach does not talk about reponsibility itself but about our *attributions* of responsibility, making the relevant question "Is there a single criterion for attributing responsibility?" However, keeping conceptual analysis at bay does not mean that we cannot include a normative aspect of epistemology or that the issue of correctness is simply ignored. Even a descriptive approach can have normative consequences, here an account of what is the correct procedure. Nelkin (2007, 246) seems to be taking this approach when she assumes:

> Fit Assumption – The criteria for moral responsibility attributions fit well with all (or most) of our ordinary judgments.

This refers only to moral responsibility attributions and not to being responsible; the *concept* of responsibility is not mentioned at all. It is this assumption that seems to be threatened by the view of Doris, Knobe and Woolfolk (2007) who interpret the data of our ordinary judgments as showing that they are not made according to a single criterion and therefore there cannot be a single *invariant* criterion for *ascriptions* of moral responsibility (a view they call *invariantism*);

[3]When Mele (2003, 334) describes his project as "to construct a viable theory about how agents produce their intentional actions, *as I* (and many philosophers of action, I believe) *conceived of intentional actions*" [italics original] he seems to be moving towards this approach, but in a few sentences this has been re-construed as conceptual analysis of a 'core' concept. It seems that Mele has not really entered into the spirit of this approach.

there are different criteria depending on (e. g., psychologically and non-rationally selected by) non-formal features, hence the asymmetrical responses to formally symmetrical cases. That there are different criteria for *ascribing* the concept of responsibility (the first possibility mentioned above, called by them *variantism*) does not entail (the second and third possibilities mentioned above) that there are different concepts of responsibility or that the concept of responsibility is variantist, i. e., contextual.

In attributing to them the latter view and then refuting it, Warmke makes a straw man of their position, taking it as an conceptual analysis of responsibility rather than an empirical thesis about responsibility attributions. Warmke (2010, 2) puts the assumption as:

> *Conservativist Assumption*: The conditions for being morally responsible for an action should accord with all (or most) of our ordinary judgments about the conditions under which an agent is morally responsible and we can discover these conditions by considering these ordinary judgments.

This differs from the Fit Assumption in two ways.

Firstly, despite acknowledging this as a methodological assumption it should be noted that he refers here to the conditions for *being* morally responsible, i. e., the concept. Then he objects validly that nothing follows about the concept of responsibility unless we assume firstly that the asymmetrical responses are *correct* in that agents are correctly held responsible, and secondly that (most) agents that are correctly held responsible would also be responsible (as, arguably, they would be in a reactive attitudes theory). This is a problem for his own Conservativist Assumption but not for the Fit Assumption.

Secondly, the Conservativist Assumption claims that the conditions for being responsible can be discovered in the data. Obviously, the Fit Assumption does not say anything about discovering conditions. However, the *method of cases* itself does license an inference from a case satisfying a set of formal conditions to another case satisfying the same set of conditions (this is what is meant by their being 'relevantly similar') in the following way: the common methodology of the theoretician has been to present cases, state their own ordinary judgment about those cases, and by abstracting away the specific details of the cases, sort out a formal set of conditions for responsibility that is considered to be an adequate criterion for all cases independent of context, conditions that set out what similarities are relevant to judgments of responsibility. However, the side-effect case presents scenarios that are said to be relevantly similar, and the concrete/abstract cases present exactly the same scenario but described differently. Why, then, having decided that the agent was not responsible in one of these cases is this classificatory judgment not transferred from the 'source' to the 'target'? The data seem to show that being judged to be relevantly similar is not a stable basis for any inference from one case to another, or to put it another way, moral intuitions are not sufficiently coherent to determine what differences are or are not relevant; they resist formalization. This seems to be the principal threat posed by the data to the theory: it is tacitly to give up on the *method of cases*, at least as a means of theorizing about responsibility.

That Doris, Knobe and Woolfolk (2007) need not be interpreted as talking about variantism with regard to the concept of responsibility (the third possibility) does not mean to say that it is not a viable point of view and, as shown above, it is a possible avenue for explaining the data. To show that the concept of responsibility involved is variantist requires showing that more than one pattern of responses is actually correct, and if only one one pattern of responses is actually correct then the concept is invariantist. I will call these *metaphysical variantism* and *metaphysical invariantism*; these metaphysical theses require, in each case, an error theory for

whatever patterns of responses are incorrect. Additionally, the metaphysical variantist would need to show that there is a single variantist concept rather than multiple invariantist concepts, such as responsibility as attributability and as accountability described by Watson (1996). But here it seems that the data helps the metaphysical variantist because the data is characterized by asymmetries, so for these asymmetries to be explained by multiple concepts it means that one and the same subject must shift from one concept to the other in considering one and the same scenario, which seems unlikely. In summary, both invariantism and variantism with regard to the concept of responsibility is consistent with variantism as Doris, Knobe and Woolfolk understand it which I will call *methodological variantism*. However, one moral we can draw from Warmke's critique is that even *methodological variantism* does not follow if all but one of the ascriptions are *performance errors*, defined as *mis*applications of a single criterion or *mal*functions of the same psychological process, perhaps caused by an affective bias.

