In *Philosophy of Action: 5 Questions*, Aguilar, J & Buckareff, A (eds.)
London: Automatic Press.

Answers to Five Questions

Joshua Knobe

[For a volume in which a variety of different philosophers
were each asked to answer the same five questions]

1. Why were you initially drawn to theorizing about action and agency?

Back when I was a college freshman, I started working as a research assistant to a young graduate student named Bertram Malle. I hadn't actually known very much about Malle's work when I first signed up for the position, but as luck would have it, he was a brilliant researcher with an innovative new approach.

Malle was interested in understanding people's ordinary intuitions about intentional action – the way in which people's ascriptions of belief, desire, awareness and so forth ultimately feed into the process by which people determine whether or not a behavior was performed intentionally. Of course, this sort of question had already been pursued in countless philosophy papers, but Malle wanted to use a different approach. He wanted to study the problem *experimentally*. We ran a number of experiments together and then co-authored a paper, which was published in a social psychology journal.

Yet although I found this type of research interesting and engaging, my real passion lay elsewhere. I was obsessively reading Nietzsche, along with some heavy doses of Kierkegaard, Hume, Marx, Wittgenstein and Aristotle. What I really wanted to do was to continue working on the very same sorts of questions I saw addressed in these thinkers.

And so, after graduation, it seemed like the natural thing to do would be to get some sort of day job and then devote myself to writing philosophy. I made my way from one position to another – working with homeless people in Texas, teaching English in Mexico, translating documents in Germany – but all the while, I was writing out philosophical papers. Most of these papers were fairly awful, but I do think that I was wrestling with some important issues. I was especially interested in the idea that people's ordinary understanding of each other was suffused with moral notions but that it might be possible to create a new, quite different form of understanding which had no moral character and was simply an attempt to make sense of why people do the things they do.

By this point, I was leading a kind of intellectual double life. On one hand, I was continuing my work with Malle, ultimately coauthoring six papers in social psychology. On the other, I was writing out philosophical meditations on the role of morality in ordinary thought. I would struggle with these philosophical papers for months, and then – when I was finally satisfied – I would put them in my desk drawer. It never really occurred to me to try showing this work to anyone.

But then something strange happened. The philosopher Alfred Mele wrote a reply to the paper Malle and I had published back when I was an undergraduate. Mele went over each aspect of our paper in detail, arguing that certain parts were mistaken, others were on the right track, and still others required further data before the relevant claims could be properly evaluated. But there was one aspect of his discussion that I found especially striking. Oddly enough, Mele tied these empirical questions about the concept of intentional action back to the very same questions I had been exploring in my more philosophical work.

In essence, Mele pointed out that our analysis of intentional action referred only to purely psychological states (belief, desire, etc.) and did not accord *moral* features any role in the concept of intentional action itself. This, he said, was exactly as it should be. On his view, the concept of intentional action was a purely psychological one, and any influence of people's moral judgments on their intuitions about whether a behavior was performed intentionally would have to be some kind of error.

To be perfectly honest, I had never actually made a conscious decision not to include moral features in the analysis. (The reason I had analyzed the concept of intentional action in terms of purely psychological features was just that it had never occurred to me to consider any other approach.) Mele was therefore moving the discussion in an important new direction. Where there had once been only an unarticulated assumption, there was now an explicit thesis that could be subjected to empirical tests.

But as soon as I saw my earlier assumption written out as an explicit thesis and declared to be right, I was overcome with the sense that it just had to be wrong. And so, I returned to experimentation – this time taking the empirical work seriously as a contribution to philosophy. My aim was to disprove the view I had argued for earlier, to show that the concept of intentional action could not be properly understood until one grasped the role of *moral* considerations. I began to cherish the hope that, this time, I might be able to publish the results in a philosophy journal.

2. What do you consider to be your own most important contribution(s) to theorizing about action and agency, and why?

Much of my work has been concerned with the relationship between two different ways of thinking about action. On one hand, there are questions about action that are explicitly *moral* – questions about right and wrong, praise and blame, and so forth. Then, on the other, there are questions that do not at first appear to be moral questions but seem instead to have a purely descriptive character – questions about intention and intentional action, about the agent's reasons for acting, about act individuation. A question now arises about the relationship between these two kinds of questions.

One obvious view would be that the relationship between people's thoughts about these two kinds of questions is, in an important sense, *unidirectional*. On this view, people first figure out what the agent intended, what her reasons were, what she caused. Then they use the answers to these purely descriptive questions to figure out how to address moral questions, such as whether or not the agent is to blame. But the relationship does not also go the other way. It does not happen, e.g., that people first figure out that the agent is blameworthy and then use their judgment about the agent's blameworthiness to determine precisely what she might have caused.

Plausible though it may seem, this view does not appear to be correct. Instead, it seems that the relationship here is *bidirectional*. People certainly do use information about intentions, reasons and causes to figure out whether an agent is to blame – but, surprisingly enough, it seems that they also use moral judgments to get at questions about intentions, reasons and causes.

