The Animal Studies Repository

From the SelectedWorks of Kristin Andrews, Ph.D.

2004

How to Learn from Our Mistakes: Explanation and Moral Justification

Kristin Andrews, *Animal Studies Repository*

Available at: https://works.bepress.com/kristin_andrews/17/

How to Learn from Our Mistakes: Explanation and Moral Justification

Kristin Andrews
York University – Department of Philosophy

ABSTRACT

A new approach to developing models of folk psychology is suggested, namely that different models exist for different folk psychological practices. This point is made through an example: the explanation and justification of morally heinous actions. Human folk psychology in this area is prone to a specific error of conflating an explanation for behaviour with a justification of it. An analysis of the error leads me to conclude that simulation is used to generate both explanations and justifications of heinous acts. It is needed in both these cases because most of us lack theoretical information about evil actors. I will argue that it is difficult to simulate such acts, and hence difficult to develop explanations for behaviour widely accepted as evil. This difficulty explains the judgements made against successful simulators by those who don't succeed, and so explains the common problem of conflating an explanation with a justification.

Introduction

The primary focus of work in folk psychology has been, appropriately enough, the understanding of other minds. Starting with the observation that we are excellent predictors of behaviour, the questions in the spotlight have been: Why are we such good predictors of behaviour? and How do we generate these predictions? For fans of folk psychology, it is the successes that are emphasized.

However, there is much to be learned from our failures as well. Though we often make accurate predictions, sometimes we don't. Our performance is also mixed when it comes to another folk psychological practice—namely, explanation. Sometimes we give ostensibly correct explanations of people's actions, but other times we miss the mark completely. Our failures are at least as interesting as our successes.

Failures in our ability to predict and explain human behaviour have received some attention,[1] but the bulk of it has come from those who argue for the elimination of folk psychology altogether. Suffice to say, merely pointing out that there are failures in our folk psychological means of prediction and explanation does not entail that eliminativism is true, or that our folk psychology is a radically false account of the motivations behind human behaviour. Even if folk psychology is not a robust predictive device, it may serve society well as a means of constructing stories that cultivate feelings of empathy for others. Or it may be an effective predictive heuristic that results in accurate enough predictions.[2] And even if we don't currently make the best predictions of behaviour, it may be possible that we can learn to make better ones. By recognizing the biases and the errors we make in reasoning we may be able to overcome certain dispositions and improve our ability to predict. Thus, there is no direct link between the claim that we are not good predictors of human behaviour, and an eliminativist position with respect to folk psychology.

I believe it is useful to look at errors in folk psychological reasoning in order to understand how we think about other minds, and as a means for improving our reasoning skills. In this paper I will focus on one common error—the conflation of an explanation of an immoral act with a moral justification. The more heinous the act, it seems, the more adamant people are that there cannot be an explanation for it. This phenomenon was clearly evident in the media reports following the terrorist attacks on the USA in 2001. Those who attempted to explain why someone might be motivated to become a terrorist were often taken to be sympathizers. When someone offers an explanation for inconceivably immoral behaviour (e.g. Hitler was a genocidal dictator because his parents were cruel), there is a tendency to think either that the explainer isn't too concerned about the moral transgression (and thus is herself morally suspect), or is an active supporter of the immoral act (and thus is, in common parlance, evil). In either case, the very act of generating an explanation for immoral behaviour is thought to entail a kind of defence of it, and some may even think that there cannot be a psychological explanation for particularly heinous actions. However, not everyone agrees. I believe that there are good reasons to seek a robust explanation for acts we strongly condemn, because the explanation can be used to curtail the unwanted behaviour. Both psychological and scientific explanations are useful as means to prepare for, deal with or even prevent the phenomena being explained. Knowing why rain falls allows us to seed clouds and keep our crops watered. And knowing how the body functions allows us to cure disease. Just as technology that allows us to control the physical world comes from scientific explanation, an explanation of horrific human behaviour may help us prevent it.

An analysis of the error leads me to conclude that simulation is used to generate both explanation and justification of heinous acts. It is needed in both these cases because, thankfully, heinous acts are relatively rare and most of us don't have much in the way of theoretical information about or experience with evil actors. I will argue that it is difficult to simulate such acts, and hence difficult to develop explanations for behaviour widely accepted as evil. This difficulty explains the judgements made against successful simulators by those who don't succeed, and so explains the common problem of conflating an explanation with a justification.

There are two points to bear in mind. First, we cannot generalize from this description of explaining heinous behaviour to the claim that every instance of explaining or justifying behaviour involves simulation. This point will be argued for in the next section. Second, I think that the folk can benefit from the knowledge of how we go about explaining immoral behaviour. Over the last 30 years social psychologists have been uncovering a host of different biases and errors that humans make in reasoning. There is evidence that as people come to learn about these biases and situational effects that influence behaviour, the biases begin to lose their power.[3] Contra the eliminativist's claim that folk psychology has not developed in the past 2000 years, we see that even in the past 100 years the folk's conception of psychological motivation changed to absorb the claims of clinical psychology, especially Freudian psychology. By uncovering and publicizing mistakes in our reasoning about other minds, today's folk can come to learn from our mistakes as well.

Background

The long-held view that our understanding of other minds took the shape of a theory came under attack in the 1980s. According to the traditional account of folk psychology, we attribute beliefs and desires to others and then appeal to an implicit folk theory in order to predict and explain others' behaviour (e.g. Dennett 1987; Fodor 1992; Gopnik and Wellman 1992; Sellars 1956). Subsequent critiques of the traditional view, which came to be known as the theory-theory, involved claims to the effect that it was inaccurate, too vague and non-universalizable (see e.g. Churchland 1981; Goldman 1989).