Let me illustrate the possible strategies involved with an example. The side-effect asymmetry was discovered by eliciting folk intuitions on the following vignettes (Doris, Knobe and Woolfolk 2007, 193–94):

> The HARM scenario
> 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, 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 program."
> They started the new program. Sure enough, the environment was harmed.
>
> The HELP scenario
> 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 it will also help the environment."
> The chairman of the board answered, "I don't care at all about helping 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.

Subjects were asked whether the chairman was blameworthy in the HARM scenario and praiseworthy in the HELP scenario. Obviously the possible answers are:

i) No blame/no praise.
ii) Blame/praise.
iii) Blame/no praise.
iv) No blame/praise.

We can forget about (iv) since this is counter-intuitive and hardly anyone answers in this way.

According to Knobe, the scenarios are identical in all features relevant to the chairman's relation to his act. Therefore, on the invariantist assumption of a single criterion we should get a symmetrical response, that is to say, either (i) or (ii). On the further assumption that one of the conditions for ascription of responsibility for an act is that the agent performed the act intentionally and in the vignettes the act is only a side-effect, the expected response is (i). However, most people by far give the asymmetrical response (iii). However, what is not always noted is that this does not actually support the claim that ascriptions of responsibility are variantist, because the same criterion that includes the act's being intentional seems

to have been applied in both cases; subjects do not take themselves to be in violation of the intentionality condition since they are also prepared to say that the chairman harmed the environment intentionally but not that he helped the environment intentionally. Here, there is a single criterion for ascriptions of responsibility but, arguably, not for ascriptions of intentionality since in both cases the chairman had identical psychological states and so whether or not the action was performed intentionally should give a symmetrical response, but in fact the intentional/unintentional asymmetry mirrored the praise/blame asymmetry.

So, if we take the intentional/unintentional asymmetry as the prior explanandum, and we can explain it without referring to responsibility, then methodological invariantism with respect to responsibility follows, or at least the data do not provide any evidence against it. This preserves the traditional idea that only acts performed intentionally are blameworthy and that the intentionality judgment is prior – the question then is whether there is invariantism (methodological or metaphysical) with respect to intentionality; the data suggest that there is not. In contrast, Knobe takes the praise/blame asymmetry as the prior explanandum, explains it through variantism, and concludes that our responsibility judgments in some way influence our intentionality judgments and that the variance in the latter is explained by the variance in the former. This would imply an objectionable circularity when construed as saying that the concept of responsibility referred to intentionality and the concept of intentionality referred to responsibility, but construed only as saying that we can infer whether an action is intentional if we know whether the agent is responsible for it and vice versa, then circularity only occurs if these are parts of *a single chain of reasoning*. That there are some occasions when we infer from intentionality to responsibility and other occasions (other chains of reasoning) when we infer from responsibility to intentionality is perfectly acceptable and orthogonal to the issue of whether there is a single criterion for ascribing responsibility or intentionality.[4]

What does this mean for the *metaphysical invariantist* who does have a concept of responsibility and is trying to test his concept against the data? If he takes the intentionality condition as part of his concept then one might initially think that he is committed to (i) and must explain away (ii) and (iii) as errors. However, as I have shown above (iii) need not be an error if he modifies his concept of intentionality, and then it is (i) and (ii) that need to be explained away, perhaps by variantism with respect to intentionality or more simply as performance errors.

[4]This is most clearly the case when one inference is deductive and the other (weaker) inference is abductive. In itself, this is not circular, and provided that what is being inferred from is itself justified there is no error (unless you consider abductive reasoning erroneous) in drawing these inferences. The kind of circularity or error that I have in mind is where the interderivability of two propositions is erroneously taken to entail that they are both true or both justified, as would occur if the judgment of responsibility for an action is made *unjustifiedly*, an inference (itself justified) made that the action was performed intentionally, and then an inference made (also itself justified) back to the judgment of responsibility, as if by travelling in this circle something unjustified has or could become justified.