My first experiments in this area were on people's use of the concept of intentional action. To assess the role of moral considerations, I constructed pairs of vignettes, such that the two elements of each pair were almost exactly the same except that one involved an agent doing something morally bad while the other involved an agent doing something morally good. Here, for example, is the 'morally bad' element of one of these pairs:

> 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.

After reading this vignette, try asking yourself: ‘Did the chairman *intentionally* harm the environment?’

To form the ‘morally good’ version of the vignette, we can leave almost everything the same but simply replace the word ‘harm’ with ‘help.’ The vignette then becomes:

> 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.

After reading this second vignette, try asking yourself the corresponding question: ‘Did the chairman *intentionally* help the environment?’

Comparing responses in these two cases, one comes upon a surprising result. The majority of subjects who receive the first vignette say that the chairman harmed the environment

intentionally, while the majority who receive the second vignette say that the chairman helped the environment *unintentionally*. Yet it seems that the two vignettes are entirely parallel in the mental states ascribed to the agent and in the relationship between those mental states and the resulting behavior. The principal different appears to be a moral one – the agent brings about a bad effect in the first case, a good effect in the second.

Subsequent studies have shown that this effect is not limited to the concept of intentional action in particular but also arises for a wide variety of other concepts. Take the concept *deciding*. Faced with the vignettes about harming or helping the environment people are quite willing to say:

The chairman decided to harm the environment.

but not to say

The chairman decided to help the environment.

And one gets similar effects for the concept of doing something for a *reason*. People are willing to say:

The chairman harmed the environment in order to increase profits.

but not to say

The chairman helped the environment in order to increase profits.

Subsequent work has shown that one can get this same effect for the concepts *desire, in favor, opposed*, even *advocating*.

In recent years, I have been examining the impact of moral considerations on concepts that go beyond the psychological. A particularly interesting case here is the concept of *causation*. Consider in this connection the following vignette:

> The receptionist in the philosophy department keeps her desk stocked with pens. The administrative assistants are allowed to take the pens, but faculty members are supposed to buy their own.
>
> The administrative assistants typically do take the pens. Unfortunately, so do the faculty members. The receptionist has repeatedly emailed them reminders that only administrative assistants are allowed to take the pens.
>
> On Monday morning, one of the administrative assistants encounters Professor Smith walking past the receptionist's desk. Both take pens. Later that day, the receptionist needs to take an important message… but she has a problem. There are no pens left on her desk.

And now ask yourself whether you agree or disagree with the sentences:

> The professor caused the problem.
>
> The administrative assistant caused the problem.

When Ben Fraser and I gave these sentences to experimental subjects, most agreed with the first, but disagreed with the second. Yet, from a purely scientific standpoint, it seemed that the professor and the administrative assistant stand in precisely the same relation to the event that results. The key difference is just that the professor is doing something wrong while the administrative assistant is doing exactly what she is supposed to do. Somehow people's moral judgments appear to be shaping even their understanding of the causal relations among events.

In my view, these effects reveal something fundamental about the human capacity to understand action. It seems that this capacity is not designed to deliver the same sort of understanding we seek when we are engaged, e.g., in a scientific investigation. Instead, our most

basic concepts for understanding action appear to be suffused through and through with moral considerations.

3. What other sub-disciplines in philosophy and non-philosophical disciplines stand to benefit the most from philosophical work on the nature of action and agency, and how might such engagement be accomplished?

In my view, we have much to gain from a greater dialogue between work in philosophy of action and work in the history of philosophy. Of course, research in the history of philosophy might profit from an examination of recent ideas in action theory, but there is probably even more potential for influence in the opposite direction.

The great philosophers of the past developed quite sophisticated theories about how people's minds actually worked. They asked about whether the mind was composed of separate parts and how these parts might interact. They discussed the ways in which reason and emotion might shape our moral understanding. Above all, they were concerned with questions about the moral implications of facts about human nature.

For reasons I don't quite understand, the twentieth century witnessed a peculiar decline in interest in these traditional issues. Scholars continued to work on questions about precisely what the great philosophers meant in their discussions of these questions, but there was surprisingly little interest in actually pursuing original research on the topics themselves. This has always struck me as a shame.

To take just one prominent example, Spinoza provides a complex and, I think, highly promising account of the relationship between emotion and scientific understanding. One

implication of this account is that if we truly come to understand the ways in which a person's actions are determined by prior events, we can continue to feel love for that person but will no longer feel anger at her transgressions. The key question that arises now is, *Is that actually true?* Does anger diminish when one comes to have an understanding of the causal chain that led up to an action? Is it true, as Spinoza elsewhere suggests, that we would no longer blame people for their actions if we knew precisely why those actions were performed? These questions lie right at the heart of the debate concerning freedom of the will, and I think it would be foolish to let our scruples about disciplinary boundaries get in the way of a wholehearted attempt to go after them.

