Knowing Mental States: The Asymmetry of Psychological Prediction and Explanation

Kristin Andrews


Introduction

The Machiavellian intelligence hypothesis promotes the view that cognition, and with it consciousness, is an evolutionary adaptation that arose due to the complexities of social life (Humphrey 1976, Whiten and Byrne 1988). According to this view, we have knowledge of our own minds because we needed to have knowledge of others’ minds. We needed to be able to predict what others were going to do in the future so that we could act to thwart others’ behaviors when they interfered with our own goals. If the Machiavellian intelligence hypothesis is right, our self-consciousness cannot be understood without also understanding our consciousness of others’ mental states.

Some interpretations of the two major theories of how minds understand minds, the theory theory and the simulation theory, conform to this requirement of the Machiavellian intelligence thesis. Though both the theory theory and the simulation theory originated as accounts of how we understand other minds, it has been suggested that both simulation theory (e.g. Bolton 1995, Gordon 1995) and the theory theory (e.g. Frith and Happé 1999, Bolton 1995, Gopnick and Meltzoff 1994, Carruthers 1996) can also provide an account of how we understand our own minds. The classical account of how we know the contents of our own minds emphasized the role of introspection, and it
was thought that we have private and direct access to mental states. However, in recent years a significant amount of evidence about the human tendency towards self-deception, overconfidence, bias, and rationalization has served to undermine the view that we are infallible with regard to our own mental states (Mele 2001, Gilovich 1991, Nisbett and Wilson 1977). We can be wrong about our own mental states (e.g. one might deny being jealous even though it is clear to everyone else that she is jealous), we can make incorrect predictions about what we would do in a counterfactual situation, and we can generate incorrect explanations for the things we have already done (e.g. one might think she hired Smith because he was the best candidate for the job, though her action was considerably affected by her unacknowledged racism). Given this, a number of suggestions have come from the theory of mind literature to suggest that we know our own mental states the same way we know the mental states of others, through appeal to a theory or by using mental simulation. The role played by introspection, given both these accounts, is very limited.

I contend, however, that neither the simulation theory nor the theory theory provide us with an acceptable account of how we understand our minds or the minds of others. Both theories fail due to their dependence on what I refer to as the symmetry of psychological prediction and explanation.

When a person predicts what someone else is going to do next, she demonstrates some understanding of that person. And when a person explains why someone acted in a certain way, she also demonstrates understanding. The same point can be made about understanding one’s self. If a person had no competence in predicting or explaining her own behavior, we would be inclined to say that she didn’t know her own mind.
Perhaps because both explanation and prediction are key components to understanding, philosophers and psychologists often portray these two abilities as though they arise from the same competence, and sometimes they are taken to *be* the same competence. When explanation and prediction are associated in this way, they are taken to be two expressions of a single cognitive capacity that differ from one another only pragmatically. If the difference between prediction and explanation of human behavior is merely pragmatic, then anytime I predict someone’s future behavior, I would at that moment also have an explanation of the behavior. The same mechanism is used to generate both predictions and explanations, and something is a prediction rather than an explanation only if my goal were to predict rather than explain. On this view, explanations are nothing more than ‘backward predictions’ (Robinson and Mitchell 1995), and we can describe this thesis as the symmetry of psychological explanation and prediction.

Following from this line of reasoning, some have thought that any account of understanding must be able to handle both prediction and explanation (e.g. Churchland 1995; 1989; Gordon 1995, 2000; Stich and Ravenscroft 1996; Fodor 1987, 1991). Both the simulation theory and the theory theory are used to explicate prediction and explanation. Stich and Ravenscroft write, “The simulation theory provides a real alternative to the prevailing explanatory strategy in cognitive science [i.e. the theory theory] for explaining our capacity to *predict and explain* other people’s behavior” (1996, 133, my emphasis). Though Stich and Ravenscroft reject the simulation theory as the best account of psychological prediction and explanation, underlying their discussion is the implicit acceptance of the symmetry thesis. The inputs used to predict a behavior,
regardless of whether the prediction is generated through a simulation or through appeal to a theory, become the explanation of that behavior. If the symmetry thesis is correct, whenever we make a correct prediction of behavior, we have at that moment an explanation of the behavior in hand.