It is not entirely clear what Knobe means by the claim that ascriptions of responsibility and intentionality affect each other. Both he and Wright and Bengson (2009) seem to regard it as problematic but this seems to me to be either a mistake or they are arguing for a variantist concept after all. In the chairman case it could be argued (but I am not saying that anyone does) that some subjects use variantist criteria of intentionality and then make inferences about responsibility, and that other subjects use variantist criteria of responsibility and then make inferences about intentionality. Here we would have variantist criteria of application of both concepts, but in any particular case the directionality of the inference is one way; it is not that a judgment about responsibility is made on the basis of a judgment about intentionality itself made on the basis of a judgment about responsibility. Which out of

i) (responsibility → intentionality) exclusive-OR (intentionality → responsibility)
ii) (responsibility → intentionality → responsibility)
iii) (intentionality → responsibility → intentionality)

is being argued for? While (ii) and (iii) involve circularity, (i) does not when interpreted as concerning ascriptions.

This shows that the data as such is not inconsistent with invariantism of either variety. The threat posed by the data is actually different, and this is the assumption, explicit in the Conservativist Assumption but also implicit in the method of cases as such, that there is a coherent set of intuitions to begin with from which we can *discover* in ordinary judgments the conditions for responsibility or for ascribing responsibility, or in other words, that the folk concept can be the basis of our theorizing, and if this is the case then it is difficult to know how the theorist can get started. The cost of *variantism* then is to make the discovery of standards of correctness in ordinary judgments impossible, undermining firstly the otherwise plausible assumption that most of our ordinary judgments are correct and secondly an analogical form of reasoning that is used in moral and especially in legal contexts.

My claim will be that once we understand the nature of moral reasoning we will understand better the patterns of moral ascriptions they lead to. I accept the assumptions that only one pattern (I choose the asymmetrical pattern) of responses is correct and, by adopting a version of the reactive attitudes theory, that agents that are correctly held responsible (which is most of those held responsible) would also be responsible. The data is eliciting reactive attitudes and is to be explained by looking at the kind of reasoning from which those attitudes issue. My claim – and the essence of my position – is that the nature of moral reasoning is dialectical; it concerns the exchange and evaluation of reasons. This will be the subject of the next section.

2 The Solution

2.1 The Theory

The reactive commitment theory is the reactive attitude theory dialectified. A commitment is attributed to a speaker by his audience when the speaker makes the linguistic performance of an assertion. This is fair because in taking what is spoken as an assertion the audience must presuppose that all the conditions of satisfaction of the speech act of asserting have been satisfied, in particular the sincerity condition that requires that the speaker believe the propositional content of his utterance. A commitment is not, however, the same as a belief; if the utterance is *mis*taken as an assertion, for instance if the speaker is acting deceptively, then a commitment is still attributable to the speaker. Thus, a commitment can be thought of as being like a contract between speaker and audience and is entered into in the cases of both genuine assertions and its *misfires*.

In this way, discussions where reasons are given in support of a proposition p are modelled dialectically in terms of generating commitments and the speaker trying to show that commitment to p is a consequence of other commitments shared by speaker and audience. This is basically the process of proving, e. g., by natural deduction, a conclusion from given premises, but construed dialogically as a series of assertions of lemmas and questionings of those lemmas until commitments are appealed to that cannot be denied by the questioner (the audience) without violating their own contractual obligation (their shared commitments). The speaker then wins the discussion by appealing to the dialectical rule called the *closure rule* that prevents introduction of or appeal to commitments that are not shared and/or continuing to hold a commitment set that has been shown to be logically inconsistent. On the other hand, if p is not shown to follow from the shared commitments then the audience appeals to the *closure rule* to win the discussion. Such appeals are not often modelled as dialectical moves in themselves, but my view is that they should be seen also as generating a commitment, but that this is a *reactive commitment*. The reactive commitment can be seen as a permission (which may or may not be exercised) to impose sanctions for breaking a contractual obligation. Unlike the

kind of commitment discussed first where a commitment is generated by the mere performance of an assertion, a reactive commitment is only generated if the closure rule is used *correctly*.