The eliminativist criticisms of the theory-theory opened the door to competition in the form of a new theory that purported to avoid these criticisms. This new theory, dubbed simulation theory, was thought to offer an adequate account of our ability to make folk psychological predictions. Among the supporters of the different theories described as simulation accounts are Jane Heal (1995), Alvin Goldman (1989), Robert Gordon (1986) and Paul Harris (1989). According to the simulation theory broadly understood, one can understand others by using one's self as a model. In order to predict what someone is going to do, I feed pretend inputs (which may include beliefs, desires, facts about the environment and social situation, past events, etc.) into my own decision-making mechanism. My decision-making mechanism is what leads me to act, and the output is usually behaviour. However, when I mentally simulate another, I take my decision-making mechanism off-line and output information rather than behaviour. This information allows me to make a prediction about the target's behaviour. The simulation theory was thought to avoid the eliminativist criticisms of the theory-theory primarily because it rejects the notion that we all share some tacit theory of human psychology. For if our folk psychology is not theoretical, and only vague or inaccurate theories should be eliminated, then folk psychology isn't the sort of thing that can be eliminated.

After some 25 years of debate, there is a growing recognition that both theory and simulation might be part of a larger folk psychological repertoire. The attribution of beliefs and desires by appeal to a theory, an act of mental simulation, generalization from past behaviours and appeal to character traits have all been seen as relevant to the prediction and explanation of others' behaviour. Josef Perner was one of the first to argue that we should move beyond the dichotomy when he wrote, 'the future must lie in a mixture of simulation and theory use. However, what this mixture is and how it operates must be specified in some detail before any testable predictions can be derived' (Perner 1996, 103). A survey of the published literature shows that the hybrid view is beginning to dominate. For example, Nichols and Stich have developed a model of folk psychological reasoning that includes both theory and simulation elements (Nichols and Stich 2003) and Currie and Ravenscroft (2002) admit that there is a theoretical aspect to some simulations. The perceived sharp divide between theory and simulation has been further undermined by those who have pointed out that the two views are not necessarily inconsistent, insofar as non-propositional accounts of theories are correct. If theories can be models rather than appropriately organized sets of propositions, then the simulation theory can be seen as theoretical (Giere 1996; Stich and Nichols 1995).

Given the realization that both simulation and theory play a role in our reasoning about others' mental states, the next natural step is to recognize that different folk psychological practices may utilize different mechanisms. Folk psychology involves not just predicting behaviour, but also explaining, justifying, understanding and co-ordinating behaviour. One problematic assumption during the heyday of the debate between simulation and theory was that the same mechanism must be used for both prediction and explanation. Thus, advocates of theory-theory and simulation theory implicitly embraced the symmetry of psychological prediction and explanation, the view that explanations are backward predictions and that the same cognitive architecture serves as the foundation for our predictive and explanatory behaviour. The symmetry thesis takes prediction and explanation to be two sides of the same coin, arising from the same mechanisms, and all predictions automatically provide an explanation of the act. However, as it was with the covering-law account of scientific explanation, there are counter-examples to the symmetry of psychological prediction and explanation. For example, I can predict that Frank will speak out in the department meeting, because he always has something to say. But I don't know why he feels the need to speak on every issue. Perhaps he is overly dedicated to the department, or perhaps he craves the spotlight. In addition, though some kinds of explanations will allow for future predictions, others will not, due to the pragmatic nature of psychological explanation.[4]

The point can be generalized: what we learn about one folk psychological practice may not apply to others, and hence we cannot examine folk psychological behaviour in a department meeting situation and blithely apply it to a terrorist hunt. Those of us interested in folk psychology must ask more narrow questions than the ones previously asked. For example, Paul Bloom restricts the focus to the folk psychological skills that children must have in order to acquire language. He argues that in order to learn new words, children must attribute beliefs because children must make inferences about others' intentions in order to realize that the term being used refers to an object in the world (Bloom 2000). From this one claim, however, we need not infer that children attribute beliefs in all their other folk psychological practices.

The conclusion that I draw from these recent developments is that the next step in the investigation is to examine particular cases of folk psychological practices, and to ask more narrowly focused questions. Rather than asking about the mechanism subsuming our folk psychological practices, we can ask how people explain or predict or justify or understand or co-ordinate behaviour. And rather than asking how people explain behavior generally, we can compare how people explain familiar versus unfamiliar behaviour, acceptable versus unacceptable behaviour, behaviour of friends and family versus behaviour of strangers, etc. This paper takes up that challenge by examining a specific act of folk psychology, namely the construction of explanations for actions judged as wicked. The answers we find in this case cannot be directly applied to other cases, given the above reasons. But they can be used to help us realize the difficulties associated with generating such explanations, and prevent us from judging too quickly those who do explain behaviours that we've condemned as evil.

Explaining Immoral Behaviour

Relevant to the current analysis is the fact that people have some difficulty in generating explanations for acts they strongly condemn. The more one is horrified by an action, the more difficult it is to explain it. A clear example of this observation is the response to the 2001 terrorist attacks in the USA. What passed as explanations were trait attributions that served to denounce the actors; they were evil, monsters and so forth. Essayists who published articles offering explanations of the terrorist attacks based on US support of Israel, its stationing of troops in Saudi Arabia, or the US-sponsored sanctions against Iraq were seen by many as condoning the attacks. The criticism was not directed at the truth of the facts presented, but was based on the belief that such explanations implied sympathetic feelings toward the terrorists. Worse yet, some thought that the explanations functioned as a justification of the attacks. To explain the attacks is not to excuse them, but the two were often conflated in public debate.

