The Baltic International Yearbook of
Cognition, Logic and Communication

October 2012 Volume 7: *Morality and the Cognitive Sciences*
pages 1-18 DOI: 10.4148/biyclc.v7i0.1774

Stefano Cossara
Paris-Sorbonne University

Cognitive Science, Moral Responsibility and the Self

A Reply To Knobe And Nichols

Abstract: In their "Free Will and the Bounds of the Self", Knobe and Nichols try to get at the root of the discomfort that people feel when confronted with the picture of the mind that characterizes contemporary cognitive science in order to establish whether such discomfort is warranted or not. Their conclusion is that people's puzzlement cannot be dismissed as a product of confusion, for it stems from some fundamental aspects of their conception of the self. In this paper I suggest, contrary to their conclusion, that there is a sense in which the skeptical worries about responsibility elicited by the computer model of the mind do result from confusion. Those worries can be traced back to an irrational over-generalization concerning the scope of cognitive science and the alleged exhaustiveness of the range of facts formulated in its vocabulary.

1. INTRODUCTION

In their recent contribution 'Free Will and the Bounds of the Self' (2011), Joshua Knobe and Shaun Nichols try to account for the difficulty that ordinary people seem to encounter in ascribing responsibility to agents whose behavior is described in terms of the standard cognitive science model of the mind. The conclusion they draw is interesting and apparently unhappy: this difficulty is not the result of a confusion, of a superficial muddle to be clarified via conceptual analysis. On the contrary, it derives from stable and fundamental aspects of the folk understanding of the self. It seems that, as a consequence of this unfavorable state of affairs, we are bound to remain prisoners of a paradox: on the one hand we are convinced that we are the causes of our actions, on the other the correctness of the computer model of the mind seems to entail that this conviction is fallacious.

In this paper I will contend that the idea that the computer model of the mind entails that we are not the genuine source of our actions is indeed the result of a confusion, though of a different kind from the one hinted at by Knobe and Nichols. In order to make this point, in section 3 I will criticize the thesis, endorsed by Nichols and Knobe, that the skeptical doubts about responsibility elicited by cognitive science do stem from fundamental aspects of the ordinary understanding of the self. Knobe and Nichols reach this conclusion by showing that skeptical doubts are a product of the same stable psychological structures that implement ordinary ascriptions of responsibility; I will suggest that this fact cannot ground any conclusion concerning whether or not they are confused. Moreover, in section 4 I will maintain that recourse to theoretical constructs such as psychological structures is unnecessary in this case, for an inquiry into the way that words are used in ordinary practice is sufficient to dispel the sense that some ordinary ascriptions of responsibility are bizarre or puzzling. Skeptical worries, by contrast, bring with themselves a more robust sense of puzzlement, but they ensue from an irrational over-generalization concerning the scope of cognitive science and the exhaustiveness of its vocabulary. Therefore, they can be dismissed as the product of confusion; or so I will argue in section 5. In section 6 I will sum up and draw some conclusions. In order to make all of this intelligible, however, it is necessary to have a clear picture of Knobe and Nichols's approach to the issue. Providing that picture will be the aim of the next section.

2. KNOBE AND NICHOLS ON RESPONSIBILITY AND THE SELF

The point of departure of Knobe and Nichols's research is an examination of the two main families of philosophical accounts of free will and moral responsibility. One is centered on the distinction between de-

terminism and indeterminism, and on claims concerning the laws that govern our universe. Philosophers such as Peter van Inwagen (1983) have argued that free will and moral responsibility may be impossible in a world governed by deterministic laws. However, Knobe and Nichols are skeptical that arguments from this family actually get at the root of people's intuitive worry about free will. And as they are interested in people's ordinary views, they turn to the second major family of philosophical accounts. This second family is concerned with the self and with the worry that it might turn out not to be the genuine source of human action. The problem here is not determinism: agents may fail to be the source of their actions not because they are determined, but because they are not, in some relevant sense, the source of their actions.[1] An example may help to clarify this point: let us suppose that it is discovered that John's actions are entirely determined by the states of his brain. People may experience the worry that John is not responsible for his actions because if his brain controls all of his actions, it is not really John who decides what to do. In other words, people's worry may be that an agent's self is causally idle, and that it has no significant impact on the agent's actions. This worry concerns the threat of epiphenomenalism rather than determinism, and this line of thought is supported by a certain amount of research in experimental philosophy (Nahmias 2006; Nahmias et al. 2007; Nahmias & Murray 2011).