This is linked to the reactive attitudes theory in the following way: an agent is responsible if there is a practice of *praise* or *blame* connected to his action. This is summed up in the phrase "an actor is responsible if she is held responsible." A moral sanction is analyzable as a speech act of blame.[5] To be a genuine blaming and not a misfire certain conditions must be satisfied, three of which are of paramount importance: the blamer must have *permission* to blame (there is a reactive commitment), the blamer must have *power* to blame (it must be physically possible that what is permitted can be done – a person who makes a 'threat' that he has no means to carry out does not really perform the speech act of threatening), and the blamer must not take the person blamed to be *morally unlucky*. It is *not* a condition that the blamer has the reactive attitude associated with blame, although he must have *some* reactive attitude, e. g., towards the discussion itself. A would-be blamer who has the reactive attitude associated with blame and desires to blame but lacks, for instance, the power to blame, cannot really blame. His attempts to blame would misfire – he would be merely letting off steam, 'playing' at blame. To answer "Yes" to the question "Is the chairman blameworthy?" in a questionnaire is not the same as blaming. Similarly, if he takes the agent to be morally unlucky but justifies blaming them on instrumental/pragmatic grounds (e. g., as a social deterrent) then he does not really blame even if, perhaps, this is the correct decision procedure. In such situations the agent is not responsible. However, the fact that there is a reactive commitment and yet the agent is not blamed does not necessarily mean that the agent is not responsible. There are cases where the conditions are met and yet the reasoner does not blame because of personal reasons such as the feeling that he is in no position to judge. It should be noted here that the reactive commitment makes blame permissible and not obligatory, or perhaps it would be better to say that the obligation to blame generated is one that can be defeated. In this situation the agent *is* responsible.

An agent is negatively responsible if open to blame and positively responsible if open to praise, or to put it another way, if the dialectical rules governing the kind of dialogue where an agent is obliged to justify his actions to a questioner *could* result in blame or praise respectively. This is not, of course, to suggest that such a dialogue actually takes place. The claim is rather that a reasoner reflecting on what reactive attitude to take, if any, reasons as if such a discussion was taking place. Moral reasoning is inherently dialogical.

The concept of responsibility given above is invariantist. It differs from most invariantist concepts because it considers the perspective of the judge as well as the actor, which allows for different judges to correctly give different responsibility ascriptions to the same actor. How this deals with the data of the first section will be the subject of the next section.

2.2 Accommodating the Theory to the Data

Explaining away the data has become something of a cottage industry, generating an impressive number of hypotheses which, even more impressively, have been found to have empirical support by their proponents. Hypotheses vary along the following axes:

1. Prior explanandum
    - Intentionality/unintentionality asymmetry is explanatorily prior (e. g., Hindriks)

[5]Note that the questioner cannot, in general, be sanctioned if the speaker wins the discussion, since the questioner who only questions commitments (in other words one who questions *p* without arguing for *not-p*) has not violated any rules unless they continue to question *p* after the discussion has been won. It is the one being questioned who has most of the obligations, e. g., the obligation to defend *p*. The dialectical rules govern how the obligations are divided and met in the course of the discussion.

- Praise/blame asymmetry is explanatorily prior (e. g., early Knobe)
- Some other asymmetry of which the praise/blame asymmetry is an epiphenomenon is explanatorily prior (e. g., Machery)

2. Internalist/externalist
   - Satisfiable by facts about the agent only (in the spirit of traditional accounts)
   - Not satisfiable by facts about the agent only

First, let us look at Knobe's position. He has since abandoned the position that the moral goodness or badness of the side-effect selects different psychological processes (a variantist position with regard to ascriptions of responsibility) and seems to have opted for conceptual revision of intentionality to include as a normative component whether some action is morally good or bad. This is an *externalist* criterion and seems to have a variantist concept of intentionality as its corollary. It is not a corollary of the superficially similar *internalist* criterion of whether the actor takes that thing to be good or bad (a psychological fact) as given, for example, in Hindriks' (2008, 638) *invariantist* account of intentional action where *S A*-s intentionally when *A* is a foreseen side-effect of *B* and *S B*-s in spite of the fact that he believes his expecting *A*-ing constitutes a normative reason against *B*-ing, and in Machery (2008) who treats the moral status as an epiphenomenon and has as his internalist criterion whether the side-effect is perceived by the actor as a cost. Considerations of cost (harming the environment is perceived by the actor as a cost but helping the environment is not a cost) would play the same role and produce the same results as Hindriks' account, without referring to anything normative such as moral badness.

Is there a case that can decide between hypotheses that take the actor's viewpoint and those that take the subject's viewpoint? Note that in the HARM condition the chairman, or at least his board, do recognize the moral badness of harming the environment, since in "it will increase our profits *but* harm the environment" the 'but' generates the implicature that what follows it is, from the speaker's point of view, a *contra* reason opposed to the *pro* reason preceding the 'but'. The alternative hypotheses seem to make the same predictions for this case.[6]

But consider the following:

> A terrorist discovers that someone has planted a bomb in a nightclub. There are lots of Americans … who will be injured or killed if the bomb goes off. The terrorist says to himself, "Whoever planted that bomb in the nightclub did a good thing. Americans are evil! The world will be a better place when more of them are injured or dead."
>
> Later, the terrorist discovers that his only son, whom he loves dearly, is in the nightclub as well. If the bomb goes off, his son will certainly be injured or killed. The terrorist then says to himself, "The only way I can save my son is to defuse the bomb. But if I defuse the bomb, I'll be saving those evil Americans as well. … What should I do?" After carefully considering the matter, he thinks to himself, "I know it is wrong to save Americans, but I can't rescue my son without saving those Americans as well. I guess I'll just have to defuse the bomb."
>
> He defuses the bomb, and all of the Americans are saved (Knobe 2007, 99-100).