One might object to this example, and explain the difficulty Americans have with making this distinction by noting that some may have felt that the explanations were a condemnation of the US government or US citizens. People who felt this way may have been uncomfortable or unhappy with condemning the USA at the same time the country was being attacked from the outside. Thus, one might think that this case is different from a general case of providing explanation of terribly immoral actions, because the explanation in this case could be seen as a personal criticism. It may be that the individuals who rejected the proposed explanations by identifying them as justifications were simply trying to avoid the criticisms they thought were implied by those explanations. Even if this is the case, I believe that the phenomenon can be seen in other instances as well.

Consider, for example, some of the responses to Alliance Atlantis' decision to produce a mini-series about Hitler's early life, his upbringing and early political career. A well-made movie about Hitler's upbringing may help to explain his intense hatreds, and may help us to understand his motivations. It seems this was the worry of critics such as Abraham Foxman, national director of the Anti-Defamation League. In response to the announcement, he protested, 'Why the need or the desire to make this monster human? The judgment of history is that he was evil . . . Why trivialize the judgment of history by focusing on his childhood and adolescence?'

It seems that some people would prefer to think that no psychological explanation is possible for horrible acts. There may be comfort in thinking that our world could not allow for such evil, and that the act is an anomaly rather than something to be explained. Of course we know this idea to be false; if every event has a cause, then there is a causal explanation for every event. And if the event to be explained is an action, that explanation should include reference to the actor.

Compare the two examples given above with examples from fiction. A good film or novel can invite or even force us to understand the motivations of the characters, even when the characters are terrible people or do horrible things. A film such as The Silence of the Lambs, which asks us to enter the mind of the killer, or Happiness, which presents the world from the perspective of a paedophile, provides us with the opportunity to explain the characters' behaviours. We understand their motivations, and we can generate explanations for their actions. However, corresponding to that ability are strong negative emotions; when we see Hannibal Lector escape from his cage by brutally attacking the police officers, and stringing their mutilated bodies up on prison cell bars, our prior identification with Hannibal Lector makes the scene even more difficult to watch.

Similarly, when we read Dostoyevsky's Crime and Punishment we are privy to Raskolnikov's inner dialogue, and so we have a full understanding of his actions. His murder of the pawnbroker was not understood by the lawyers who tried his case. They could only conclude that 'he had been led to the murder through his shallow and cowardly nature. . .' But because we are given epistemic access, the readers understand that he was trying to save his sister's life, and his family honour. This understanding doesn't lead the reader to condone Raskolnikov's actions, but it makes them comprehensible, and this is what distinguished the readers' view of Raskolnikov from that of the lawyers.

These cases demonstrate a difference in one's ability and willingness to understand the motivations behind an immoral act. What do we look for when we ask for an explanation of behaviour? We think that people act for reasons, and those reasons are seen as playing a causal role; the attitude causes the action.

If an explanation attributes a belief set and a desire set that is seen as causally sufficient for an agent to do what she did, we are typically satisfied with the explanation as a psychological one.[5] But often we do not present explanations in this form. Research in social psychology indicates that explanations which refer to character traits, past behaviours, situational factors and group stereotypes can also help us to make sense of behavior (Kunda 2002). Whereas the tradition from the philosophy of science literature emphasizes the deductive nature of explanation, with beliefs and desires as the explanans, the social psychology literature suggests a coherence model of explanation. The aspects of the explanation must support one another rather than compete, and there should be a positive constraint between the explanans and explanandum (Thagard and Kunda 1997). A person's past behaviours (e.g. she always brings wine to the party) and her traits (she is very generous) cohere with current behaviour. We understand that behaviour; it makes sense. Behaviour which we can make sense of is also behaviour that we can provide explanations for. These explanations may or may not be accurate in various ways, but they appear to satisfy the folk.

So in order to explain why Raskolnikov killed the pawnbroker and then hid the loot never to return, we can talk about his beliefs and desires. Raskolnikov desired to save his sister from marrying a man she didn't love, he believed that his sister was marrying for money and he felt responsible for her plight. We can also talk about his traits: Raskolnikov was a proud man of somewhat limited intellect who was unable to take all the consequences into account when performing his utilitarian calculation. Both these approaches help us to understand his actions, and if we had access to his lawyers, we might be able to alleviate their confusion about his motives by telling them these facts (or by handing them Dostoyevsky's novel).

Explanation and Justification[6]

Though some have argued that we provide explanations for actions in order to justify them (e.g. McDowell 1985; Morton 2003), there is good reason to reject this claim. Actions explained by appeals to weakness of the will do not, for example, serve to justify those actions. The folk are quite capable of making the distinction between explanation and justification in some domains. For example, one can explain how a magician seems to pull a rabbit out of thin air without implying any value judgement about the trick. We will also ask for explanations of morally neutral behaviour such as a fashion faux pas. We ask, ‘Why did he wear that horrible tie?’ and are told ‘He’s colour blind’ or ‘He has bad taste’ or ‘His mother gave it to him’. The different explanations will likely have different effects on the questioner—some may resonate as a good reason, and hence also serve as a justification for wearing the tie. If the underlying belief is that a mother’s present must be worn, even if it is an ugly tie, then the final explanation will also act as a justification. However, ‘He has bad taste’ doesn’t justify his choice of tie. It does make sense as an explanation, though, because it refers to a character trait that might cohere with other things that we know about him. That explanation will only make sense, however, if there is no contradicting evidence about the person’s taste. If he has never worn such ugly clothing in the past, and has a beautifully decorated home, etc., then that explanation will not suffice. It works as an explanation if we think the trait attribution is correct and we also see the causal connection between the trait and the behaviour to be explained.