Why do people experience similar concerns? A classical line of response is that people's worries concerning epiphenomenalism are due to confusion. What 'confusion' means in this case will soon become clear. For the moment, suffice it to say that this is exactly the line of thought that Knobe and Nichols are going to challenge. Their hypothesis is that people's intuitions are not merely the result of confusion, but reflect something "deep and fundamental" (Knobe & Nichols 2011, p.536) about the concepts they employ, and more notably about the way they ordinarily think about the self. The core of their work consists in the empirical validation of this hypothesis. They start by examining the different conceptions of the self developed in the history of philosophy. They identify three main conceptions: the bodily conception, the psychological conception and the executive conception of the self. Each conception differs from the others in what it takes to count

as the self.[2] According to the bodily conception, the self is constituted by the agent's body, i.e. by everything that is contained in the skin. According to the psychological conception, only an agent's mental states and psychological processes—her memories, convictions, aspirations, etc.—constitute her self. According to the executive conception, the self is a kind of commanding-faculty that stands over and above the specific mental states and psychological processes one may have. As should be evident, each conception is more restrictive than the previous one as to what counts as the self. When one adopts a bodily conception, any action performed by an agent's body can be read as having the agent as its legitimate source. If one adopts a psychological conception, however, the agent can be taken to be the genuine source of her actions only if those actions originate from her mental states. If one adopts an executive conception, the agent is the legitimate source of those actions that directly originate from her 'core' self, and not from her body or her mental states. Knobe and Nichols are convinced that, far from being pristine inventions of philosophers, those conceptions represent important strands of commonsense thinking about the self. They argue that people are not committed to one single conception of the self. Rather, they tend to shift from one conception to the other depending on the way in which they approach the problem in specific cases. This entails that people do not have any fixed view about what lies inside or outside the self, nor any fixed sense of the relationship between the agents, their bodies and their mental states.

As we will see, people's judgments and concerns about responsibility seem to be connected to the kind of conception of the self they adopt. But what determines the kind of conception they adopt in specific cases? According to Knobe and Nichols, an important factor is the kind of perspective that people take in observing a specific situation. When people adopt a zoomed-in perspective and focus on the details of a certain action, they will adopt in most cases an executive conception of the self. This entails that they take an agent's self to be confined to something that stands over and above her body and her mental states; and they ascribe responsibility to the agent for her actions only if they can trace those actions back to such a separate self, independent from her body and mental states. On the contrary, people are more promiscuous in assessing what comprises the self, and

more liberal in ascribing responsibility,[3] when they adopt a zoomed-out perspective. When they observe an agent acting within a broader context, they typically ascribe responsibility to the agent even though her actions are described as stemming from her mental states.

The bulk of Knobe and Nichols's work consists in providing empirical evidence for the hypothesis just sketched: when people adopt a zoomed-out perspective, they take an agent's mental states (and more notably her emotions) to be a part of the agent's self, and hence they ascribe responsibility to the agent for the outcomes of her actions even when those actions are determined by her emotions. On the contrary, when they take a zoomed-in perspective and consider an agent's behavior in isolation, they adopt a conception according to which the agent's emotions do not count as a part of her self, and thus conclude that she is not responsible for the outcomes of actions stemming from her emotions.