Faced with this vignette, most subjects say that the terrorist *did not* save the Americans "intentionally" but that his saving the Americans can be explained by his reasons.

[6]Machery's hypothesis makes different predictions when the difference between the conditions is non-moral. The original studies did not show the asymmetry in non-moral cases, but these results have been contested (Machery 2008).

According to Knobe, this case creates problems for the externalist hypothesis that the *subject's* judgment of the badness of the action influences the ascription of intentionality. His reasoning seems to be that ascriptions of intentionality depend on the same psychological factors as ascriptions of reasons to the agent unless the outcome is morally bad, yet here the outcome is morally good and yet subjects are inclined to say that the agent's actions can be explained in terms of the reasons described but not that he saved the Americans intentionally. In other words, where the outcome is morally good (as perceived by subjects) then ascriptions of intentionality should track ascription of reason explanation. Thus, Knobe considers this to be a counter-example to his own hypothesis. I think that the premise Knobe relies on here is too simplistic; there can be reasons explanations without intentions. For instance, I intend to write a paper on experimental philosophy, and my having this intention explains why I wrote this paper and why there is such a paper before you. Does it explain why I wrote a *long* paper? This is a contrast question I find difficult to answer: I certainly did not set out with the intention of writing a long paper – it just turned out that way. But even if the reasons explanation of why I wrote a paper does not explain *fully* why I wrote a long paper, it does not seem to me that the explanation is false. When we *A* we bring about a host of effects and this bringing about can accurately be described as our actions *X, Y,* and Z. A full explanation of our *A*-ing is still a partial (arguably a so-called *straight*) explanation of our *X*-ing. So, I find it questionable whether this is really a counter-example: the terrorist has a reasons explanation for saving the Americans because he has a reasons explanation for saving his son and knew that he could not perform the one action without the other.

In contrast, this case does seem to work against hypotheses where it is how the actor perceives the moral status that is the question. Given that the actor perceived saving the Americans as a bad thing or as a cost, then subjects should say that the actor saved the Americans intentionally (consistently with the chairman in the HARM scenario) but they do not. This case does seem to decide in favor of taking the subject's point of view and variantism.

The result seems to be the expected one on Wright and Bengson's approach: given that we have judged the outcome as good, we should not judge it to have been performed intentionally or otherwise we would be forced to infer that the agent is positively responsible, i. e., praiseworthy. And we obviously do not want to consider the agent in this case as praiseworthy. They give the following formula (Wright and Bengson 2009, 27):

Good/bad action + intentionality = positively/negatively responsible actor

The formula can be used as a mathematical formula would, that is to say, it can be solved for whichever term is unknown. If the action has been judged as bad and as performed intentionally then the actor's being negatively responsible can be calculated by using the formula from left to right. On the other hand, if it is the 'intentionality' term that is missing and awaiting judgment then we can apply the formula to calculate it if we have the judgments on the action and on responsibility (and analogously when the missing judgment is whether the action is good or bad). The inference from the known values to the unknown value is non-deductive and seems to vary depending on which side of the '/' applies to the particular case. If the action is good and the actor is positively responsible (i. e., open to praise) then it can be inferred (defeasibly) that it was performed intentionally, and when the action is bad and the actor is negatively responsible (i. e., open to blame) it can be inferred (arguably more weakly) that it was performed intentionally.

Wright and Bengson exploit a perceived asymmetry in praise and blame that explains the praise/blame asymmetry. This is common to many different accounts of different types. Ac-

cording to McCann what we are blaming in the HARM condition is the chairman's attitude. This attitude violates a Kantian perfect duty not to harm the environment, whereas the corollary duty of helping the environment is an imperfect duty over which we have certain freedom to attend to as and when we wish. He sums up the chairman's attitude to harming the environment by saying that the chairman "means it" (McCann 2005). Similarly, Nadelhoffer stresses that what we praise or blame is in the first instance the agent and that "insofar as subjects judge that an *agent* is blameworthy, they are more inclined to say that any *negative* side effects brought about by the agent are intentional and any *positive* side effects brought about by the agent are not intentional" (Nadelhoffer 2004, 180); in both the HARM and HELP conditions the chairman is morally reprehensible because he does not take the moral considerations as motivating reasons for or against his action and this explains the intentional/unintentional asymmetry. If the agent were morally praiseworthy, says Nadelhoffer, then positive side-effects would be said to have been brought about intentionally, and this was in fact supported by empirical evidence.

