Forthcoming in *Journal of Cognition and Culture*

Experimental Philosophy and Folk Concepts: Methodological Considerations[1]

Joshua Knobe and Arudra Burra

A reply to comments from Alfred Mele, Fred Adams, Gilbert Harman, Adam Morton, Liane Young, Fiery Cushman, Ralph Adolphs, Daniel Tranel & Marc Hauser, and Charles Kalish.

Experimental philosophy is a comparatively new field of research, and it is only natural that many of the key methodological questions have not even been asked, much less answered. In responding to the comments of our critics, we therefore find ourselves brushing up against difficult questions about the aims and techniques of our whole enterprise. We will do our best to address these issues here, but the field is progressing at a rapid clip, and we suspect that it will be possible to provide more adequate answers a few years down the line.

1. First, we need to get clear about what it is to provide a theory of a folk concept. **Alfred Mele** offers the following suggestion:

> By an analysis of a concept of X, I mean a statement of individually necessary and jointly sufficient conditions of a thing's being an X.

The problem with such an account is that it seems to say nothing about people's *concepts*. (It would tell us, not about people's concepts, but about the actual properties in the world that these concepts pick out.[2]) We take it that a person's concept is a particular type of mental representation. Hence, a theory about people's concepts must be a theory about particular types of mental representations that people possess.

To see the difference between a theory about properties in the world and a theory about people's concepts, one need only consider a definition like the following:

> A triangle is a polygon whose number of sides is equal to the cube root of 27.

This definition shows that it is possible to give necessary and sufficient conditions for being a triangle in terms of cube roots, but it does not thereby prove that people's *concept* of cube roots plays any role in their *concept* of triangles.

If we wanted to know whether people's concept of cube roots actually did play a role in people's concept of triangles, we could take into account a number of different types of considerations. One approach would be to actually conduct experimental studies. (For example, we could use priming studies to check to see whether hearing the

---

[1] We are greatly indebted to Stephen Stich both for numerous conversations on the topics discussed here and for his pioneering philosophical work on the philosophical importance (or lack thereof) of people's ordinary intuitions.

[2] Of course, these two types of inquiry often turn out to be mutually supportive. Mele's own work in action theory is a case in point. Although it was originally offered as part of an attempt to understand certain properties in the world, it has subsequently proved enormously fruitful and influential in research on people's concepts.

word ‘triangle’ activated people’s concept of cube roots.) Another approach, though, would be simply to look at the definition itself. When we see that the mention of cube roots serves only to make the definition more complicated, we get at least some prima facie reason to suppose that cube roots don’t actually figure in any way in the relevant concept.

Much the same could be said about the relationship between people’s concept of intention and their concept of intentional action. Mele has suggested that it might be possible to create an extremely complex analysis of intentional action that draws in an essential way on the concept of intention. For example:

> *PA2. S* intentionally *A*-s if and only if *S A*-s and either (a) *S*’s *A*-ing is caused in the right way by an intention to *A* or (b) *S* performs some action *B* that is caused in the right way by an intention whose plan component represents *S*’s *A*-ing as a goal relative to *S*’s intended *B*-ing and *S*’s *B*-ing appropriately generates *S*’s *A*-ing

But even if this analysis does yield necessary and sufficient conditions for intentional action, it does not thereby prove that the *concept* of intention plays any role in the *concept* of intentional action. Indeed, the sheer complexity of the analysis gives us at least some prima facie evidence that the two concepts are not as closely linked as one might have thought.

2. **Fred Adams** argues in favor of the hypothesis that the effect of moral considerations on people’s intentional action ascriptions is entirely a matter of pragmatics. This is an important hypothesis, and there has already been a considerable amount of discussion in the literature about whether it can adequately explain the behavioral data (Adams & Steadman 2004a, 2004b; Knobe 2004; Nichols & Ulatowski 2006). We have nothing new to contribute to that discussion here. Instead, we want to focus on questions about precisely how the debate should be framed.