In the four experimental studies that constitute the core of their work, Knobe and Nichols present participants with vignettes describing an agent that accomplishes simple actions, and ask them whether the agent is responsible or not for her actions and their outcomes. In the first study, half of the subjects are assigned to a 'choice-cause' condition, whereas the other half are assigned to an 'emotion-cause' condition. Subjects in the former condition receive the following scenario (Knobe & Nichols 2011, p. 544):

> Suppose John's eye blinks rapidly because he wants to send a signal to a friend across the room.

Subjects in the latter condition receive the following scenario (ibid.):

> Suppose John's eye blinks rapidly because he is so startled and upset.

Subjects in both conditions are asked to express their agreement on a scale from 1 ("disagree") to 7 ("agree") with the following statement (ibid.):

> John caused his eye to blink.

As should be clear, both scenarios prompt people to zoom-in and therefore to adopt an executive conception of the self, according to which

John's emotions do not count as a part of John himself. As a consequence, most participants in the choice-cause condition tend to agree that John caused his eye to blink, whereas most people in the emotion-cause condition tend to disagree with that statement.

Study 2 aims to show that the same result obtains with mental states other than emotions. All subjects are presented with the following scenario (ibid., p. 545):

> John's hand trembled because he thought about asking his boss for a promotion.

Half of the subjects are asked whether

> John caused his hand to tremble.

The rest of the subjects are asked whether

> John's thoughts caused his hand to tremble.

As expected, subjects tend to agree with the statement that John's thoughts caused his hand to tremble, but not with the statement that John caused his hand to tremble. Knobe and Nichols take this result to suggest that people adopting the executive conception take an agent's thoughts, and not only her emotions, to be something distinct from the agent's self. Study 3 presents a scenario in which John has a disease in the nerves of his arm. He experiences a sudden spasm, his arm twitches, and his hand ends up pushing a glass off the table. The glass strikes the floor causing a loud crashing noise. Subjects in the zoomed-in condition are asked whether they agree or disagree with the sentence (ibid., p. 546):

> John caused his arm to twitch.

Subjects in the zoomed-out condition are asked whether they agree or not with the sentence

> John caused the loud noise (ibid., p. 547).

Subjects in the zoom-in condition tend to disagree with the claim that John caused his arm to twitch, whereas subjects in the zoomed-out condition tend to agree with the claim that John caused the loud noise.

Once more, according to Knobe and Nichols people adopt a thinner conception of the self in the zoomed-in condition, which leads them to conclude that John has not caused the twitching. On the contrary, the thicker conception of the self they adopt in the zoomed-out condition leads them to conclude that John caused the noise. According to Knobe and Nichols, the divergence between the two responses must be traced back to the fact that there are two different 'objects' that are identified as 'John' in the two cases: in the zoomed-out condition 'John' includes John's bodily processes, in the zoomed-in condition he does not.

In study 4 Knobe and Nichols examine all possible combinations of the choice/emotion variable and the zoom-in/zoom-out variable. In the choice-cause version of the zoomed-in scenario, participants are presented with the following vignette:

> A bee lands next to John and his hand withdraws.
>
> Now suppose you learn that John's hand withdrew because he is afraid of bees (ibid., p. 548).

In the emotion-cause version of the zoomed-in scenario participants receive a similar vignette:

> A bee lands next to John and his hand trembles.
>
> Now suppose you learn that John's hand trembled because he is afraid of bees (ibid.).

In both cases participants are asked to say whether they agree that John caused his hand to move. The setting is parallel in the zoomed-out cases, except that John's movement knocks over a glass of milk. As expected, subjects in the zoomed-out condition tend to say that John is responsible for the outcome in any case, whereas subjects in the zoomed-in condition tend to say that John causes his hand to move only in the choice-condition.