Thus, not all explanations serve as justifications. But the disjunction between the two is even stronger than that, because in some cases we offer an explanation in order to denounce an act. If we further consider the case of the man who wears an ugly tie, we see that rather than serving as a justification, the explanation for the bad tie in terms of the wearer’s bad taste comes very close to providing a condemnation of his donning the tie. In this case it isn’t a moral condemnation, but an aesthetic one.

This defence of the claim that the psychological explanations we provide in our daily lives can be distinguished from endorsements need hardly be given, however. The explanation of Raskolnikov’s behaviour I gave in the previous section suffices as a counterexample to the claim that we provide explanations in order to justify behaviour. I don’t think that Raskolnikov did the right thing just because I understand why he acted. That is what makes him an anti-hero. It is not the case that whenever we ask for an explanation we want a justification; sometimes we just want to know.

Nonetheless, it is true that in informal conversation we often use the terms ‘explanation’ and ‘justification’ interchangeably, as in ‘Look, I can explain!’ when we are confronted with a disapproving glare. As well, when attempting to justify an action we often neglect to mention the relevant value. I may justify my speeding to a police officer by saying that I believe my passenger is about to give birth, and I want to get her to the hospital in time. This is an explanation for my speeding, but it becomes a justification given the unstated value claim that getting someone to the hospital in time for her to give birth is more valuable than driving at a legal speed.

Though we sometimes make errors and neglect to make the distinction between explanations and justifications, at the same time we accept that such a distinction exists. Because there is a certain similarity between an explanation and a justification for someone’s action, it is all the more important to distinguish between the two. A justification is a rationale for action, and sometimes an argument that the act was right. When I attempt to justify an action that you see as immoral, I am trying to convince you that you’re mistaken, and that the act is not immoral. While some justifications consist only of the normative claim (e.g. because it’s allowed) or a description of the situation (e.g. because it made more people happy), the sorts of justifications relevant to the issue at hand are the narrative accounts a third party provides in order to condone an action. Given that the mistake consists of conflating an explanation with a justification, where the explanation is seen as an account of the causal antecedents that serve as reasons for the behaviour, it suggests that we take some explanations and some justifications to be of the same form. The question prompted by this mistake is why we sometimes have difficulty accepting a causal explanation as a mere explanation rather than a defence of the behaviour.

When offering a justification in terms of antecedent features, I tell you what I believe about the cause of the event, and try to convince you that my beliefs are true. In addition, I try to convince you to share my value. The realm of value is where explanations and justifications differ; a justification is an explanation plus a value claim. We desire to justify an action because we want to show that it is morally permissible, pragmatically reasonable, or otherwise acceptable. And because we give explanations for actions that we don't want to justify, there is a conceptual wedge between explanations and justifications. Not all reason-giving should be seen as normative.

Theories of Folk Psychology

Now that the problem is set out, we can attempt to account for the phenomenon by looking at how current theories of folk psychology might account for the difficulty people sometimes have with explaining highly immoral actions. I will examine four alternative accounts of folk psychological prediction and explanation: induction, trait attribution, theory-theory and simulation theory. I don't mean to suggest that one of these four theories will fully account for our ability or inability to explain immoral actions; that should be evident by the background discussion. Rather, the idea is that an examination of each method's capacity to account for our ability to explain immoral behaviour will indicate the extent to which each method may play a role. The goal is to come up with a plausible description of what subserves our ability to explain heinous behaviour.

Induction

Elsewhere I have challenged the notion that humans are excellent predictors of behavior (Andrews 2003). Though we do enjoy some success in predicting what others will do, I argue that our successes are limited to certain domains. We are most successful at predicting behaviour in contexts that we are familiar with. The stock examples of folk psychological prediction, such as getting a cold beer from the refrigerator, are perfectly accounted for by mere inductive generalization. I cannot predict of Joe Smith, whom I have never seen before, that he will get a beer when he goes to his refrigerator. If I have had no previous exposure to Mr Smith, I may not even know whether his refrigerator contains beer or medical waste.

I can predict that at least some of my students will show up to class on Tuesday. And I can predict that my spouse will eat a meal tomorrow. These sorts of predictions are amazingly reliable, but once we begin leaving familiar territory (what would my spouse do when confronted with a hostage situation?) or once we look for more specific or exact predictions (what time will Pavel show up to class on Tuesday?) our rate of success drops dramatically. Given that the majority of our accurate predictions are in domains with which we are familiar, and involve situations and actors we are familiar with in some way or other, it becomes easy to make the argument that these successful acts of prediction are not dependent on a robust understanding of folk psychology, but only on past experiences that can yield inductive generalizations.