If we look at prediction and explanation in the physical world, we see that humans are able to predict that many things occur in a certain way without having an explanation of the phenomenon. I understand that there is a force of gravity, and that dropped objects fall to the ground. I expect that solid objects will not pass through other solid objects, and I expect that the sun will rise tomorrow. I can base these predictions on my experience of the world, and on the things that I have been taught. We would not assume that the method I use to predict the rising sun will provide an explanation why the sun rises every morning. I can make this prediction through statistical induction, however induction will not tell me that the earth spins on its axis, and so it cannot give me an explanation for the rising sun. Prediction and explanation are not symmetrical in our folk physics, so why should we expect things to be any different in our folk psychology?

The symmetry thesis should be rejected in folk psychology, just as it was rejected in the philosophy of science half a century ago. I will argue that psychological prediction and explanation are not symmetrical, and in so doing I will make clear that there is no paradigm case of explanation. Instead, we provide different kinds of explanations given different circumstances. Because there are different sorts of explanations, we can have no clean mapping from explanation to prediction; there is no isomorphism between all cases of prediction and explanation. A major problem with past work in this area is that generalizations have been made about explanation *simpliciter* based on characteristics of
one particular variety of explanation. To fully understand the ability to generate explanations, we must first recognize the different varieties of explanation.

I will also argue that there are different methods of predicting behavior. Contrary to the dominant positions, we do not always need to attribute beliefs and desires in order to predict what a person will do next. Though this may be obvious in the case of predicting one’s own behavior, it may not be as obvious in the case of predicting the behavior of others. In both cases it often suffices to refer to past events, and use statistical induction to conclude that a person will behave in the future the way she did in the past. Because predictions made on the basis of statistical induction will not offer a reason why someone behaved as she did, statistical induction about behavior is not symmetrical with explanation.

The symmetry thesis

Like people in the social world, scientists are concerned with both explaining and predicting the occurrences in the natural world. However, there has rarely been much agreement about what a scientific explanation looks like. For a short while, the symmetry thesis was embraced by philosophers of science with the introduction of Hempel and Oppenheim’s deductive-nomological account of explanation (Hempel and Oppenheim 1948). They argued that an explanation consists of a general covering law and initial conditions, the explanans, which deductively entails the phenomenon to be explained, the expandum. One can also use the explanans to predict an occurrence of the phenomenon, for the expandum follows logically from it. To both explain and predict the
explandum, one needs merely to refer to the relevant initial conditions and general law that deductively entails it.

However, soon enough several counterexamples to the symmetry thesis were found. For one, not all predictions that fit this model acted as explanations. For example, sailors understood that there is a constant conjunction between the phases of the moon and the tides, and they used this law to predict the tides. However, until Newton, no one knew why there was this constant conjunction, and there was no explanation of the movement of the tides. Another celebrated counterexample to the symmetry thesis in scientific explanation is Bromberger’s flagpole. Though the length of the flagpole’s shadow can be explained by reference to the height of the pole, the position of the sun, and the general law of rectilinear propagation of light, the height of the flagpole is not itself explained by these other facts. These facts do not resolve the question of why the flagpole is a certain height. However, this information can be used to deduce, or predict, the height of the flagpole.\(^1\) Thus, not all predictions serve as explanations. The symmetry thesis was further damaged with the realization that not all scientific explanations allow us to make predictions about how things will be in the future. To make this point, Scriven noted that though evolutionary biology explains what has evolved in the past, it does not serve to predict what will evolve in the future. Thus, not only is it the case that not all predictions provide explanations, but it is also true that not all explanations allow us to predict what will happen. Scriven gave us an example of an explanation that does not have any predictive force (Scriven 1959). Once it was realized

\(^1\) For a discussion of many other problems with the symmetry thesis in scientific explanation, see Salmon 1989.
that there are predictions which do not provide explanations, and explanations which do not serve to make predictions, the symmetry thesis was rejected.

However, when we turn to psychological prediction and explanation, it seems that the symmetry thesis is widely accepted. For example, according to Daniel Dennett’s intentional stance, in order for us to predict the behavior of other people (and other intentional systems generally) we must recognize their goals, beliefs and desires (Dennett 1987). From that we can infer how they will behave in the future. Thus, in order to explain why someone acted as she did, we need only to refer to her appropriate goal, belief and desire. For Dennett, reference to the appropriate beliefs and desires serve both to predict and to explain.