The aforementioned studies support a theory that predicts that people will have problems ascribing responsibility when taking a zoomed-in perspective on behaviors stemming from an agent's mental states and psychological processes. Knobe and Nichols use this theory in order to explain why cognitive science seems to represent a threat to free will. By means of a further experiment, they show that people's ordinary conception of the human mind is significantly different from the

computer model that characterizes cognitive science. More notably, people tend to think that a human agent can perform an action even though all of her mental states tell her to do otherwise, but that a computer cannot perform an action if all of its software tells it to do otherwise. Thus, people's ordinary understanding of human actions seems to involve a separate self, something distinct from mental states and psychological processes. This also explains why people find typical models in cognitive science disquieting and confusing when they are brought to bear on attributions of responsibility. Those models describe human behavior in detail, thus prompting the adoption of a zoomed-in perspective and of an executive conception of the self. At the same time, such descriptions do not mention anything beyond mental states and psychological processes, and leave no room for anything like a separate self. When adopting this perspective, people feel that all of those states and processes fall outside the bounds of the self, and end up thinking that the self has no impact on human action, and that agents cannot be morally responsible for their conduct.

3. RESPONSIBILITY AND CONFUSION

Because Knobe and Nichols's argument unfolds through a number of empirical studies, it is not always easy to follow. Let us take stock. They remark (ibid., p. 552) that

> it has been a recurring theme in philosophy that a complete scientific explanation for human action would exclude the possibility of free will. An old and persistent line of response to this worry is that it stems from a confusion.

Their goal is to show that this line of response is incorrect. But what does the term 'confusion' specifically mean in this case? The best place to look for an answer to this question is the kind of account that Knobe and Nichols put forward as an alternative to the confusion hypothesis. This is an account in which failure to ascribe causal responsibility to agents in ordinary practice is explained in terms of 'fundamental' aspects of people's understanding of the self. They suggest that the idea that the skeptical doubts about responsibility that people experience when confronted with the computer model of the mind are a result of

confusion is not compatible with the fact that those doubts seem to be connected with stable, fundamental features of the ordinary practice of ascribing responsibility. Knobe and Nichols seem to be suggesting that one might take seriously Dennett's (1984) idea that people's failure to identify the self with a set of psychological states and cognitive processes is a result of confusion only if such a failure comprises a kind of exception to the rule embodied in regular ascriptions of responsibility. But Knobe and Nichols believe that their studies show that there is no way to differentiate the exception from the rule in these cases: people's 'negative'[4] ascriptions of responsibility in the ordinary context originate from the same stable psychological structures that are the source of positive ascriptions. Moreover, those very psychological structures (in particular the different conceptions of the self which commonsense seems to include) are the source of the hyperbolic skeptical worries about responsibility elicited by the computer model of the mind. As a consequence, it is not possible to identify any real asymmetry between those worries and usual ascriptions of responsibility. But in the absence of any asymmetry, saying that the former are the result of confusion seems to be unwarranted.

I think this is a pretty faithful way of explaining the (not entirely explicit) line of argument endorsed by Knobe and Nichols against the idea of confusion. It must be clear, however, that this line of argument takes as its premises some claims that are in no way watertight. To get started, it must be noted that the notion of confusion employed by Knobe and Nichols has a strong normative component. Not accidentally, at the beginning of their article Knobe and Nichols declare that their aim is "to get at the sources of this discomfort and thereby gain some insight into whether or not it is warranted." (ibid., p. 531). If the claim that the correctness of the computer model of the mind entails that we are never responsible for our actions were shown to be a product of confusion, it could be dismissed as lacking epistemic warrant. But if the notion of confusion is thereby normatively-laden, it is not clear that the psychological treatment suggested by Knobe and Nichols can provide the grounds to reach the kind of conclusion they mean to reach. Knobe and Nichols seem to maintain that the fact that two claims are implemented by what, within a psychological theory, can be recognized as structures that are equally stable entails the impossibility of recognizing either claim as the product of confusion. But is it really so? Maybe whether a claim is or is not the product of confusion depends on factors other than the psychological structures that implement it. Moreover, it seems quite clear that the degree of epistemic warrant possessed by a claim (to which, as already said, the notion of confusion seems to be related) *certainly* depends on factors other than the cognitive structures devoted to its implementation, at least when those structures are identified by a psychological theory in terms of their being stable or, as Knobe and Nichols prefer to say, 'fundamental'. To say the contrary would be tantamount to maintaining that all the claims implemented by equally stable cognitive structures have the same degree of warrant, which looks quite implausible.

