# Interpretative Explanations

G. F. Schueler

Commonsense explanations of actions, in terms of the agent's reasons, hopes, desires and the like, are on their face frequently teleological in form. They specify the goals, purposes or points of the things we do. In this they seem sharply different from other sorts of commonsense explanations of events, as well as from the sorts of explanations found in sciences such as physics and chemistry, all of which are causal, and of course not teleological. But actions are often simply constituted by events involving the agent of the action. And these events are obviously open to causal explanation as long as we describe them in terms of their physical or chemical makeup. So there is a puzzle here. How can commonsense explanations of actions, which are apparently teleological and hence not causal in form, actually explain these actions?

In this paper I will argue that what I will call 'interpretative explanations' are both central to explanations of human action and irreducibly different in form from other commonsense explanations of events, as well as from explanations found in paradigm 'hard' sciences such as physics. If this is right it turns out that, as a consequence of this different form, it is a mistake to think that interpretative explanations are somehow reducible to (or explicable in terms of) causal explanations. What I mean by an 'interpretative explanation' will be brought out in the course of the discussion. But we can start with an example.

## 1.

We sometimes misinterpret what others are doing. Many years ago Andy Griffith did a comic routine where he described something he had witnessed on a college campus. Two groups of students, each dressed in colorful costumes, were performing some sort of ritual in a cow pasture. Each group would have a short meeting to discuss and vote on some topic, and then the ones selected to present the conclusions of the group would line up facing the other group. After a brief moment of silence, one person on each side would yell out its opinion and then a fight would break out which had to be broken up by people in striped shirts. Then the whole thing was repeated. The title of Griffith's piece was 'What It Was, Was Football'.[1]

Griffith was just being funny, of course, but the possibility of misunderstanding in this way is a real one. Finding out that 'what it was, was football' would explain the events on the field to a foreigner who really was unaware that this was what was being witnessed, in a perfectly ordinary sense of 'explain'. It is that sense of this term that I will say involves giving an interpretation and that this paper will explore.

In his routine, Griffith describes the actions of the players as if he doesn't know that they are playing football, but he knows they are doing something. That allows him to pretend to understand the players as performing intentional actions, just not the ones characteristic of football. He pretends to misinterpreted what they are doing. But we can

imagine an observer who does even worse than that. Suppose that the observer is unaware not just that it is a game that is being played but even that the organisms she is watching are performing any intentional actions at all. She is, lets suppose, an alien from outer space (flown in especially to work in philosophical examples) who sees the events on the field simply as very complex interactions of some of the local fauna.[2] Of course these events really are complex interactions of some of the local fauna. So this won't prevent her from describing with complete accuracy, and to any level of detail her observational powers allow, everything that happens on the field. It is just that she won't describe them as intentional actions. She won't interpret what she sees in this way.

This suggests that there are at least two rather different kinds of mistakes one could make here. In the case satirized by Griffith the observer sees that he is observing people who are performing intentional actions. He simply fails to realize what actions they are performing. But one might also make the more serious mistake of not realizing that intentional actions were being performed at all. This would be to understand the behavior being observed in the way we often look on the behavior of lower animals, insects for instance: complicated behavior produced by complex brain responses to the environment but not intentional actions. If that were the only correct way to look at behavior, as some philosophers have held, it would follow that the mistake satirized by Griffith would not be any more mistaken than any other interpretation. If absolutely no intentional characterizations correctly apply to anything, then those students on the field are no more playing football than they are having brief discussions and then fighting with each other. On such a view both characterizations of what is going on are equally mistaken. Rather than pursuing this issue now[3], however, I will start by assuming the reality of the mistake satirized by Griffith, where the form of the mistake seems to be that the observer misinterprets the actions she is observing while realizing that they are indeed intentional actions.

So what would have gone wrong if an observer, seeing what is in fact a football game, takes it as some sort of ritualized debate followed by fisticuffs, in the way Griffith pretended to? Some of the errors Griffith pretended to make can just be set aside. We need to distinguish errors of interpretation from those based on mistakes about the underlying facts being interpreted. Here is an example. Suppose I am at what seems to me a very boring party. I manage to catch the eye of my wife, who is across the room, and she gives me the sort of 'rolling back of the eyes' look that I take to mean that she can hardly wait to leave. So I invent an excuse to give the hosts and drag her away. Once we are out the door though she is incensed; she was having a great time. I was mistaken in thinking she wanted to leave.