Implicit advocacy of the symmetry thesis can also be seen in Fodor’s writing when he speaks of “explanation/prediction” (Fodor 1991, 19), and when he moves from a discussion of “predictive adequacy” to “explanation” in such a way to present them as if they were one and the same (Fodor 1987). In Schiffer’s criticism of Fodor, he notes the requirement of symmetry for the theory theory when he writes “the conceptual roles of our propositional-attitude concepts secure that typically when we move directly from a belief that a person has such-and-such attitudes to a belief that he will do a certain act we also move to the belief that the person will do the act because he has those attitudes (Schiffer 1991, 12-13). It is the ‘because’ that provides the explanation, according to Schiffer.

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 behavior 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, as is the case in prediction, the behavior
is the input, and the theory is used to determine which sets of beliefs and desires can
cause that behavior. Thus, for the theory-theory, prediction of intentional behavior and
psychological explanation are symmetrical.

The simulation theory also advocates symmetry of prediction and explanation.
Somewhat early in the debate between simulation theory and the theory theory, Paul
Churchland (1989) criticized the simulation theory by suggesting that even if it describes
an excellent method for predicting behavior, simulation cannot be used to offer an
explanation of behavior. He argued that an appropriate explanation cannot be generated
until we learn why the simulation model behaves as it does, and the real explanation of
human behavior would come from an understanding of the model’s function. Churchland
thought that for the simulation theory to be a viable alternative to theory theory, it must
also provide for explanations of behavior. This is because one of the simulation theory’s
strengths over theory theory was though to be its relative simplicity. However,
simulation would not be simpler if it were unable to account for explanation, but merely
prediction. Because humans do explain and predict, if simulation only accounted for our
ability to provide predictions, and the theory theory accounted for both prediction and
explanation, any claim that simulation was simpler would have to be abandoned given
that an entirely new account of explanation would be required to exist alongside the
simulation account of prediction. Thus, symmetry of prediction and explanation became
a virtue that any account of folk psychology must demonstrate.
Both Goldman (1995) and Gordon (1995) accepted the requirement of symmetry when they took pains to defend simulation from Churchland’s criticism. They both argued that Churchland was wrong; simulation does provide explanations of behavior by generating a number of possible belief/desire sets that could have caused the action, and thus, like the theory theory, enjoys the virtue of symmetry.

Goldman, for example, writes, "explanation can consist of telling a story that eliminates various alternative hypotheses about how the event in question came about, or could have come about. This is done by citing a specific set of goals and beliefs, which implicitly rules out the indefinitely many alternative desire-and-belief sets that might have led to the action" (Goldman 1995, 89). That is, one generates a number of possible belief/desire sets that could have caused the action and then tests these by using them as inputs in a simulation. When we arrive at a set of beliefs and desires that does produce the behavior in question, we take that set as a possible explanation of the behavior.

Gordon offers a similar account of explanation. He shows how model simulations can explain the behavior of what is being modeled because models are manipulable. If we have a model that can be manipulated until the object behavior is exhibited, we can determine which features interact to cause the behavior. He writes “in explaining one’s own behavior, it would seem that one can—without invoking or using laws or theories—simulate in imagination various counterfactual conditions and test their influence by methods akin to Mill’s…And one can perform such thought experiments not just in one’s own case but also within the context of a simulation of another” (Gordon 1995, 115). Gordon has argued that explanations of intentional behavior 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 which take facts about a person’s situation rather than her beliefs and desires as inputs to the simulation can provide satisfactory explanations. Thus both Gordon and Goldman take simulation theory, like theory theory, to be consistent with the view that explanation is merely backwards prediction.

The requirement of symmetry is strongly advocated by other defenders of simulation theory as well. Boterill writes:

“Perhaps the most commonly urged ground for regarding a theory as implicit in our folk psychological practices is that we use our knowledge of folk psychology to explain and predict the actions and reactions of others. Certainly explanation and prediction are two of the chief functions of theories. Yet this falls well short of providing a compelling case for theory-theory. Give or take the availability of the specific information required for predictive success, we can say: no theory without explanation and prediction. But we should hesitate to assert: no explanation or prediction without theory” (Boterill 1996, 107).