Thus, it is not evident that tracing ascriptions of responsibility back to the psychological structures that implement them has much to say about their being or not being the result of confusion, at least not if confusion is read in the normatively-laden sense that is relevant to Knobe and Nichols's project. Why then should one pursue this route? Nichols and Knobe seem to suggest that there is no other way to make sense of ordinary ascriptions of responsibility: postulating theoretical entities such as cognitive structures is necessary to account for our ordinary practice of ascribing responsibility. They say that ordinary ascriptions of responsibility are sometimes 'puzzling', and that they "have not been able to come up with any alternative hypothesis that can explain the full pattern of intuitions revealed in these studies." (ibid., p. 549).

However, it seems to me that ordinary ascriptions of responsibility may only look puzzling if one fails to attend to the specific and sometimes idiosyncratic features of our ordinary conceptual practice. This is a risk that the inquiry conducted by Knobe and Nichols certainly runs, guided as it is by the norms of elegance and simplicity that are typical of scientific theorizing. In the next section, however, I will show that it is possible to make sense of our ordinary practice of ascribing responsibility without relying on psychological theory. I will conclude that, beyond not being very suited for the normative task of assessing whether people's discomfort with cognitive science is or is not warranted, theory construction is in this case also unnecessary, for nothing is really puzzling once the proper elucidation has been achieved.[5]

4. ORDINARY PRACTICE AND THE UNNECESSARINESS OF THEORY

The aim of the four studies presented in Knobe and Nichols's paper is to assess people's intuitions concerning responsibility as elicited by different vignettes, with the purpose of gaining insight into the underlying psychological structures. The descriptions thereby presented to the experimental subjects are quite short and focused on ordinary life settings. It seems to me that these features make it possible to read the subjects' replies as providing insight into their ways of using words, rather than into hidden features of their psychology. In this section I will present this alternative, 'open to view' interpretation of the data provided by Nichols and Knobe. Just because our ordinary practice is open to view, my interpretation will add nothing to what we already know about it; in a sense, it will be a kind of reminder of things we already know. The point I want to argue for is that there is nothing particularly puzzling in this practice; the feeling of puzzlement that Knobe and Nichols seem to experience when confronted with some of their subjects' replies fades away once proper attention is devoted to the particular ways in which the words are used.

In study 1 subjects are asked whether John causes his eye to blink. People tend to respond affirmatively when John's eye blinks in order to send a signal to a friend (choice-cause case), negatively when his eye blinks because he is startled and upset (emotion-cause case). Knobe and Nichols do not explicitly state that responses to this case are puzzling. Nevertheless, they think the responses require a theoretical explanation: on their view, participants in the emotion-cause condition tend not to ascribe causal responsibility to John because they adopt an executive conception of the self, according to which John's emotions do not count as a part of John. However, this result can be explained more simply by making appeal to the fact that, even if in both the choice-cause case and in the emotion-cause case people are asked to express their agreement about whether John causes his eye to blink, 'blinking' refers to two different actions in the two cases. In the former, it refers to a voluntary action (winking at a friend), in the latter to an involuntary bodily movement (an eye blinking because of a high level of emotional activation). It seems plausible and natural that people may distinguish between the two cases, and identify the agent (John) as the cause of a voluntary action but not of an involuntary bodily movement. Thus, there is nothing particularly strange in the difference between the two.