The reactive commitment theory gives the same result. Assuming that the subjects in the study thought that saving Americans is a good thing, the question is whether the agent is praiseworthy in bringing it about, and this depends on what reasons the agent can give. For praise to apply those reasons must promote the moral values contained in the shared commitments by appeal to propositions contained in the shared commitments. Now, the subject would presumably consider the agent's saving his son to promote shared moral values and so praise him for that, but as for saving the Americans the best that the agent can sincerely say is that it was a side-effect of his saving his son. Because of the way that the scenario is presented the subject knows also that it was an undesired side-effect. So, whether the agent finds the side-effect undesirable (as in this case) or he simply doesn't care (as in the case of the chairman) he loses the discussion; he is not properly motivated. To put it another way, the agent was morally lucky from the subject's point of view that bad reasons led to a good result.

What does it say with regard to whether the chairman is blameworthy for harming the environment? Again, the agent cannot defend himself with the right kind of reasons and loses the discussion. But couldn't the chairman claim that he was morally *un*lucky? Subjects who give the "no praise/no blame" response probably reason something like this. But as Feltz and Cokely (2009a, 345) point out, what is often important from the group's point of view is social harmony, and this allows for an instrumental value of reactive attitudes and a looser interpretation of moral luck. Such subjects will tend to take foreknowledge as sufficient (given other conditions) for ascriptions of intentionality and respond also that the agent is blameworthy – they will give the asymmetrical "no praise/blame" and "unintentional/intentional" response. This could, in fact, be the correct decision procedure. But if the agent really is morally unlucky then although the subject would have the reactive attitude associated with blame and say in their questionnaires that the chairman was responsible this would not correspond to any reactive commitment. In the case as described, it is not clear whether the chairman is or is not the beneficiary of moral luck, but we do not need to decide this metaphysical issue in order to explain the pattern of responses: some such decisions will be correct, others will not.

What kind of explanatory hypothesis is this? It takes the praise/blame asymmetry as its explanandum and gives a single *invariantist* set of conditions for attributions of responsibility. Variation is explained by the fact that subjects may reason from different commitment sets and also have different ideas of what may defeat their obligation to sanction. It depends on facts about the judge as well as on facts about the agent, but – functionalized by the notion of a discussion governed by dialectical rules – this does not seem to lead to any problematic form of variantism.

On the connection between ascriptions of responsibility and intentionality I find Wright and Bengson's account of the inference between responsibility and intentionality attractive, and probably this can be also be given a dialectical turn. For instance, the closure rule could generate a commitment that the agent acted intentionally. A commitment, it has already been said, is not a belief, so it is not implied that either of the speakers believe that the agent acted intentionally. However, realizing that they have this commitment they may feel that it is necessary, when faced with the question, to indicate some kind of endorsement. This is a defeasible obligation and is overridden if you take having the intention, as opposed to mere foreknowledge, to be a necessary condition of ascribing intentionality. To put it another way, it could be a commitment of a type that can be retracted without sanction.

We must now consider the rest of the data. I will state my hypothesis before going through each asymmetry in turn. We have already seen in the case of the side-effect asymmetry that the man who foresees certain good or bad side-effects, but does not count those effects as reasons for acting, is lucky if those side-effects are good. My claim now is that moral luck is the common thread through all of the data.

A: The deterministic agent whose causal history is composed of only good acts is lucky.
B(i): The person who acts from emotion is lucky if his act turns out to be a good one (and presumably, that his emotion is a 'positive' one).
B(iii): The drunk motorist who gets home without hitting anything is lucky.

The issue of moral luck links all of these asymmetries. Asking a subject to praise such an act is to ask them to make a performative contradiction; it is a fallacy of many questions in that it demands a direct answer when its presupposition is not satisfied, to apply a concept that is not applicable. What we judge in such cases is not the act but, by reasoning dialogically about his or her reasons, the agent.

First, (A) the abstract/concrete asymmetry. Nahmias et al have done the experiment of describing the same deterministic background and varying whether an action is described as caused by his intentional states or by his neurophysical states, notwithstanding the fact that each are equally determined. They discovered that respondents were ready to attribute responsibility where psychological language was used but not when physical language was used. This seems to show that people are generally amenable to compatibilism between determinism and responsibility, contrary to the claims of incompatibilists, but not to the compatibilism between mechanism (the reduction of the description to a physicalist language) and responsibility. When examined more closely the folk conception of responsibility is not incompatibilist, it is argued, but only seems to be (Nahmias, Coates, and Kvaran 2007). Against this, Knobe argued that the folk conception of responsibility was incompatibilist and that what the studies (backed up with studies of his own) showed was that when presented 'concrete' cases subjects tended to exhibit compatibilist tendencies, but that when presented abstract cases subjects tended to exhibit incompatibilist tendencies. This in turn was explained by the fact that concrete cases elicited high affect which biased the psychological process; compatibilist intuitions were performance errors. But Nelkin (2007, 255–56) turns this around:

> With no other information given, people tend to assimilate determinism to coercion, but this suggestion is concealed when an intentional action is described in concrete terms. Determinism is also sometimes assimilated in people's minds to reductionism. But determinism does not preclude intentional, rational action. The concrete case avoids these faulty assimilations.