Adams describes the debate by saying that his own view is that the effect should be understood in terms of ‘pragmatics’ whereas Knobe’s view is that the effect should be understood in terms of ‘semantics’ (e.g., Adams and Steadman 2004b). This does not seem quite right to us. After all, pragmatics and semantics are two aspects of our use of *language*. So if one says that the debate is between those who explain the phenomenon in terms of pragmatics and those who explain it in terms of semantics, one seems to be saying that everyone agrees that the phenomenon has something to do with language and the only disagreement is about which particular aspect of language is relevant here.

Our own view, however, is that the effect has no particular relation to language either way. Rather, it is best understood in terms of people’s *concept* of intentional action[3]. Our assumption here is that concepts are not entirely dependent on language. Researchers frequently study the use of concepts in non-linguistic creatures such as animals and infants, and it is often supposed that people make use of concepts even when they aren’t employing any of the corresponding natural language terms. It seems reasonable to suppose that people are continually using their concept of intentional action to classify behaviors into different types, even when they are not making any use at all of the English word ‘intentional’ or its synonyms.

[3] Hence, Malle (this volume) gets it exactly right when he refers to our view as a ‘conceptual solution.’

When people learn English, they learn a particular sort of connection between the word 'intentional' and the concept of intentional action. The existence of this connection makes it far easier for us to study their concept itself, and most existing studies of the concept of intentional action have in some way involved the word 'intentional' or one of its synonyms. But one should not get confused between the techniques one uses to study a phenomenon and the actual phenomenon under study. (Even if astronomers always used telescopes to acquire their data, we would not say that astronomy was the study of telescopes.)

3. In our original paper, we drew a distinction between cases in which the meaning of a word is *derived* from the meaning of some other word and cases in which the meaning of a word is simply *primitive*. **Gilbert Harman** claims that the basic distinction we are trying to draw here is a mistaken one. He suggests that it almost never happens that a word can truly be defined in terms of some other word but that we can quite often make sense of a concept by showing how it is related to various other concepts.

We fear that we did not do a very good job of explaining what we meant to say on this issue in our original article. With any luck, when we explain the point more clearly, it will become evident that we are not in fact in any disagreement with Harman's view.

To get a sense for the notion of derivation that we are invoking here, consider the relationship between the meaning of 'slowly' and the meaning of 'slow.' One might suggest that we come to understand the meaning of these two words by simply mapping each of them onto the corresponding concept, with 'slowly' being mapped onto one concept and 'slow' being mapped onto another, completely separate concept. But that view does not seem at all plausible. On the contrary, it seems likely that we map the word 'slow' onto a particular concept and then make use of a rule that enables us to derive the meaning of 'slowly' from the meaning of 'slow.'

Now consider the words 'lovely' and 'love.' Here the relationship between the two meanings seems quite different. It does not seem likely that we have a rule that enables us to compute the meaning of 'lovely' based on the meaning of 'love.' Rather, it seems that we have two distinct concepts ─ the concept *lovely* and the concept *love* ─ and that we learn the meanings of the two words by mapping each one onto the relevant concept.

Of course, psychologists might choose to investigate the underlying nature of the concept *lovely*, and this investigation might involve studying the relationships this concept bears to various other concepts. But such an investigation would not have any particular relation to the study of *words*, and it would be unlikely to be aided in any significant respect by advances in linguistics. After all, the investigation would not be concerned with the ways in which words are mapped onto concepts but rather with the nature of the concepts themselves.

In light of these distinctions, perhaps it is possible to reach a better understanding of the claim we were making about 'intentionally' and 'intention.' Our claim was that the relationship between these words is not like the relationship between 'slowly' and 'slow' but rather like the relationship between 'lovely' and 'love.' It is not the case that we learn the meaning of 'intention' and then have a rule that enables us to derive the meaning of 'intentionally.' Rather, what we have here are just two separate concepts,

each associated with a different word. (It is in this sense that we regard the meaning of 'intentionally' as a primitive.)

As Harman explains, it may turn out that the best way of investigating the underlying nature of the concept *intentionally* is to explore the relationship between this concept and various other concepts. We hope it is now clear why our own view does not involve a rejection of this claim.