One of two things might have happened. It could be that she rolled her eyes all right but she wasn't thereby signaling that she wanted to leave. (Maybe she just at that moment noticed the chandelier above her head.) The other thing that could have happened is she didn't roll her eyes at all. A trick of the light only made me think she had. It was not that I misinterpreted what I saw. Rather I did not see what I thought I did. This second sort of error, where she did not in fact roll her eyes, is not an error of interpretation on my part but a factual error about what I saw. The first sort of error though was an error of interpretation.

The most straightforward way to draw this distinction is by saying that the first sort of error involves misattributing at least one intentional state, such as my wife

meaning something by rolling her eyes, while the second sort need not. The second sort might involve only misattributions of non-intentional states, such as whether her eyes moved in a certain way. Our outer space visitor, who never attributes any intentional states to the objects she observes on this backward planet, might still make no mistakes of the second sort. Depending on her observational powers, she might be completely accurate in her description of non-intentional states, properties and the like.

As I described Griffith's story, it involves lots of errors of the second, non-interpretative sort. The football itself for example, doesn't even get mentioned.[4] So to have an example of a purely interpretative mistake of the sort I want to discuss we will either need to do some re-working of Griffith's story or just use another example, such as my misinterpreting my wife's rolling of her eyes, or perhaps Wittgenstein's example of a set of yells and foot stampings, performed by members of some foreign culture, which can be interpreted as moves in a chess game.[5] I am just going to assume here that at least sometimes all the non-interpretative mistakes can be eliminated by adjusting the alternative story. That is I am going to assume for now that there can be purely interpretative mistakes.[6] The question I want to ask is what has gone wrong when the observer makes such a completely interpretative mistake, that is, where she gets none of the underlying facts wrong but still misinterprets what is going on.

An interpretive mistake of this sort will at least involve misattributions of some intentional states to the people on the field. For two teams to be playing a game of football, the players must have many of a very large but indefinite set of intentional states. Similarly, for two groups to be engaging in a ritualized form of debate which involves short statements of position followed by fights, a very different set of intentional states is required. In specifying that only interpretative errors are involved though I am supposing that none of the 'underlying' physical states, movements and the like have been mistaken by the observer. So none of the things our space alien observes, such as the movements the players make, or the sounds that come from their mouths, are in dispute between the correct interpretation and the mistaken one. Though what actions these movements constitute and what these sounds mean will be of course different in the two interpretations. What the football interpretation holds is the quarterback calling signals, for instance, the ritualized-debate interpretation presumably will have to say is some sort of reference to a text or debate position.

So I am assuming that the two competing interpretations are consistent with, and intended to be based on, exactly the same set of 'underlying' facts, events, states of the players, etc. Of course while much of each interpretation will involve assigning different intentional states to the people involved, there will also be other intentional states of the various agents that are the same in each of the two interpretations, such as beliefs about the color of the grass. That is, both interpretations will assign them (though of course not the space alien, who doesn't assign any intentional states to the objects she sees). But the point is that the various beliefs, actions, and thoughts ascribed to the players, coaches and officials by each of these two interpretations will be claimed to supervene on the same set of underlying facts, which will include only movements, sounds, and the like.[7]

2.

I will explain below why I think this is not problematic, indeed not even uncommon, that is, why it is not always the case that a mistake about one or another

underlying fact will serve to distinguish the correct from the mistaken interpretation. But first it might be worth examining whether what I am assuming violates the principle of supervenience as philosophers have used it. Even if that were true I can't see that it affects the argument I want to make, but in any case it is not true.

To say that one state supervenes on some other states, as when a mental state is claimed to supervene on some physical states of the brain, is to say that there can be no difference in the supervening state without some difference in the underlying states on which it supervenes. 'A set of properties A supervenes upon another set B just in case no two things can differ with respect to A-properties without differing with respect to their B-properties.'[8] Someone might think that this means that the elements of the football interpretation and the elements of the ritualized debate interpretation could not possibly be held to supervene on the same underlying facts. Since the first interpretation is correct and the second incorrect, the thought would be, there must be some difference in the underlying facts that distinguishes the two. But that would be a mistake.

It is true that there can't be two sets of events which are exactly the same in all relevant respects but one of which is correctly described as a football game and the other of which is not. But we are not dealing with two (correctly interpreted) sets of events here, only one set, interpreted in two different ways, one of which is mistaken. To see that this difference is important it might help to recall that a claim of supervenience is not the same as a claim of entailment, or indeed of <u>any other</u> regular connection such as would hold if the underlying properties were connected to the supervening property by a scientific law. A claim that one set of properties supervenes on another set is merely a claim about a certain relation between those sets of properties. It says nothing about why this relation holds. As Kim says at one point, the mere fact of 'supervenience leaves open the question of what grounds or accounts for it…' 'Supervenience is not a metaphysically deep, explanatory relation; it is merely a phenomenological relation about patterns of property covariation.'[9] If there is a nomological connection, or even a logical entailment, between the underlying and supervening properties, then of course that would be explanatory as well, but such connections go beyond mere supervenience.