Knobe and Nichols explicitly label people's intuitions as 'puzzling' in study 3. The scenario describes John as having a disease in the nerves of his arm. He experiences a sudden spasm, his arm twitches, and his hand ends up pushing a glass off the table. The glass strikes the floor causing a loud crashing noise. Participants in the zoomed-in condition tend to say that John did not cause his arm to twitch, whereas participants in the zoomed-out condition tend to say that John did cause the loud noise. The result is apparently puzzling because it is the twitching that causes the loud noise; thus it seems that people take John to be the cause of a distant outcome, and not of the more proximate outcome that causes the more distant one. However, the puzzlement dissolves once it is recognized that the verb 'to cause' is used differently in the two contexts, i.e. when it refers to an agent's bodily processes or to events that are external with respect to an agent's body. It would sound strange to say that a person causes involuntary bodily processes such as her heartbeat or her breathing; similarly, it would be odd to say that a person has caused an *involuntary* contraction of her muscles. At the same time, it is natural to say that a person caused a *voluntary* contraction of her muscles. Thus, it seems that voluntariness makes a difference to the use of the verb 'to cause' with respect to bodily processes. The same is not true in regard to events that are external with respect to an agent's body. In this case, the fact that some outcomes are brought about *involuntarily* does not prevent us from saying that an agent caused them. It does not sound odd, for example, to say that I have accidentally caused the glass to fall. To resume: it seems that people usually say that an agent has caused her bodily processes only when those processes correspond to voluntary actions; on the contrary, the agent is typically said to cause events that are external with respect to her body, regardless of her bringing them about in a voluntary or an involuntary fashion. As a consequence, it is not surprising that John can be said to (involuntarily) cause (cause*) a loud noise by pushing a glass off the table, even though he pushes the glass off the table because of the (involuntary) twitching of his arm, which he does not cause (cause**).[6]

In study 4 Knobe and Nichols seem to identify two apparent sources of puzzlement. Participants in the zoomed-in condition are asked whether

John caused his hand to move as a consequence of a bee landing next to him. Among those subjects, participants in the choice-cause condition respond affirmatively, participants in the emotion-cause condition respond negatively. This result seems to puzzle Knobe and Nichols, for "John performs exactly the same behavior in the two cases" (ibid., p. 548). But does John really perform the *same* behavior? The former scenario describes his hand as *withdrawing*, the latter as *trembling*. Thus, in the former case the movement seems to be part of a voluntary action, which John chooses to perform because he is afraid of bees. In the latter, the movement is involuntary and directly determined by John's fear. As in study 1, people take John to cause a voluntary action but not an involuntary bodily movement; there is nothing particularly puzzling in the difference between the two.

Participants in the zoomed-out condition are asked whether John caused the milk to spill as a consequence of a bee's landing next to him. In this case, subjects tend to respond affirmatively regardless of whether they are assigned to the choice-cause condition (where John's hand withdraws) or to the emotion-cause condition (where John's hand trembles). As in study 3, subjects in the zoomed-in condition are asked whether John causes a bodily movement, whereas subjects in the zoomed-out condition are asked whether he causes an event that is external to his body. As we have seen in the discussion of study 3, the verb 'to cause' is used differently in the two cases: an agent is said to cause external events, which come about as consequences of her bodily movements (be they voluntary or involuntary). On the contrary, she is said to cause bodily movements or processes only in case they comprise voluntary actions. Once it is recognized that the same word is used differently in different contexts, the sense that people's responses to similar scenarios are puzzling (and that the only way to make sense of them is by appeal to a psychological theory) starts to fade away.

5. COGNITIVE SCIENCE, RESPONSIBILITY AND SKEPTICAL WORRIES

While there is nothing particularly puzzling in failures to ascribe responsibility to agents in ordinary practice, it is certainly puzzling that people may sometimes be driven to think that agents are *never* the genuine sources of their actions, and can never be morally responsible for them. According to Knobe and Nichols, this is what happens when people are confronted with the computer model of the mind. In this case, people may find themselves thinking something like:

> "If the mind actually does work like that, it seems like we could never truly be morally responsible for anything we did. After all, we would never be free to choose any behavior other than the one we actually performed. Our behaviors would just follow inevitably from certain facts about the configuration of the states and processes within us." (ibid., p. 530)

In Knobe and Nichols's view, the problem here is that models in cognitive science typically present detailed accounts of agents' behaviors described in terms of mental states and psychological processes; descriptions of that kind promote the adoption of a zoomed-in perspective, and therefore of an executive conception of the self, which allows for ascriptions of responsibility only when a separate self can be identified over and above mental states and psychological processes. However, typical cognitive science accounts do not leave room for such a self. Indeed, they mention no agent at all, but only mental states and psychological processes.