This quote shows the force of the virtue criterion, and also the fallacy: since we use folk psychology to both predict and explain behavior, any viable theory of our folk psychology must provide for both prediction and explanation. The fallacy comes from thinking that the same mechanism must be at work in both cases. Boterill points out the reason why simulation theorists have worked to develop simulation accounts of explanation. His reasoning is that any good theory must both predict and explain, and the theory theory seems to have a natural advantage in the explanation condition. In order to avoid loosing the battle, the responses from Goldman and Gordon were to develop simulation accounts of explanation. However, there is another solution, and this is to deny the symmetry of prediction and explanation.
**Prediction**

According to our commonsense understanding of the mind, beliefs and desires cause our behaviors, and so we can predict the behavior of other people by determining what beliefs and desires they have. Discussions of folk psychology normally begin by noting how successful humans are at predicting behavior. And we are quite good, in many cases. We predict behavior all the time, even sometimes when we’re not aware of it. In most of our daily interactions with people on the street, we are in a continual state of anticipating behavior. When walking down a city street we are confronted with hundreds of people with whom we interact, however briefly. We must dodge people approaching us, and move aside for those who want to pass. Our verbal and non-verbal interaction with other people is a constant dance choreographed by our society and our human impulse to anticipate similarity. If we were not skilled at predicting behavior, we would be paralyzed in the world, unable to even order a cup of coffee.

But how good are we, really, at predicting behavior? How far do these abilities stretch? I can predict, with reasonable accuracy, that if you say you will bring a bottle of wine over for dinner, that you will arrive with a bottle of wine. I may even be able to predict to some extent what kind of wine you will bring. If I tell you that I am preparing steak for dinner, I might think that you will bring a bottle of red wine. If I know that you are a traditional wine drinker, who chooses red for beef and white for fish, then this prediction is easy to make. In making such predictions I rely on my knowledge of your habits, and I need not know anything about your wine-beliefs. Maybe the correct explanation for your behavior is that you believe drinking red wine with beef is socially required, or perhaps you believe that the fruit and tannins of a robust red accentuate the
taste of the meat. Perhaps you want to have the best possible aesthetic experience, so you buy the rare Bordeaux, or maybe instead you buy it with the desire to impress me with your good taste and financial success. I can predict what you will do, if I know your habits, but I may not have a very good understanding of why you buy a certain variety of wine. I could generate any number of possible explanations for your behavior, even if I were able to predict that behavior with great success.

In other cases, it is much more difficult to predict behavior. If you have never been to dinner before, I may be unable to predict what variety of wine you will bring. And if we were walking down the street on the way to the wine store when a mugger confronts us, I would have very little idea of how either of us would behave. We might run, or hand over our money, or try to speak to the potential mugger, or one of us may try a self-defense move. In unfamiliar situations, we have difficulties predicting behavior. If I have never observed a person in a particular situation, such as drinking wine or being confronted by a mugger, I won’t have a good idea of how that person might behave in those respective situations, just as I would have a hard time predicting how I might act in a strange or complex situation.

We are most successful when predicting simple behaviors; that you will do what you say you will do, and that you will behave in the future as you have behaved in the past. When we start trying to predict what someone will do in complex or unusual situation, our success rate drops.

For example, suppose I need to get $100 from the bank. I can go inside the bank, and fill out a withdrawal slip, and ask the bank teller for $100 from my account. Because I’ve done this many times before, and because I was taught that this is normal behavior, I
can predict that the teller will give me $100. I know what the teller is going to do because I know that tellers generally honor withdrawal requests by giving the requested amount of money, so long as the customer has sufficient funds to cover the withdrawal. I don’t need to attribute beliefs to the teller in order to predict his behavior, and I don’t need to imagine being him. I know that the teller will give me the money in the same way that I know that the automatic teller machine will give me the requested funds: they always (or, at least usually) have in the past. And I know it for the same reasons: we learn to expect a particular action in a particular circumstance after having experienced many cases of the conjunction. If I wanted to explain the teller’s willingness to give me $100, I might well appeal to his beliefs and desires. But it doesn’t follow that we need to attribute beliefs and desires to predict his behavior.

The bank teller story is an example of a very simple prediction of behavior, and it is true that we are much more successful when predicting simple behaviors than complex behaviors. Dennett, in defending the intentional stance account of prediction, recognizes its limitations when he writes “[The intentional stance] is notoriously unable to predict the exact purchase and sell decisions of stock traders, for instance, or the exact sequence of words a politician will utter when making a scheduled speech, but one’s confidence can be very high indeed about slightly less specific predictions: that the particular trader will not buy utilities today, or that the politician will side with the unions against his party, for example” (Dennett 1987, 24). However, we wouldn’t need to attribute beliefs and desires to predict that the trader will not buy utilities today if we know that she never buys utilities when they are over $50 a share, even if we have no idea why. In fact, if we were to attribute beliefs and desires to the trader, we must first be familiar with her past
behavior; we need to understand the trader, and understanding will only come with exposure to regularities, either of the individual or of the kind of trader she is. We need to have some understanding of a person before we are able to make correct belief/desire attributions.