And if there is no such definitional or nomological connection between underlying and supervening facts, the mere claim that the one supervenes on the other carries with it no requirement that denying a supervening fact one must deny one of the underlying facts. The requirement is there, when it is, only because of the connection that explains the supervenience, not because of the supervenience relation itself. On exactly these grounds, I want to claim that *so far as the relation of supervenience goes* someone can without logical or nomological error deny, for instance, that what she is observing is a football game and yet accept all the underlying facts on which its being a football game supervene. If she is making an error, which in this case she is, it need not be that error.

It might help to take a different sort of case, one where it seems clearer that there really is no logical or nomological connection between the underlying facts and the supervening one. So suppose that you and I both find ourselves in court, facing the same judge, charged with the same crime. Discussing our cases, we discover that the various circumstances of our crimes are exactly the same in all relevant respects. Each of us is charged with doing something unfair to a student, lets suppose, and it turns out to be exactly the same sort of thing in exactly the same sort of class to exactly the same sort of

student (etc.). Your case is called first, all the relevant facts come out, and you are found not guilty. My case is next, all the relevant facts are brought out again, but I am found guilty. Considerations of 'cosmic' or 'poetic' justice aside, something must have gone wrong. The judge has been inconsistent. If all the relevant facts are the same in both cases then either both of us have been unfair to our student or neither has been. Fairness and unfairness supervene on the facts. There cannot be a difference as to the fairness of how we treated our students without some difference in the relevant facts of our two cases.

Notice however that this tells us nothing about whether what you and I have done is actually unfair. The fact that fairness and unfairness supervene on the facts, and that the facts are the same in each case, entails that either we both treated our student unfairly or that neither of us did. But nothing in this says which it is. The judge would have been consistent, and not violated any consideration of supervenience, whichever decision she had made, as long as she decided both cases the same way. By the same token, two judges, both looking at exactly the same set of underlying facts, and in complete agreement as to what those facts are, can still disagree as to whether the correct interpretation of the law and of the applicable principles of fairness require a verdict of 'fair' or 'unfair' in our two cases. The judge hearing our cases is mistaken about one case and since the facts of our two cases are the same what makes the one ruling mistaken and the other one correct cannot be a mistake about any of those facts. That is what I am assuming to happen between the football and ritualized debate interpretations of the events Griffith witnessed. Some have held that the supervenience of the moral on the non-moral, and the resulting possibility of such disagreements in moral evaluations, argues for non-cognitivism about the ascriptions of moral concepts.[10] But that doesn't seem at all plausible if, as I am claiming, exactly the same thing applies to football games.[11]

This assumption about the two interpretations by itself yields an interesting conclusion, which is part of the reason it will be worth looking at it more carefully below. Since the underlying states and events will of course be held to interact causally in exactly the same way under both these interpretations, the difference between the two interpretations – what makes one true and the other false - cannot be any causal factor, any more than it can be a physical or chemical one. Just as both interpretations will be consistent with exactly the same number of people on the field, the same colors of clothing, and the like, so both will be consistent with, because they will be claimed to supervene on, exactly the same set of underlying causal relations among the various events that take place. What the ritualized debate interpretation sees as part of a fight, the football interpretation will see as tackling the tailback for a three-yard gain. But the causal interactions between the events involving the participants will be the same under each interpretation.

So the picture is this. We have two completely different interpretations of exactly the same set of underlying facts. One, the correct one, says that a football game is in progress. The other one says that it is a ritualized series of debates, each of which is followed by fighting. We are supposing that there is no disagreement at all about the underlying facts. And my claim is that, given all this, since the elements of both interpretations will be claimed to supervene on exactly the same set of underlying facts, the difference between the correct interpretation and the incorrect one cannot be found in

those underlying facts. In particular it cannot be found in any of the causal relations included in those underlying facts. The elements of both interpretations are claimed to supervene on exactly the same set of facts, which include exactly the same causal interactions between the same events, etc. So whatever it is about these two interpretations that makes the one correct and the other incorrect, that is not where we will find it.

3.