Though induction is useful for generating predictions, there is good reason to doubt that it will provide much in the way of explanation. Though we do sometimes answer why questions with an inductive generalization ('Why was she taking her watch on and off during the lecture?', 'Well, she always does that'), such responses are often not very satisfying. In some cases it is an admission of defeat; an inductive answer to a why-question might mean that there is no psychological explanation for that behaviour (or at least no known psychological explanation), in which case a physical explanation may be more appropriate. In other cases, an inductive explanation is vacuous. Why was she acting shy at the party? Because she always acts shy at parties. This answer obscures the real question: what is it that makes her shy? There are some cases in which we will be satisfied with an inductive explanation; learning that someone has a habit of playing with her watch during lectures can allay your fears that e.g. she thought your lecture was terribly boring. Though 'she always does that' is not a full explanation for the behaviour, it can help to address the asker's concern. If explanations are answers to why-questions, as van Fraassen (1980) argues, then there is a pragmatic element to explanation. We can give different kinds of acceptable explanations for the same event depending on where our interest lies. A neuroscientist, a psychologist, and a physicist may each be looking for a different answer to the question 'Why did she play with her watch during the lecture?' because they are interested in different levels of explanation.

Given that inductive explanations are sometimes appropriate, the question for us is whether they are given to account for heinous behaviours. It seems not. Given the unique and thankfully rare nature of horrific acts, an appeal to induction in order to help construct an explanation for that behaviour would not be terribly useful. It certainly is not a common response given by the folk when attempting to provide explanations of evil behaviour.

Trait Attribution

Recent findings in social psychology have suggested yet another way in which we might explain behaviour.[7] There is evidence that we often appeal to character traits in order to predict behaviour, and also to explain and describe others (Miller 1984; Park 1986; Ross and Nisbett 1991; Winter and Uleman 1984). For example, if I put Susan in charge of bringing drinks to the party, and I want to know whether I can rely on her to being enough for everyone, I can be reassured when I realize that she has the trait of generosity. Conversely, if I want an explanation for why she brought so many bottles of liquor to the small gathering, I can also explain that behaviour based on her generous nature. I need not explain all such behaviour this way; another colleague's apparent generosity in this regard may be explained by referring to his nature as a drunkard.

The attribution of traits goes hand in hand with a number of heuristics we use when asked to understand others, whether we are predicting or explaining. Some of these heuristics are reliable, but famously, others are not.[8] The most important lesson to take from all this is that our predictions and explanations are often inaccurate, even when the object of analysis is one's self. Insofar as these heuristics help to comprise our folk psychology, there is room for improvement.

One of the heuristics that is not terribly reliable is the spontaneous attribution of traits. However, research suggests that we do view behaviour dispositionally, and we expect an individual to engage in behaviour consistent with an attributed trait across a variety of situations. Because our focus here is on explanation, I will skip over a discussion of how trait attribution may be used to make folk psychological predictions. Rather, we will look at the role trait attribution plays in the explanation of heinous behaviour.

I think it's fairly evident that the most common kind of explanation given for heinous behaviour is an attribution of a trait. Serial killers are described as evil, terrorists as monsters who live in holes under the ground. It was not uncommon to hear people explain the 9/11 terrorists by attributing a trait, which sounded very much like a curse—they did it because they were monsters, or because they were religious fanatics, or some such thing. These explanations suggest that people acted immorally because they are not like us. The paedophile molested small children because he was sick, and John Hinckley, Jr shot

Ronald Reagan because he was crazy. This quick attribution of traits in order to explain immoral behaviour is easy, and common. I'd wager that people do not have much difficulty generating these explanations, but some will reject them as too simplistic, partially for the reasons social psychologists tell us to be wary of them.

One thing left out by trait attribution explanations is any reference to the goal of the action. It is part of our folk psychology that people act for reasons, and even though the trait might help us to understand why a person took a particular action over another, it doesn't illuminate the goals that may have motivated the behaviour in the first place. It is at this point that we can turn to the two mainstays in the folk psychological literature, the theory-theory and the simulation theory. Both theories attempt to illuminate our approach to generating explanations in terms of reasons for all kinds of behaviour.

Theory–theory

If I use the theory-theory to predict what a person will do by appealing to initial belief/desire conditions C and the relevant theory T, I can infer with some degree of probability that the person will engage in the behaviour B. In contrast, when I explain behavior B, I look for appropriate initial conditions C and a theory T that implies B.

Instead of using beliefs and desires as the input, when I generate an explanation the behaviour is the input, and the theory is used to determine which sets of beliefs and desires are consistent with that behaviour. (Remember, for the theory-theory, psychological prediction and explanation are symmetrical.) Thus, to generate a justification of the behavior I would start out using the same process I use for explanation. That is, I use the above method to find out why the person acted as he did, and I then analyse the explanation to determine whether those reasons justify the behaviour.

Theory-theory can point to two places where a person might fail to accept the existence of an explanation for an atrocious act, which correspond to the two steps involved in generating explanations.[9] The explainer is either unable to generate potential belief/desire sets, or she is unable to test the belief/desire sets that are generated.

Take the case of Americans responding to suggested explanations for the 2001 terrorist attacks as though they were justifications of the events. In this instance people had difficulty distinguishing between an explanation and a justification not because they had a difficulty generating the belief/desire sets; there were any number of authors willing to offer explanations for the attacks. These writers were condemned because the explanations they offered were seen as attempts to justify the terrorist attacks. (A good example of this is the response to Susan Sontag's New Yorker article, in which she offered an explanation for the 9/11 terrorist attacks by ascribing intelligible motives to the terrorists. The New Republic responded to Sontag by publishing an article which began, 'What do Osama bin Laden, Saddam Hussein and Susan Sontag have in common?') When writers such as Sontag offered explanations, the refusal to accept their explanations was not due to a difficulty in generating potential belief/desire sets. Those had been given. The difficulty seemed to arise at the point of testing those sets by appeal to a theory.[10]

When someone who is trying to explain a behaviour fails, either because she is unable to generate the relevant belief/desire set, or unable to use her folk psychological theory to test the belief/desire set, it seems there must be a problem with the folk psychological information being used. The database of folk psychological knowledge might not contain the information necessary to make the inference. Perhaps an individual's database doesn't tell her how people are going to react when they feel that their sacred land is desecrated, or what it is like to want sex with children. Also, and plausibly, there may be a gap in the non-psychological knowledge. If someone doesn't know that there are American troops in the Holy Land, then she won't be inclined to generate a belief/desire set reflecting this fact.