In talking with other philosophers, I find that they quite often express an interest in more 'traditional' questions one sees addressed in work of Spinoza, Hume, Nietzsche and others. Still, it seems that this interest is sometimes mixed with a certain feeling of trepidation. There is a sense that one cannot address the traditional questions unless one can somehow connect the discussion back to the sorts of issues one more typically finds in contemporary journal articles. In my view, this is all a big mistake. The traditional questions truly are profound and important, and if these questions do not connect up sufficiently closely to whatever appears in the latest journal articles, well then, so much the worse for the view that one must always be relating one's work back to debates in the contemporary literature.

4. What do you regard as the most neglected issues in contemporary work on action and agency that deserve more attention?

There has been a great deal of interesting work exploring the precise patterns of people's intuitions – numerous papers revealing surprising new facts about intuitions regarding intentional action, causation, reason explanation, and so many other properties and relations – but it often seems to me that there has been a striking neglect of questions about the philosophical

significance of the various patterns that have been uncovered. In other words, there has been a neglect of questions about why it even matters whether people have this intuition or that one.

The usual answer here is that, e.g., a careful study of intuitions about the proper use of the adjective 'intentional' will help us to answer questions about which behaviors truly are intentional. I am not quite sure, though, whether this sort of claim can really answer the question. To see the worry here, imagine that a researcher is giving a talk and begins discussing a distinction which she refers to as the distinction between 'intentional' and 'unintentional.' And now suppose that a philosopher stands up and says: 'Well, technically, the distinction you are discussing there is not actually the distinction between intentional and unintentional behavior. Our analysis shows that the intentional/unintentional distinction is actually a somewhat different distinction from the one you have in mind.' At least to a first glance, it might be difficult to see how this could be anything more than some kind of nit-picky point. Something further has to be said before we could understand why it was supposed to be philosophically important.

But perhaps these issues are only distracting us from the things that are most deeply significant here. Amidst all the talk about how exactly one can use intuitions to get at various properties and relations within the world, there has been a striking lack of attention to another, far more straightforward way in which intuitions can be philosophically significant. Our intuitions do not merely give us information about properties and relations within the world; they also give us information about *ourselves*. Thus it seems that the fact that moral considerations play such an important role in so many concepts might be telling us something deeply important about the sorts of creatures we are.

What we face now is an enormous untapped opportunity. We already have before us an incredibly rich and nuanced understanding of the patterns in people's intuitions. The thing to do

now is to *philosophize* about those patterns, to ask about their meaning, their significance, what they might have to tell us about human life.

5. What are the most important open problems in philosophical theorizing about action and agency, and what are the prospects for progress?

When people consider the idea of a deterministic universe, they often find themselves pulled in a number of different directions. Something draws them to the view that people in such a universe could not possibly be morally responsible for their behavior, but it seems that there is also something that draws them toward the opposite view – the view that, even in such a universe, people could still be morally responsible. Faced with this conflict between opposing intuitions, where exactly should we put our trust? The question is a difficult one, but it seems that one helpful way to make progress here would be to think about why exactly we have come to have the intuitions we do.

One of the most exciting developments in this domain in recent years has been the emergence of systematic empirical studies that really help us to get a handle on the origins of people's philosophical views. These studies paint a complex portrait of people's ordinary intuitions about freedom of the will. When the question is posed in a way designed to trigger emotional responses, people tend to say that an agent can be morally responsible even if her actions are entirely determined. However, when the question is posed in a way designed to trigger abstract theoretical reasoning, people tend to reach precisely the opposite conclusion – that people in a deterministic universe cannot possibly be responsible for anything. This pattern of results suggests a particular hypothesis about the origins of our conflicting intuitions. Perhaps

our emotions are pulling us toward compatibilism, while our capacity for more abstract theoretical reasoning is somehow pulling us toward incompatibilism.

This incompatibilist strand in people's thinking appears to be remarkably robust. One recent study examined the intuitions of people in Hong Kong, India, Colombia and the United States. When the question was framed in a way designed to promote abstract theoretical cognition, subjects in all four of these cultures said that no one in a deterministic universe could be morally responsible. Many of these subjects had presumably never thought about the free will problem before the moment the experiment began, and yet when the problem is put before them, people in these very different cultures somehow all converge on the same answer.

What on earth could be going on here? The answer is not yet known, but just in the past year or so, there has been a surge of interesting work that explores the processes by which people ordinarily arrive at these intuitions. Misenheimer showed that people's intuitions about moral responsibility vary depending on whether they are explicitly told that something 'caused' the agent to perform the action; Nichols and Roskies showed that intuitions vary depending on whether the event is described as taking place in the actual world or in the counterfactual world; Sias showed that intuitions vary depending on whether people are explicitly told that the agent 'decided' to perform the action; Nahmias, Coates, and Kvaran showed that intuitions vary depending on whether subjects are told that the behavior is determined by psychological facts and neurological facts. Each of these studies gives us a tantalizing clue about the nature of people's intuitions here, but I do not think that any of these studies, taken in isolation, can tell us everything we need to know about the nature of the underlying phenomenon. What is needed now is an integrative theory that can account for the full range of data and explain why it is that people have the intuitions they do.