4. Still, we cannot say that we are entirely in agreement with Harman's arguments. There remains an issue on which his claims conflict with ours. Specifically, Harman suggests that we have just as much reason to suppose that the meaning of 'intentionally' is derived from the meaning of 'intention' as we have to suppose that the meanings of certain other morphologically complex words are derived from the meanings of their component morphemes. The basic argument here is a simple one. Harman suggests that we are not actually able to provide strict *definitions* of these other words that appeal to the meanings of their component morphemes ─ all we can do is show certain rough relations in meaning between the two. Similarly, we can't actually define 'intentional' in terms of 'intention,' but we can show certain rough relations in meaning, and that is all that should be expected here.

This is an important and powerful mode of argument, but we do not quite agree with the way Harman applies it in this case.[4] It seems to us that the key question here is not whether it is possible to come up with strict definitions but whether it is possible to develop substantive theories that offer interesting and testable predictions. We really do have such theories on hand for other morphologically complex words, but attempts to develop comparable theories for 'intentional' have not yielded much success thus far.

For an example of a successful application of a substantive theory, consider the relationship between 'presidential' and 'president.' As Harman would surely point out, we are not able to provide a strict definition of 'presidential' in terms of 'president,' but that does not mean that we are unable to say anything of value here. In this case, it can be shown that 'presidential' can be understood either as a *relational adjective* or as a *qualitative adjective*. From this fact, one can derive some surprising predictions about the way the meaning of 'presidential' will be related to the meaning of 'president.' Consider, e.g., the very different meanings we assign to 'presidential' in (1) and (2):

(1) That is the presidential signature.

(2) That signature is presidential.

Here, (1) can be interpreted to mean that the signature was made by the president, but (2) cannot be interpreted in this way. (It can only be interpreted to mean that the signature itself has a certain 'presidential' quality.) Note that we do not arrive at this observation just by trying out a number of possible sentences and looking at how our intuitions vary from one case to the next. Instead, we were able to generate a prediction directly from a general theory about the meanings of adjectives derived from nouns.

Similar remarks apply to the relation we find between 'lustfully' and 'lust,' 'lovingly' and 'love,' 'compassionately' and 'compassion.' Harman correctly points out

[4] We are grateful to Gilbert Harman for numerous conversations regarding these issues. Although he seems not to agree with our conclusions, his comments have proved tremendously helpful in the development of our own views.

that we are not able to provide strict definitions of these adverbs in terms of the corresponding nouns. Still, it does seem that we are able to provide an interesting and successful theory here. All three of these adverbs are *manner adverbs*, and from that alone, one can derive certain predictions about the relation between their meanings and the meanings of their component morphemes. One can predict that the relationship in meaning between 'lustfully' and 'lust' will be similar to the relationship between 'lovingly' and 'love' or between 'compassionately' and 'compassion.' One can also predict that these three adverbs will play similar syntactic roles ─ and here again, the prediction is confirmed. Thus, one would tend to say 'She gazed at him lustfully' rather than *'She lustfully gazed at him,' and similarly for the other two adverbs. (Compare the perfectly acceptable sentence: 'She intentionally gazed at him.') What we have here is a successful research program examining the relationship in meaning between a whole class of adverbs and a whole class of nouns.

In the case of 'intentionally' and 'intention,' one does not find anything remotely comparable. There are no general theories from which we can derive surprising or interesting new predictions. Nor are there insights that enable us to provide unifying explanations of a whole array of apparently unrelated phenomena. Instead, what we have is an effort to accommodate each new piece of data by putting together ever more complex analyses of one word in terms of the other.

Harman suggests that, although we cannot provide a strict definition of 'intentionally' in terms of 'intention,' we may be able to gain valuable insight into the meanings of both of these words by examining their relation to each other. Clearly, this is an empirical claim, and future research might show it to be correct. At this point, however, it seems to be an empirical claim with no supporting evidence. It simply isn't the case that interesting or surprising predictions are being derived from theories about the relationship between 'intentionally' and 'intention.' Rather, what happens is that surprising predictions are generated using other approaches and that researchers then scramble to accommodate the new data by adding ever more clauses and subclauses to definitions that in some way incorporate the word 'intention.'

Of course, the considerations adduced here do not constitute a decisive refutation of the claim that the meaning of 'intentionally' is derived from the meaning of 'intention.' Our point is simply that there is strong evidence for the view that the meanings of certain words are derived from the meanings of their component morphemes but that we have yet to see any evidence for a comparable claim about the meaning of 'intentionally.'