We are good at simple predictions, and not so good with complex or detailed predictions, I contend, because we use past regularities to make the large bulk of our predictions. We do not need to simulate or to appeal to a theory of folk psychology when walking down a city street, or expecting a person to pick up a dropped wallet, because we know that these are the sorts of things that we do. And, given that we use fewer variables when we predict behavior using statistical induction about behavior than we do when attributing beliefs and desires, our predictive power when using the inductive methodology is greater. This is not to say that all predictions of human actions can be made through statistical induction, because not all behavior is predictable through appeal to past regularities. We sometimes appeal to our own behavior in order to predict the behavior of others, and use an overt simulation in order to predict behavior, asking ourselves “What would I do if I were him?” And we also frequently predict behavior by attributing beliefs and desires, especially in complex or unusual circumstances. For example, a realtor who is trying to sell a house may try to discover the desires and beliefs of the customer in order to make the most convincing case for buying that particular house. There are many different kinds of behavior that we are capable of predicting to different degrees. It is easier to predict the behavior of the teller in the bank than it is to predict what he would do when confronted with a moral dilemma. It is often easier to predict one’s own behavior, or what one's family and friends will do, than to predict what
strangers would do in the same situation. And it is not always easier to predict the behavior of someone similar to oneself than someone dissimilar. If I saw a racist in a position where she could save only one of two drowning children equidistant from her, I would predict that she would save the child of her own race, and leave the other. If I placed myself in the same situation, I have little faith in any self-prediction I could now make.

My goal in this section is not to resolve the simulation theory/theory theory debate by arguing that we use neither strategy. Rather, I only mean to point out that what most people already accept, namely that we often use induction when making predictions of behavior. Perhaps we also use a folk-psychology theory, or perhaps we also simulate, or perhaps our cognitive mechanisms consist of a theory-simulation hybrid.

However, my point has not generally been accepted. Though Goldman recognizes that we do use induction to make predictions, on his account simulation is prior to induction. He writes “I am not saying, it should be emphasized, that simulation is the only method used for interpersonal mental ascriptions, or for the prediction of behavior. Clearly, there are regularities about behavior and individual differences that can be leaned purely inductively” (Goldman 1995, 83). And later he writes “When a mature cognizer has constructed, by simulation, many similar instances of certain action-interpretation patters, she may develop generalization or other inductively formed representations (schemas, scripts, and so forth) that can trigger analogous interpretations by application of those ‘knowledge structures’ alone, sans simulation” (Goldman 1995, 88). On Goldman’s view, the work done by statistical induction about behavior depends on having previously engaged in successful simulations. Because of this, our ability to
predict through induction is derived from our prior simulations, and simulation remains fundamental (Goldman 1995).

The sort of induction I am talking about is not derivative of simulation or theory or anything else but recognition of past behavior. I can predict that you will buy a double espresso every day at noon, not because I ever had to simulate your coffee-buying behavior, but because I noticed that you have bought an espresso every day at noon, and I have no reason to think today will be any different. I can predict your behavior the same way I can predict the leaves falling from the trees in autumn: it’s always (or usually) been like that before.

For some predictions we make, it is simplest to use the inductive method, and in some cases it will be necessary to use the inductive method. For example, if I am trying to predict the behavior of someone utterly unlike me, such as a visiting alien from another planet, I will not know what his beliefs and desires are, and I will need to study the subject for a while in order to determine what his behaviors are like. Simulation theory should only work with agents who are relevantly similar to oneself, and belief attribution should only work with agents to whom we can reliably attribute beliefs and desires. However, we can become quite good at predicting the behavior of agents quite unlike ourselves. For example, we often predict the behavior of non-human animals, and we can do this after observing them and determining how they actually do behave. We know that vervet monkeys run into the bushes when they hear a certain warning cry, and we can predict that any particular monkey will engage in this behavior in such a situation. We know this because ethologists have observed vervet monkey behavior, and they have seen this pattern of behavior time and time again. However, we don’t know the
explanation for this behavior. For those who are inclined to see animal behavior as
minded behavior, an obvious explanation would be that the vervet monkey doesn’t want
to be eaten by the eagle, or that the vervet is afraid of the eagle. And for those inclined to
think of animals as automata, an explanation may be given in stark behavioral
reinforcement terms, or evolutionary terms. Whichever way one is inclined, a number of
explanations suggest themselves, and it becomes difficult to adjudicate between them.