Some of this information is easily acquired. One can learn about foreign events by reading newspapers, attending lectures or engaging in discourse with others who are informed. But even if we have a robust understanding of the relevant details of the situation, additional psychological information that is not discursively learnable is needed. I cannot find out what it is like to be you simply by reading a list of facts about you. Though the facts will give me some further information, they don't provide the qualitative experience that might help to make sense of how those facts effect you. To generate the belief/desire sets in order to test explanations for Middle Eastern terrorism, it would be helpful to know how one would respond to the perceived occupation of the Holy Land, to the death of one's parents at the hands of soldiers or to life in a refugee camp. Likewise, to explain the paedophile's behaviour, we may have to engage in some very difficult acts of the imagination. I will return to this point shortly.

Simulation Theory

First, let us take a look at how simulation theory generates explanations for behaviour. Both Goldman and Gordon have argued that simulation can produce explanations just as well as the theory-theory. In his account of simulated explanation, Goldman argues that an explanation involves telling a story that refers to the actor's beliefs, goals and desires, which will allow the simulator to reject alternative hypotheses that do not conform to those belief/desire sets. Possible explanations are run through the simulation until the behaviour to be explained is outputted, at which point that belief/desire set is taken as a possible explanation (Goldman 1989).

Gordon agrees that simulation can be used to provide explanations, and though his language is different, his account is very similar to Goldman's. Instead of belief/desire sets, Gordon talks of models that can be manipulated until the target behaviour is exhibited. The model allows us to determine which features interact to cause the target behavior (Gordon 1992). Gordon has argued that explanations of intentional behaviour need not refer to a person's psychological state, but that facts about a person's 'epistemic horizon' can and do serve equally well as explanations (Gordon 2000). Thus, even versions of simulation theory that take facts about a person's situation rather than her beliefs and desires as inputs to the simulation are thought to provide satisfactory explanations.

Of course there will be many possible belief/desire sets that are compatible with any particular behaviour, and which, through simulation, would lead to the action we are attempting to explain. In order to adjudicate between those, Goldman suggests we assume that the agent is like us, that we share relevant psychological features. Then we would accept explanations that seem natural to us, and reject those that are less natural (Goldman 1989). Goldman even suggests that the simulator should assume shared basic likings and desires, unless there is reason to think otherwise. As Stephen Stich puts it, a simulator is like a king's taster—there must be a certain degree of similarity between the simulator and the person being simulated, or the simulation will not go through. With an elephant as king's taster, there is no guarantee that a bit of poison won't slip by.

A simulator who is unwilling or unable to consider explanations for immoral actions is likely to cause problems for folk psychological theories. The simulator who cannot consider Sontag's explanation, for example, might be unable to consider it because she cannot imagine being a person able to commit mass murder for political purposes. Gordon's account of explanation places emphasis on the imaginative identification with the actor whose behaviour is to be explained. The breakdown in a simulated explanation would most certainly come when there is little or no identification between the simulator and the actor.

When errors arise, the simulation theorist could look at Goldman's suggestion that we assume that the other person has the same basic likings. Because I do not have the same likings as a paedophile, my simulation would go wrong if I didn't adjust my own basic likings to reflect his, and that could be a demanding task. Moreover, if the simulator and the simulated are in radically dissimilar environments, the differences between the two may be so great as to render the simulation extremely difficult if not impossible. For example, if I have my basic needs met, I may be unable to understand what it is like not to have those needs met. If I am not discriminated against based on my race, religion or gender, I may not be sufficiently similar to one who is discriminated against to get a simulation started. Certainly in such cases it would be difficult to come to grasp the what-it-is-like knowledge that is used to make a simulation. Rather than admit my inability to simulate, then, I may conclude that there is no explanation at all for the extreme behaviour of people whose race, religion, gender or basic likings are different from my own.

Discussion

The possible breakdowns in generating an explanation using the theory-theory are at the level of gaps in the theory, and most problematically, gaps in the psychological information. And regardless of whether simulation theory or theory-theory is used, the explainer will need to have basic non-psychological information; I will be wholly unsuccessful predicting the religious behaviour of someone I know to be Christian unless I am familiar with some of the basic practices of Christianity, for example. But this information can be learned discursively. The primary difference between folk psychology as simulation and folk psychology as theory is in how the psychological information is gained. Theory-theory says little about this; we simply have the facts about what people will do when they think and feel in certain ways. Simulation theory places emphasis on this aspect of folk psychological understanding: we learn what it is like to be another by identifying with them, or through some act of the imagination.

Film is an interesting case in comparison because we can generate explanations for heinous fictional behaviour (we wouldn't be so intrigued by films such as Happiness or novels like Crime and Punishment if the immoral characters were not believable). Some good films will fill in the relevant gaps to provide the audience with all the non-psychological knowledge needed to follow the story. The same is true for gaps in our psychological knowledge. I may not know what it is like to be a refugee, but a well-made film may help to close those gaps in my theory by giving me a taste of what refugees think, feel and do in particular circumstances.