These assimilations being faulty, it is incompatibilist intuitions that are performance errors.

I will lay my cards on the table: I am an incompatibilist.[7] Granted that we may only feel ourselves coerced when it is another agent who prevents us from acting according to our will, or perhaps when there is a physical impediment to our so acting, this seems to me only an empirical fact about what we have to consider in everyday contexts and not a conceptual truth about responsibility. In the philosophical context I see no difference between our intentions being brought about by events out of our control and physical movements being brought about in the same way.

Scenarios described in physicalist language are simply not the kind of things about which making moral judgments is sensible or useful; "not responsible" in the abstract scenario here should be taken as a paraphrase of "responsibility is not an applicable concept in this scenario." In other words, asking for a moral judgment is at best infelicitous and at worst fallacious. For the same scenario described in psychological language, the agent can provide reasons and excuses of the type that the psychological description at least suggests and for that reason a dialogue where they are discussed can be imagined to take place because of which we may have a reactive attitude. But because there is no reactive commitment corresponding to this reactive attitude, this intuition (that the agent is responsible in the deterministic but non-reductionist scenario) is false.

On B(i) the emotion asymmetry Nelkin (2007, 252) claims that negative feelings can interfere with the reasons-responsive mechanism and cause us to fail to recognize the right reasons. Positive feelings, e. g., empathy, may in fact help us to see the right reasons. I find it oddly optimistic that so-called negative feelings can only hinder an accurate appraisal but positive feelings do not, and Nelkin does not seem to provide an argument for this claim. I am inclined to see this as a cultural bias – some groups may accept high emotion as a mitigating factor while others may not. This is one of those things that cannot be decided beforehand by a conceptual analysis but only emerges from the actual dialogue and what the commitment sets allow for.

The bad act seems to us worse if done calmly because the agent fails to appeal to the right values in giving reasons, and this is worse in the case where the agent has reasons and is not behaving non-rationally. On the other hand, if the presence of high emotion is among the shared commitments as an acceptable mitigating factor, then the agent may present a successful defence. Or, because of empathy, or because the agent has shown remorse and/or is unlikely to repeat the transgression, or the agent has performed the speech act of apologizing, or because the reasoner does not think he is in a position to judge (thus defeating his obligation to sanction) nobody actually blames the person even though the person is and sometimes is even explicitly recognized to be responsible.

On B(iii) the severity asymmetry Nelkin (2007, 254) comments that there cannot be moral luck. Intuitions that there must be – that attempted murder is not as blameworthy as murder – are due to under-description of the cases. If the situation is the same, the intuition should be the same. The asymmetry with regard to the negligent motorist could also be due to confusing degrees of responsibility with degrees of compensation. Thus, Nelkin seems to take the asymmetrical response to be incorrect in such a scenario.

But the negligent motorist case seems to differ from the chairman case in the following respects alone: (i) the chairman knew that the environment *would* be harmed, while the motorist knew only that some accident *might* occur, and (ii) the chairman explicitly considered, and excluded from his motives, the harm that would be caused, while the motorist might not

[7] I am also an introvert. If Feltz and Cokely (2009a; 2009b) are right there is a strong correlation between these things. In saying that the assimilation of determinism to coercion is faulty and attributing the intuitions to the fact that this is made clear in the non-reductionist scenario, Nelkin seems to be taking a compatibilist view.

have considered at the time he parked the car that an accident might result although he knew this in a *dispositional* sense. Now, the severity of the accident affects (i) since the worse the outcome the less likely you should think it able to occur. This is basically decision theory where sometimes you take a small chance because of a correspondingly large reward, or refuse to take a chance because, although you will probably win, the consequences of losing are severe. The more severe the side-effect, the more the motorist case is like the chairman case. However, in the motorist case the side-effect was not, we may suppose, foreseen, but was foreseeable. Obviously, it cannot be a moral requirement to foresee all the consequences of our actions. Whether foreseeability is considered as enough to avoid being morally unlucky depends, as before, on the extent to which social considerations play a part and the extent to which the motorist satisfices his duty to take precautions. Consideration of the latter may occur in two ways: it may be the case that the judge does not find the motorist's behaviour sanctionable and that having the brakelines checked would have been supererogatory, or it may be that the judge does find it sanctionable but defeats the sanction in virtue of some defeating obligation such as not believing himself to have the standing to sanction the motorist, as would probably occur if the subject behaved in a similar manner. On the other hand, if there are strict rules about when brakelines should be checked that the motorist has not observed then he will be held responsible.