The goal of this section is to point out that some predictions of intentional
behavior appeal to statistical induction about behavior, and do not take the mentality of
the subject into account, either through simulation or through the use of a theory. It also
seems that the predictions made through induction, because they are simple and involve
fewer variables, are quite reliable. In complex situations, our ability to predict accurately
breaks down, and it is in these situations that the inductive method fails as well.

**Explanation**

Explanation has typically played a less central role in the discussions of folk
psychology. For the most part, empirical studies relating to this area have focused on the
human ability to predict behavior, and accounts of psychological explanation are derived
from those results. Though there have been a few studies comparing our ability to predict
and explain behavior, it seems to me that the varieties of explanation have been glossed
over. There is a distinctive pragmatic nature to explanations that doesn’t seem to be
addressed in this literature.

A few years ago, a number of psychologists presented the hypothesis that
explanation is cognitively less demanding than prediction (Mitchell 1994; Robinson
1994; Robinson and Mitchell 1995; Moses and Flavell 1990; Riggs, Peterson, Robinson and Mitchell 1998; Bartsch 1998). This view was meant to challenge previous findings coming from prediction-based tasks that concluded children do not develop an understanding of other people’s beliefs until about 4 years-old. These psychologists thought that children begin to understand others at an earlier age, and that their understanding derives at least in part from their ability to explain rather than predict human behavior. Various reasons were given for thinking that explanation is simpler than prediction. Moses and Flavell, for example, argued that explanation is cognitively less demanding than prediction because when children see someone behaving in a way that is inconsistent with a desire the child expects the actor to have, that anomalous behavior will serve as a clue to the actor’s belief (Moses and Flavell 1990). However, we don’t only offer explanations when a person behaves unexpectedly. Sometime we need to offer an explanation when trying to show that the behavior is justified, or to help someone who doesn’t know the person understand him better. Moses and Flavell’s reason for thinking that explanation is cognitively simpler is only true for cases of explaining anomalous behavior. They do not show that explanation as a whole is simpler than prediction.

It seems that we need to better understand the activity of psychological explanation before drawing any conclusions about the relative difficulty of prediction and explanation, or about symmetry of prediction and explanation. Explanation of intentional behavior involves a complex set of abilities, and appropriate explanations require an understanding of what information is being requested.
The criticisms of Hempel’s D-N model of explanation and later criticisms of the inductive-statistical models of scientific explanations gave rise to a growing concern that scientific explanations are not arguments. It seems that there are similar problems with the view of psychological explanation as arguments, given that arguments have been unable to deal with the pragmatic issues that are involved in the generation of good explanations. Explanation of behavior is in some sense less an objective phenomenon than prediction of actual behavior. To judge a prediction as correct or incorrect one has an objective criterion—the actual behavior. I can easily verify my prediction of someone’s behavior as long as I’m there to observe it. Explanations are not as easy to verify. The explanations that refer to a person’s psychological state will be unverifiable in principle. I can’t be certain that George W. Bush really wants to be president, even though he says he does, but I still use this belief to give a partial explanation of why he ran for the office.

There are varieties of explanations for one's behavior; some are psychological and refer to propositions or desires, and other may refer to historical or environmental circumstances. Suppose I ask, “Why did Kurt kill himself?” You might answer, “His wife had just left him” or perhaps “He had just gotten out of drug rehab” or instead you might say “He was clinically depressed.” None of these explanations refers to Kurt’s beliefs or desires. The last explanation gives a medical diagnosis that refers to his brain chemistry, and though it may help us understand why he killed himself, it is not psychological in the sense we are interested in here because it does not refer to Kurt’s specific mental content. The context of the question might help decide which of these explanations would be acceptable at a given time. Many different sorts of explanations
could be appropriate, and often no one is better than another. However, some explanations would not be satisfactory at all. Suppose you answered my why-question by saying “Because he shot himself in the head.” That wouldn’t satisfy me at all, though this action does give a sort of explanation.

The same point can be made about the explanations one generates for his own actions. Suppose a student is asked why he is taking Professor P’s philosophy class. The student might refer to his beliefs: “I think P is a good teacher” or “I am very interested in philosophy.” Or he might give other explanations such as “It fit into my schedule” or “My advisor told me to”. Given that these explanations are consistent with one another, they might all be accurate explanations for the student’s behavior.