An important difference between film and reality, it seems to me, is that while we are asked to identify to some degree with characters of fiction, it is rare that we are invited to identify with actual people who have acted immorally. Certainly the media reports of heinous actions often do the opposite, and help us to see the perpetrator as someone with whom we cannot identify. Whether the person is labelled with a derogatory trait term, or an unflattering picture of the person is shown, the media can obstruct the audience's ability to imagine being that person.

Even when I know all the discursive facts about a situation, it may be that I lack some knowledge. The what-it-is-like-to-be experience of a terrorist or a paedophile is something that may be closed to me. I cannot gain such knowledge by being told facts about the terrorist's past, or the political situation in the terrorist's country. Rather, I can only begin to learn it by having certain kinds of experiences. Once I've had the appropriate experiences, I can add that knowledge to my database (if using theory to generate explanations), but to fill the gaps in the theory, it seems that simulations are necessary. That is, I must be able to successfully simulate the terrorist to some degree before I can generate the psychological explanation.

This may help to explain why some people conflate explanation with justification. The lay analysis seems to be, if one is able to identify with a person seen as evil in order to provide an explanation, then there must be some evil in her as well. Thus you have people claiming that Susan Sontag, who was able to engage in a degree of identification with bin Laden, is also a terrorist. However, as we should have learned from watching film, identifying with someone does not require agreeing with him, or condoning his actions. When explanations for immoral actions are offered that do not explicitly condemn the actor, it is assumed that the explanation serves to justify. I may offer a fairly bad explanation of someone's evil act by saying that he is crazy, and no one thinks that serves as a justification. But I may not say that he is poor or he was abused or anything that might generate emotions of sympathy or empathy for the perpetrator.

The problem we have with conflating explanation and justification does not come from an inductive or trait-attribution method of explaining behaviour, but from a lack of knowledge that can only be filled by engaging in a simulation. The knowledge about the motives of someone very different from you requires an act of imagination. Clearly there are different folk psychologies across cultures and subcultures, and in order to understand someone from one of these different cultures we need to be able to expand our own folk psychological knowledge. It is one thing to be told facts about people from these other cultures, and another to believe those facts. Without spending a lot of time around people who are immoral and learning through proximity how they think, what is required is a series of mental simulations. What I suggest here is that the failure to identify explanations for heinous actions can be seen as a gap in our folk psychological theory, but a gap that can only be filled by engaging in a simulation. It seems to me that there is an essential role for simulation in constructing explanations of behaviour perceived to be immoral. How we go about simulating in order to gain the what-it-is-like information is another issue, and though I have suggested above that film or other methods of evocative story-telling may help us with this, that issue is open to further exploration.

Some people may fail to consider plausible explanations for atrocious behavior because it is terribly difficult to do so. It is so difficult because the act of simulation requires us to imagine being a heinous actor. The degree of difficulty leads to the mistake of denying the existence of an explanation altogether. I hope that the folk can learn from this mistake by first learning how we explain behaviour. That information, which is discursively learnable, can be added to the body of folk psychological knowledge. Our folk psychology has changed over time, and there is no reason why the current findings of philosophy and psychology can't be incorporated into an existing folk psychology just as the concept of the subconscious was added in response to the theories of Freud. Along with this information will come implicit permission to engage in those difficult acts of the imagination needed in order to gain the non-discursively learnable knowledge about why people do terrible things. Such knowledge is essential if we are to control heinous behaviour successfully. If we can do that, we will have truly learned from our mistakes.

ACKNOWLEDGEMENTS

Thanks to James Harold, Brian Huss and Christopher Eliot for helpful comments on drafts of these papers. I would also like to thank the audience at CENSS 2002, the philosophy department at the University of Cincinnati and two anonymous reviewers for this journal.

NOTES

1. See, for example, the debate over the cognitive penetrability of mental simulations. It has been argued that if we fail to predict a common behaviour, we must not be using our own cognitive mechanisms to make that prediction, and thus we must not be simulating (Stich and Nichols

1995). This is the case with many behaviour patterns uncovered by social psychologists such as the endowment effect (placing a higher monetary value on objects when we own them) and the fundamental attribution error (minor situational factors can have a dramatic effect on behaviour, as in the case of Milgram’s infamous experiment). One response to this criticism is that the failed prediction may be caused by imputing faulty initial states (e.g. Gopnik and Wellman 1992).

2. As an anonymous reviewer for this journal pointed out, the eliminativists seem to hold folk psychology to too high a standard.
3. For example, Beaman and colleagues demonstrated that the more hurried a person is, the less likely she is to offer aid to a person obviously in need (even if the reason for hurrying is not a very good one). However, they also found that people who were given a lecture on this effect were less likely to be influenced by it even two weeks later (Beaman et al. 1978). It is still not clear how much education is needed to immunize people against such biases, and whether ongoing instruction is required.
4. See Andrews (2003) for a discussion of these issues.
5. Of course, there is good reason to think that no finite set of beliefs and desires can determine any particular action. There could always be an unknown defeater outside the set that would cause the actor to engage in some other act. For a discussion of this point, see Morton (2003). However, this metaphysical point is entirely consistent with the claim that humans do in fact understand belief/desire sets to be sufficient for predicting or explaining behaviour. As we know from Nisbett and Ross (1985) and Kunda (2002) our heuristics are not limited to those that guarantee truth.
6. My focus in this section is on the justification of behaviour rather than justification more broadly construed, so I am not interested in epistemic issues dealing with justified belief. However, I aminterested in more than moral justification. While justifications of behavior are value claims, they are not all ethical claims. A behaviour might be justified on aesthetic grounds, for example. The issues at hand is whether a behaviour is acceptable on some grounds rather than whether it is truth oriented or reliable.
7. Trait attribution and induction might at first appear to be describing the same method of predicting behaviour. Since we may come to attribute a certain trait to someone after observing her behaviour for a while, induction certainly has a role to play. However, one can come to attribute traits to people via other methods as well. For example, if I am told that a person that I have just met has a certain trait, I may use that information in order to predict what he will do, even though I have no independent evidence that the person actually behaves in a way consistent with that trait.
8. Our decisions are not based on a good understanding of probabilities; we often ignore base rates, we underestimate the importance of consensus information and the power of situations to influence behaviour, we ignore sample size, and so forth (Kunda 2002).
9. Of course it is possible that there are some other problems; for example, the person might have a theory of mind deficit. However, here I am only concerned with people whose cognitive capacities are considered normal.
10. One might also reject Sontag’s explanation because they think the facts she presented are false. However, that wasn’t the criticism. It seems that it was the mere attempt at providing an explanation that illuminated the cause of the attacks which was so offensive.