5. **Adam Morton** points out that, although we say quite a bit about the cognitive processes underlying people's concept of intentional action, we say nothing at all about the *semantics* of that concept. The basic question here is a simple one. People use certain criteria to determine whether or not a behavior was performed intentionally, but it seems that those criteria are not themselves sufficient to pick out the property to which their concept of intentional action corresponds. Hence, even if we knew everything there was to know about the cognitive processes underlying the concept of intentional action, there would still be an open question as to which property this concept picked out. To answer this latter question would be to determine which actions truly *are* intentional.

Morton offers a series of ingenious suggestions about how one might address the relevant issues here.

Before turning to the substance of Morton's comments, we should emphasize that the usual reason for investigating folk concepts has nothing to do with learning about which properties these concepts pick out. We do not study the concepts of folk physics as part of an attempt to understand physical properties. Nor do we study folk biology as part of an attempt to understand the properties of biological systems. We study these folk concepts because we are interested in questions about how people ordinarily come to grips with certain aspects of their environments.

In the particular case under discussion here, it seems that we are touching on profound questions about the nature of people's ordinary understanding of psychological states. The data appear to suggest that moral judgments actually play a role in some of the fundamental concepts of folk psychology. Thus, the experimental results provide some initial evidence for the view that folk psychology is in certain respects radically unlike a scientific theory. These findings thereby touch on a number of important questions about the relationship between our ordinary self-understanding and the kind of understanding one seeks in the sciences (see, e.g., Knobe forthcoming).[5]

By contrast, it is not at all obvious why questions about the semantics of our folk concepts should be considered interesting or important. If we already know about the cognitive processes underlying these concepts, and if we already know about the relevant properties in the world, why is it also important to figure out which concept picks out which property?

Perhaps an example will be helpful here. Adams (1986) and Mele and Moser (1994) have each offered an analysis of the concept of intentional action. Each of these analyses picks out a property, and we could easily assign names to each of the properties picked out. We could then proceed to address important questions in cognitive science or moral philosophy by employing only these new names, each of which unambiguously picks out a particular property in the world. The only remaining debate here is about which of the two properties is picked out by our ordinary concept of intentional action. But why exactly is this issue interesting or important?

Pending an answer to this question, we propose that it might be best to focus not so much on semantic issues as on the fundamental cognitive processes underlying people's concepts.

6. In their extraordinary contribution to this volume, **Liane Young, Fiery Cushman, Ralph Adolphs, Daniel Tranel and Marc Hauser** show that the asymmetry between good and bad side-effects arises even among people who have sustained damage to the ventromedial prefrontal cortex (VMPC). In other words, when VMPC patients are given stories about agents who knowingly bring about particular side-effects, they respond exactly like normals ─ saying that the morally bad side-effects were brought about 'intentionally' but that the morally good side-effects were brought about 'unintentionally.' This is an extremely surprising finding, with a number of important implications for the study of moral cognition.

---

[5] Morton himself has made a number of important contributions to the literature on this topic. See especially his *The Importance of Being Understood: Folk Psychology as Ethics* (Morton 2003) and his superb recent paper 'Folk Psychology Does Not Exist' (Morton forthcoming).

Previous experiments have found that VMPC patients show deficits in emotional processing. Since VMPC patients show the normal response to cases involving good or bad side-effects, Young and colleagues conclude that the pattern of responses shown in normals is not due to emotional processing.

If correct, this conclusion would have implications that go far beyond the questions we had hoped to address in our original study. Many researchers assume that most of the key aspects of moral cognition can be explained either in terms of conscious moral principles or in terms of immediate emotional reactions. Since no one thinks that the asymmetry we find in cases of side-effects is due to a conscious moral principle, it is sometimes assumed that the asymmetry must be the product of an immediate emotional reaction. The results of the present study call that assumption into question. They suggest that the results might be due *neither* to a conscious principle *nor* to an emotional reaction. One then faces a difficult question as to how the results might have arisen. (One possibility would be that they are due to the operation of an encapsulated 'moral module.')