Van Fraassen, recognizing the pragmatic nature of explanation in the sciences, proposed the following account of an explanation: “An explanation in not the same as a proposition, or an argument, or list of propositions; it is an answer” (Van Fraassen 1980, 137). More specifically, an explanation is an answer to a why-question. Thus, an account of explanation in the sciences must be an account of why-questions. Van Fraassen gives a story that undermines the universality of Bromberg’s criticism of the D-N model of scientific explanation, because in some cases the length of the shadow can explain the height of a tower: suppose a mad Chevalier wanted to build a tower to commemorate an ill-fated relationship he had with his maid. He erects a tower at the spot where he killed her, and makes it tall enough so the shadow will cover the place where he first declared love to her. Given this story, we can, contrary to Bromberg, explain the height of the tower by referring to the (in this case necessary) length of the shadow (Van
Fraassen 1980). Thus, explanations are not context-independent, but the correctness of the explanation depends on the context in which the question is being asked.

Let’s apply Van Fraassen’s ideas to psychological explanation. Suppose I use theory theory to predict that John will hide in a closet when he sees a person in the house. I make this prediction because I know that John is afraid and he wants to feel safe, and he feels safe in the closet. I can explain why John hid by referring to his belief and desire (that he thinks there is a burglar in the house, and he doesn’t want to be shot), but if the why question is “Why did John hide rather than attack the intruder?” the reference to the belief and desire I used to make the prediction does not answer the question. Nor would that response be a good answer to the question, “Why did John hide to avoid the burglar rather than run out of the house?” To give a good explanation, I need to know what sort of explanation is called for. If John had told me that he hides when he is frightened, or if I had seen him hide in another instance when he was afraid, I may be able to predict that he will hide in this case. However, generating this explanation does not give me any understanding of why he hides rather than runs. Unfortunately for those who seek symmetry in folk psychology, not every appropriate explanation will refer to those mechanisms used to make the prediction.

This point may be clearer for theory theory than simulation, for most versions of the theory theory do seem to accept the covering law model of prediction and explanation. The symmetry in theory theory is explicit; those features I use to predict can also be used to explain. However, the same problem arises for the simulation theory. The explanation of John’s behavior is formed by generating reasons why John would hide in the closet given that there is a burglar in the house. Perhaps I know that John has
certain beliefs and desires, and when I input these into the simulation, I see that he would hide in the closet because he was afraid of the burglar. Or perhaps I know that John will hide in the closet simply because there is a burglar in the house. The same problem of pragmatics arises regardless of whether you start the simulation with beliefs and desires or facts about the situation. If I explain John’s behavior by saying that he was afraid, I don’t have an answer to all why-questions one might ask about John’s behavior. Perhaps I don’t understand why John is afraid of the burglar, given that he is a brave fellow who has demonstrated his bravery time and time again. My why-question might not be satisfied by reference to John’s fear, because I want to understand why he feels fear. An explanation in terms of the situation falls prey to the same problems. If I didn’t know there was a burglar in the house, an appropriate explanation would be “Because there is a burglar in the house.” But, of course, this would only be a good explanation if there really were a burglar in the house; if the noises John had heard were caused by a squirrel, the appropriate explanation would be neither “Because there is a burglar in the house” or “Because there is a squirrel in the house.” To be a good explanation, we would have to refer to John’s mistaken belief about there being a burglar in the house. And if there really were a burglar in the house, the explanation of John’s behavior that cites this fact wouldn’t explain why he hid rather than doing something else. One sees symmetry of prediction and explanation when looking at certain kinds of cases, but because there are many kinds of why-questions, not all of them will be answered through reference to the inputs of the simulation that allowed the behavior to be predicted.

This point follows from Churchland’s original criticism of the simulation theory’s ability to offer explanations. It is true, given Goldman’s and Gordon’s responses, that
simulation can answer some why questions. But Churchland pointed out that simulation cannot offer explanations of why one has those beliefs and desires, or how the situation causes one to act in a certain way. Simulation doesn’t provide information about the causal interaction of the inputs of the simulation and the behavior, and because of this certain why questions such as “Why did John feel afraid now when he had never been afraid before” may not be answered through appeal to simulation.

The asymmetry is also present when I use a statistical induction about behavior to predict what someone will do. Suppose John lives in a bad neighborhood, and his house has been broken into many times while he was home. If in the past he had always hid in the closet, I can predict using statistical induction that John will hide in the closet again this time. However, I will need to appeal elsewhere to develop a plausible explanation. Most of us won’t be satisfied by the explanation “He always does that,” for when we ask for a psychological explanation we often do want to know what beliefs and desires caused the behavior we’re interested in explaining.