3 Conclusion

An agent is *responsible for* his action if the group he is *responsible to* hold him responsible (in the sense described). An agent performs the action *intentionally* either if it is intended or if it is the foreseen side-effect of an action that is intended.[8] The empirical data can be explained by the interplay of reasons, obligations, and defeaters of obligations as regulated by dialectical rules. Attributions of blameworthiness made to agents in deterministic scenarios will be incorrect, but it is perhaps misleading to say on the contrary that they are not-blameworthy, i. e., the predicate-negation. It is better to say that blameworthiness is simply inapplicable in such cases, but there is no space to argue this point here. There is variantism in so far as attributions of intentionality can in some cases be made on the basis of attributions of responsibility as shown by Wright and Bengson, and there are different criteria for good actions and bad actions. This explains how we get from the no praise/blame response to the unintentional/intentional response. But this is a rather toothless variantism that does not lead to incoherence or to any objectionable form of circularity. Any thought that it does confuses claims about concepts with claims about their criteria of application.

Outside of conditions where the concept of responsibility is applicable, our reactive attitudes (unsurprisingly) do not imply anything about responsibility, but this does not, of course, mean that we simply cease to have them. Reactive attitudes arise out of our emotional, interpersonal interactions and cannot be reproduced by theoretical reasoning, but when the scenarios are described in concrete and psychological terms we can model them dialectically and thereby reason about them. This will tell us what reactive attitude we 'ought' to have. Our intuitions about responsibility (the 'ordinary judgments' of the Fit Assumption) are largely coherent within a particular dialogue or commitment set, but can be incoherent across dialogues or

[8] This seems to me the most perspicuous way of expressing the Single Phenomenon view. For discussion see Bratman (1984), Mele (2003) and Nadelhoffer (2006).

If this definition of "intentionally" is correct then the judgment that the terrorist did not save the Americans intentionally is incorrect – he surely did. However, the inference that he is not because he is not praiseworthy, as described by Wright and Bengson, still applies because it concerns the ascription of the term "intentionally" and not its analysis.

commitment sets. It is too quick to conclude from the empirical data that these intuitions correspond to distinct psychological processes.

References

Bratman, M. (1984), 'Two faces of intention', *The Philosophical Review* **93** (3), 375–405.

Doris, J., Knobe, J. & Woolfolk, R. L. (2007), 'Variantism about responsibility' *Philosophical Perspectives* **2**, 183–214.

Feltz, A. & Cokely, E. T. (2009a), 'Do judgments about freedom and responsibility depend on who you are? Personality differences in intuitions about compatibilism and incompatibilism', *Consciousness and Cognition* **18**, 342–350.

Feltz, A. & Cokely, E. T. (2009b), 'Individual differences, judgment biases, and theory-of-mind: Deconstructing the intentional action side effect asymmetry', *Journal of Research in Personality* **43**, 18–24.

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

Knobe, J. (2007), 'Reason explanation in folk psychology', *Midwest Studies in Philosophy* **31**, 90–106.

McCann, H. (2005), 'Intentional action and intending: recent empirical studies', *Philosophical Psychology* **18** (6), 737–748.

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

Mele, A. (2003), 'Intentional action: controversies, data, and core hypotheses', *Philosophical Psychology* **16** (2), 325–340.

Nadelhoffer, T. (2004), 'On praise, side-effect, and folk ascriptions of intentionality', *Journal of Theoretical and Philosophical Psychology* **24** (2), 196–213.

Nadelhoffer, T. (2006) 'On Trying to Save the Simple View', *Mind & Language* **21** (5), 565–586.

Nahmias, E., Coates, D. J. & Kvaran, T. (2007), 'Free will, responsibility and mechanism: experiments on folk intuitions', *Midwest Studies in Philosophy* **31** (1), 214–242.

Nelkin, D. (2007), 'Do we have a coherent set of intuitions about responsibility?', *Midwest Studies in Philosophy* **31** (1), 243–259.

Watson, G. (1996), 'Two faces of responsibility', *Philosophical Topics* **24**, 227–48.

Wright, J. C. & Bengson, J. (2009), 'Asymmetries in judgments of responsibility and intentional action', *Mind & Language* **24** (1), 24–50.