But this does not exhaust the implications of the results. Another aspect of their significance stems from the fact that earlier studies showed important differences between VMPC patients and normals on a variety of tests of moral judgment. (So, for example, VMPC patients give different answers on the trolley problem and on cases that involve offering help to distant people in need; Hauser, et al. forthcoming.) In other words, VMPC patients give *different* answers on some moral questions, but they give *the same* answers on the questions under discussion here. This result provides strong evidence for the view that the mechanism underlying people's answers to the questions under discussion here is not precisely the same as the mechanism underlying their answers to other moral questions. Overall, then, the results suggest that it won't be possible to find a single underlying mechanism that explains the full range of people's immediate moral intuitions.[6]

7. **Charles Kalish** points out that it might be possible to explain people's intuitions about intentional action entirely in terms of their beliefs about the agent's psychological states. Presumably, the agent doesn't have precisely the same attitude toward harming that he would toward helping. So perhaps people's intuitions about whether the agent acted intentionally are simply due to their assumptions about his attitude and not to their beliefs about the moral status of his behavior.

Kalish suggests that we might test this hypothesis by giving subjects a story about a highly unusual agent. This agent would have moral beliefs that diverged sharply from those of the subjects themselves. We could then figure out precisely which sorts of judgments were influencing people's intuitions. Are people's intuitions only being influenced by their judgments about what the agent *believes* to be right and wrong? Or are they actually influenced by judgments about what truly *is* right and wrong?

---

[6] Liane Young (personal communication) suggests that the difference here might arise from the fact that the intentional action task involves explicitly asking subjects to answer theory-of-mind questions whereas the other tasks involved explicitly asking subjects to answer moral questions. Another possible hypothesis would be that the difference arises because the other tasks involve emotion and this one does not. Further experimental research might help to clarify these issues.

It seems to us that Kalish is on exactly the right track here. And, as it happens, the experiment he suggests has already been performed (Knobe & Kelly unpublished data, reported in Knobe 2005). The results indicated that people's intuitions are influenced, not only by judgments about the agent's psychological states, but also by judgments about what truly is right and wrong. Hence, it seems that moral judgments actually do play a certain role in at least some aspects of folk psychology.

8. One conspicuous characteristic of work in experimental philosophy thus far is that it has been highly collaborative. For the most part, researchers have not divided off into opposing camps and tried to prove each other wrong. Instead, participants in the field have worked together in an effort to solve the difficult problems emerging from recent experimental data. The present volume is a case in point. Our critics have helped to move the discussion forward in a number of essential respects, and we hope that they remain active participants in the emerging discussion on these issues.

Works Cited

Adams, F. (1986). Intention and intentional action: The simple view. *Mind and Language*, 1, 281-301.

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

Adams, F. & Steadman, A. (2004b). Intentional action and moral considerations: Still pragmatic. *Analysis,* 64, 268-276.

Hauser, M., Young, L. & Cushman, F. (forthcoming). Reviving Rawls' linguistic analogy: Operative principles and the causal structure of moral actions. In W. Sinnott-Armstrong (ed.) *Moral Psychology*.

Knobe, J. (2004). Intention, intentional action and moral considerations. *Analysis*, 64, 181-187.

Knobe, J. (2004). Folk psychology and folk morality: Response to critics. *Journal of Theoretical and Philosophical Psychology,* 24.

Knobe, J. (forthcoming). Folk psychology: Science and morals. In Hutto, D. & Ratcliffe, M. (ed.) *Folk psychology reassessed*. Kluwer/Springer Press.

Malle, B.F. (this volume). The relation between judgments of intentionality and morality. *Journal of Cognition and Culture.*

Mele, A. & Moser, P. (1994). Intentional action. *Nous*, 28, 39-68.

Morton, A. (2003). *The importance of being understood: Folk psychology as ethics*. London: Routledge.

Morton, A. (forthcoming). Folk psychology does not exist. In Hutto, D. & Ratcliffe, M. (ed.) *Folk psychology reassessed*. Kluwer/Springer Press.

Nichols, S. & Ulatowski, J. (2006). Intuitions and individual differences: The Knobe effect revisited. Unpublished manuscript. University of Utah.