But why should people conclude that responsibility is impossible, starting from the premise that it cannot be ascribed within standard cognitive science accounts? On the face of it, it is not easy to see how the premise that no agent is mentioned in those accounts should lead to the conclusion that no agent is ever responsible for her actions. That premise might at best entail that responsibility cannot be ascribed when adopting the vocabulary of cognitive science. But that vocabulary is in no way the only one available. Why do people not simply infer that cognitive science is not the right place to ascribe responsibility?[7] Indeed, it seems that in order to step from the premise that no agent is mentioned within standard cognitive science accounts to the conclusion that no agent is ever responsible for her actions, one should add one further premise, such as "All the genuine facts about agency must be expressed or expressible in the vocabulary of cognitive science". Only if one accepts this further premise will she infer the general impossibility of responsibility from the absence of agents in cognitive science models of the mind. However, this additional premise

clearly looks like an over-generalization: why should one think that the only genuine facts about an agent are those that can be expressed in the jargon of cognitive science? It seems hard to believe that such a generalization may be reasonable. Indeed, it looks quite irrational.

This seems to be a local version of the over-generalization that grounds the commitment to the general thesis of naturalism, according to which the only facts in the world must be natural facts. Here is Paul Horwich's (2010, p. 157) 'diagnostic' account of the origin and apparent appeal of naturalism:

> a) Naturalism rests on the impression that any non-natural facts would be intolerably weird.
>
> b) That impression stems from a combination of three factors: first, the singular practical and explanatory importance of naturalistic facts; second, the very broad scope of the naturalistic—the striking range and diversity of the facts that it demonstrably encompasses; and third, the feeling that reality must 'surely' be fundamentally uniform—so all facts must be naturalistic.
>
> c) This final feeling is based upon a misguided overextension of scientific norms: in particular, the norm of theoretical simplicity. For it is pretty clear (i) that the metaphysical and epistemological variety of possible facts corresponds exactly to the variety of possible meanings (i.e. of possible regularities of word-use); (ii) that the latter will certainly include many that are *non*-naturalistic; and (iii) that many of those will be socially useful and will therefore be deployed.
>
> In order to undercut the sense of 'weirdness' that can stem from our failure to naturalistically 'locate' a given phenomenon it suffices to acknowledge the evident plausibility of this diagnosis.

Similarly, it seems that only people who think that there are no facts concerning an agent beyond those expressible in the vocabulary of cognitive science might think that the impossibility of ascribing responsibility in that vocabulary entails the non-existence of responsibility. But this conclusion is based on an irrational over-generalisation. And it

seems sensible to say that a person who is drawn to generalise in the absence of reasons to do so is indeed confused. Thus, it seems that, *pace* Knobe and Nichols, the discomfort that people experience when confronted with the computer model of the mind is indeed the unwarranted product of confusion. People are confused not, as Dennett (1984) thinks, because they do not recognize that an agent is identical to the sum of her mental states and psychological processes. They are confused when their tendency to over-generalise leads them to endorse the idea that the only genuine facts concerning the agent are those expressible in the vocabulary of cognitive science. But once it is recognized that facts concerning an agent (and her being responsible for her actions) can be found beyond the scope of cognitive science, it becomes easy to get rid of the sense of puzzlement that the computer model of the mind might initially arouse.