REFERENCES

ANDREWS, KRISTIN. 2003. Knowing mental states: The asymmetry of psychological prediction and explanation. In Consciousness: New philosophical essays, edited by Q. Smith and A. Jokic. Oxford: Oxford University Press.

BEAMAN, A. L., P. J. BARNES, B. KLENTZ, and B. MCQUIRK. 1978. Increasing helping rates through information dissemination: Teaching pays. Personality and Social Psychology Bulletin 4: 406–11.

BLOOM, PAUL. 2000. How children learn the meaning of words. Cambridge, Mass.: MIT Press.

CHURCHLAND, PAUL. 1981. Eliminative materialism and the propositional attitudes. Journal of Philosophy 78: 67–90.

CURRIE, GREGORY, and IAN RAVENSCROFT. 2002. Recreative minds. Oxford: Oxford University Press.

DENNETT, DANIEL. 1987. The intentional stance. Cambridge, Mass.: MIT Press.

FODOR, JERRY. 1992. A theory of the child's theory of mind. Cognition 44: 283–96.

GIERE, RONALD. 1996. The scientist as adult. Philosophy of Science 63 (4): 538–41.

GOLDMAN, ALVIN. 1989. Interpretation psychologized. Mind and Language 7 (1/2): 161–85.

GOPNIK, ALISON, and HENRY WELLMAN. 1992. Why the child's theory of mind really is a theory. Mind and Language 7: 145–71.

GORDON, ROBERT. 1986. Folk psychology as simulation. Mind and Language 1: 158–71.

——. 1992. The simulation theory: Objections and misconceptions. Mind and Language 7 (1/2): 11–34.

——. 2000. Simulation and the explanation of action. In: Empathy and agency: The problem of understanding in the human sciences, edited by H. Ko¨ gler and K. Stueber. Boulder, Conn.: Westview Press.

HARRIS, PAUL. 1989. Children and emotion: The development of psychological understanding. Oxford: Blackwell.

HEAL, JANE. 1995. How to think about thinking. In Mental simulation: Evaluations and applications, edited by M. Davies and T. Stone. Oxford: Basil Blackwell.

KUNDA, ZIVA. 2002. Social cognition: Making sense of people. Cambridge, Mass.: MIT Press.

MCDOWELL, JOHN. 1985. Functionalism and anomalous monism. In Actions and events: Perspectives on the philosophy of Donald Davidson, edited by E. Lepore and B. McLaughlin. Oxford: Basil Blackwell.

MILLER, J. G. 1984. Culture and the development of everyday social explanation. Journal of Personality and Social Psychology 46: 961–78.

MORTON, ADAM. 2003. The importance of being understood: Folk psychology as ethics. New York: Routledge.

NICHOLS, SHAUN, and STEPHEN STICH. 2003. Mindreading: An integrated account of pretence, self-awareness, and understanding other minds. Oxford: Oxford University Press.

NISBETT, RICHARD, and LEE ROSS. 1985. Human inference: Strategies and shortcomings of social judgement. Englewood Cliffs, N.J.: Prentice Hall.

PARK, B. 1986. A method for studying the development of impressions of real people. Journal of Personality and Social Psychology 51: 907–17.

PERNER, JOSEF. 1996. Simulation as explicitation of predication-implicit knowledge about the mind: arguments for a simulation-theory mix. In Theories of theories of mind, edited by P. Carruthers and P. Smith. Cambridge: Cambridge University Press.

ROSS, L., and R. NISBETT. 1991. The person and the situation: Perspectives of social psychology. New York: McGraw-Hill.

SELLARS, WILFRED. 1956. Empiricism and the philosophy of mind. In Minnesota studies in the philosophy of science. Vol. 1, edited by H. Feigl and M. Scriven. Minneapolis: University of Minnesota Press.

STICH, STEPHEN, and SHAUN NICHOLS. 1995. Second thoughts on simulation. In Mental simulation: Evaluations and applications, edited by M. Davies and T. Stone. Oxford: Blackwell.

THAGARD, PAUL, and ZIVA KUNDA. 1997. Making sense of people: Coherence mechanisms. In Connectionist models of social reasoning and social behavior, edited by S. Read and L. C. Miller. Hillsdale, N.J.: Erlbaum.

VAN FRAASSEN, BAS. 1980. The scientific image. Oxford: Oxford University Press.

WINTER, L., and J. ULEMAN. 1984. When are social judgments made? Evidence for the spontaneousness of trait inferences. Journal of Personality and Social Psychology 47: 237–52.