A defender of symmetry might respond by saying, “You predicted that John would hide because he thought there was a burglar in the house. If someone who didn’t know about the burglar saw John hide, her question ‘Why did John hide?’ would be easily answered with the response ‘John thinks there is a burglar in the house’. And I would agree with this. I am not claiming that reference to a belief or desire than one uses to generate a prediction will never be the correct answer to a why question. Rather, I am pointing out that some why questions cannot be answered in this way, and because of this, the perceived symmetry between prediction and explanation is an illusion.
Symmetry exists between two things if it exists in all instances; if they are partially symmetrical, or sometimes symmetrical, then they are not symmetrical at all.

**Asymmetry of prediction and explanation**

I will introduce one last example to make the asymmetry of prediction and explanation clear. Kelly has just gotten home from work after a long day of fixing teeth. The three theories can produce the same prediction of Kelly’s behavior, but note the different explanations for her next action.

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Theory theory</th>
<th>Simulation Theory</th>
<th>Statistical Induction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kelly gets a beer from the refrigerator.</td>
<td>Kelly gets a beer from the refrigerator.</td>
<td>Kelly gets a beer from the refrigerator.</td>
<td></td>
</tr>
<tr>
<td>Explanation</td>
<td>Kelly thinks there is a beer in the fridge and desires a beer.</td>
<td>Kelly thinks there is a beer in the fridge and desires a beer. OR Kelly just got home from work.</td>
<td>Kelly always gets a beer from the fridge when she gets home from work.</td>
</tr>
</tbody>
</table>

The reference to Kelly’s belief and desire doesn’t provide a very full explanation of why she gets the beer. We have no information about the intention that leads her to drink a beer. Maybe she wants to relax after her hard day, or maybe she wants to forget about work, and intends to drink until she is drunk. If Kelly wanted to stop drinking, and went to a councilor to help her stop, none of the explanations given above provide a reason why she was drinking after work, and without that understanding the councilor might not be able to help her.

Due to the pragmatic nature of explanation, it is difficult to imagine any one-to-one correlation between a prediction and an explanation. An explanation can be a request
for the beliefs and desires one had before acting, or the immediate causal story, or for a cause of the tendency being exhibited. Not all explanations of human behavior will refer to mental events, as Gordon has made clear (Gordon 2000), though theory theory and some accounts of simulation theory take mental events as an integral to both prediction and explanation.

The case for symmetry becomes even weaker if you accept that many of our predictions are made by statistical induction about behavior. If it is true that we do often appeal to simple regularities to predict, then the symmetry thesis is certainly false, for inductive generalizations do not usually provide explanations. However, in some cases a generalization about behavior will be all the explanation there is. For example, you meet a fellow who often twirls his beard, and you ask his old friend why he twirls his beard. The friend might respond, “Oh, it’s just a habit.” This explanation, though it does not make reference to mental states, may be satisfactory, because it indicates that there are no beliefs or desires associated with the behavior. There may be no further psychological reason for the behavior.

If I answer a why question by stating that things are always like that, I am either reporting on my epistemic state, or I am claiming that there is no further explanation. Explanations, like justifications, must come to an end somewhere, and there are some things that we have no further explanation for. Perhaps the final explanation may refer to the laws of nature, and if you ask me to explain why the laws of nature exist as they do, you are asking a question that I cannot answer if we understand laws of nature as necessary.

I have argued that there are different kinds of psychological explanations, and that
one method we use to make predictions, statistical induction about behavior, will often not automatically generate an explanation for the behavior. Psychological prediction and explanation is only symmetrical if we use the same mechanisms to predict and explain behavior, and if the act of making a prediction automatically results in an explanation.

Churchland’s observation that simulation theory is not able to account for psychological explanations should not be taken as a criticism of simulation theory, because we should not expect symmetry between our abilities to make predictions and our ability to provide explanations. Only after it has been accepted that these are two distinct activities can the investigation into our understanding of other minds proceed.

This conclusion not only undermines both theory theory and simulation theory as accounts of how we know others’ mental states, but it also demonstrates that neither theory can provide a plausible account of how we know our own mental states. Simply knowing that I will do something does not help to illuminate the reasons I might have for acting as I do, nor need this knowledge provide any indications of my true beliefs and desires.