6. CONCLUSION

Knobe & Nichols (2011) suggest that the discomfort that ordinary people experience when confronted with the computer model of the mind cannot be the result of confusion, for it stems from fundamental features of the ordinary conception of the self. I have maintained that their defining those features as fundamental by appeal to theoretical constructs such as psychological structures prevents them from drawing conclusions on the issue of confusion, taken in the normatively-laden sense that is relevant to the current discussion. I have also maintained, *contra* Knobe and Nichols, that the very appeal to psychological constructs is unnecessary in this case, for the apparent sense of puzzlement that arises out of some ordinary ascriptions of responsibility can be dispelled by examining the specific conditions for the application of words within the ordinary discourse. I have suggested that there is more discontinuity than continuity between ordinary practice and the sceptical worries that sometimes arise out of accounts of human action framed in the vocabulary of cognitive science. Those worries seem to be the product of an irrational over-generalisation, according to which the vocabulary of cognitive science should suffice to express all the facts concerning an agent. Thus, I have argued that there is a sense in which those worries are the product of confusion.

Knobe and Nichols want to examine the threat to responsibility and free will represented by a complete scientific explanation of human action. In a sense, this very sentence contains the seeds of confusion that might foment our worries. For of course we do want a complete scientific explanation of human behaviour, i.e. one that includes all the relevant causes that help to predict and explain it. But it is hard to see how a similar explanation might jeopardise responsibility, unless one reads 'complete' as implying that there should be a vocabulary in which we can formulate an explanation that is complete in the sense of exhausting the totality of things to be said about agents. It is when we step from the former to the latter meaning of 'complete' that we end up thinking that cognitive science makes responsibility impossible. But we should not be too concerned with this worry, for it is evidently unwarranted. Once we recognise that it is the product of confusion, the threat it represents loses much of its grip.

Notes

[1] Even though determinism may sometimes serve to make this second sort of worry salient.

[2] It seems that Knobe and Nichols take the self to be by definition the source of an agent's actions.

[3] Causal responsibility, which is typically taken to be a necessary condition for moral responsibility.

[4] By the term 'negative' ascriptions I refer to judgments to the effect that an agent is not responsible for her actions. I call 'positive' ascriptions judgments to the effect that an agent is responsible for her actions.

[5] What follows is inspired by some strands of thought in Wittgenstein (1953/2001) and more specifically by the deflationary reading of Wittgenstein provided by Paul Horwich (2005; 2010).

[6] It should be noted, however, that in this case John will plausibly be recognized as causally, but not morally responsible for breaking the glass and producing the loud noise.

[7] It seems to me that people *do* indeed make this inference, at least in some cases. But for the sake of argument I will take it that Knobe and Nichols are right in describing the folk worry.

References

Dennett, D. 1984. *Elbow Room: The Varieties of Free Will Worth Wanting*. Cambridge: the MIT Press.

Horwich, P. 2005. 'Wittgenstein's meta-philosophical development'. *From a Deflationary Point of View* 159–171.

——. 2010. 'Rorty's Wittgenstein'. In A. Ahmed (ed.) 'Wittgenstein's Philosophical Investigations: A Critical Guide', 145. Cambridge: Cambridge University Press.

Knobe, J. & Nichols, S. 2011. 'Free Will and the Bounds of the Self'. In R. Kane (ed.) 'The Oxford Handbook of Free Will, 2nd Edition', 530–554. Oxford: Oxford University Press.

Nahmias, E. 2006. 'Folk Fears about Freedom and Responsibility: Determinism vs. Reductionism'. *Journal of Cognition and Culture* 6: 215–237.

Nahmias, E. & Murray, D. 2011. 'Experimental Philosophy on Free Will: An Error Theory for Incompatibilist Intuitions'. In J. Aguilar, A. Buckareff & K. Frankish (eds.) 'New Waves in Philosophy of Action', 189–216. London & New York: Palgrave-Macmillan.

Nahmias, E., Coates, J. & Kvaran, T. 2007. 'Folk Fears about Freedom and Responsibility: Determinism vs. Reductionism'. *Midwest Studies in Philosophy* 31: 214–242.

Nichols, S. & Knobe, J. 2007. 'Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions'. *Noûs* 41: 663–685.

Van Inwagen, P. 1983. *An Essay on Free Will*. Oxford: Oxford University Press.

Wittgenstein, L. 1953/2001. *Philosophical Investigations*. Blackwell Publishing.