Of course, at this point I am really only assuming this is possible in this case, even though it seems a plausible assumption. Still, how could it be so? It will help to contrast the interpretative explanations we are discussing with a situation where it is not so, that is, where a different explanatory theory requires some difference in the facts on which it is based. So consider the difference between two theories supposed to describe the motion of some object through space. Suppose we are technicians looking through the records of radar scans taken on some remote island, covering some part of the sky for the last few minutes. The radar is part of an environmental monitoring program and our job is to check these records. We notice markings indicating that the radar has detected something, but we don't know what. Maybe it is a weather balloon, maybe a rocket, maybe only a bird flying in front of the radar. That is what we need to figure out. At first we have recorded only a relatively small number of observations of whatever this object is, four or five. So all we really know is the object's position at those times. On the basis of these observations we formulate two theories of this object's motion based on what it might be, theory R, that the object is a rocket, with a smooth path (which, unknown to us, is correct) and theory B, that it is a bird, with a much more erratic path.

Since both these theories are consistent with all the observations of this object that we have when we start trying to figure out what it is, there is no evidence from these observations that supports theory R over theory B, or vice versa. So in that respect these two theories are analogous to our football v. ritualized debate interpretations of what is happening on that football field. At the same time, both theories R and B will of course be 'under-determined by the data' which supports them. Both theories make far more predictions about the position of the object in question than anyone has yet actually checked or, really, could ever check, since there will never be more than a finite number of observations and each theory makes predictions for the positions of the object at every point in time, not just the times when actual observations are made.

To see which theory is correct we have to look at the predictions each theory makes about the as yet unobserved positions of each object. Both theories are consistent with all observations so far. But the predictions the two theories make about where the object will be observed apply to all possible observations, not just the ones already made. When a new observation is made, say by making another pass with our radar, the position of the object is consistent with the predictions of theory R but not those of theory B. Of course it will always be possible to add a new feature to theory B, an 'epicycle' for instance in which the alleged bird flutters into just that position, which adjusts it to the new observation. The resulting, adjusted theory (B2) will once again be consistent with all the observations yet made but it will have the same fate as B when yet another observation is made. And then it too will have been refuted, or at least it is no longer consistent with the observations. This process can continue, of course, but if we stick

with any one of these theories, we keep finding we have to abandon or revise it as soon as new observations are made.

Contrast this with our two interpretations of what is happening on the football field. Here too both interpretations are under-determined by the data, that is, there are lots of facts that are 'brute' relative to these interpretations but that have not yet been observed. Each interpretation makes predictions about what will happen, say, in the next minute. But, if I am right, in contrast to the two theories of the object detected by our radar, none of these facts, either the ones already checked or the ones 'predicted' by the two interpretations, need be inconsistent with either interpretation.

How can that be? I suggest that it is because, unlike our two theories of the moving object, neither of the two interpretations of the events on the football field is completely determinate with respect to the underlying facts on which it is based. Both the theories of the object moving in front of the radar and the interpretations of the events on the football field are under-determined by the data that supports them. They make predictions about much more than has yet been observed. But, unlike the theories of the moving object, the interpretations of the events on the football field are indeterminate with respect to some of the underlying facts on which they supervene. Each leaves lots of possibilities open, even for the underlying facts relevant to the interpretation. That is quite different from the two theories of the moving object. Both theories R and B make predictions about the exact positions of the moving object for every point in time. That is because, whether that object is a rocket or a bird, there will not be any 'gaps' in its movement. Because of that, when conjoined to the initial observations about the positions of the object, each theory entails that the object will occupy a specific portion of space. And so each will entail specific, though of course different, claims about what observations will be made. In short, within the parameters of the theory, each of these theories is determinate in its predictions. For each area of space, each point in time, etc. each theory either predicts the object occupies that part of space or that it does not.

Nothing like this is true for the two contrasting interpretations of the events on the football field. Even though each will claim to supervene on the same set of actual underlying events, each interpretation is indeterminate within a range of possible events. Each leaves plenty of things 'open'. For each interpretation there are lots of underlying facts, relevant to the interpretation, which can either obtain or not without affecting the truth of the interpretation. If that sounds mysterious, think of how many open choices there are for those participating in either a football game or a debate tournament. Whether the team on offense calls a running play or a pass play, whether the player with the ball cuts to the left or the right, whether the defense rushes all its linebackers or drops them back in pass coverage, it is still a football game between two specific teams, etc. Similarly for the sort of debate tournament we are supposing for the alternative interpretation. Which specific debate position gets supported by the vote of the team members, for instance, would be a matter of how the members choose to vote.

So for each of these two opposing interpretations there will be plenty of underlying facts, facts on which the interpretation in question supervenes, with respect to which the interpretation is indeterminate. Whether the quarterback decides to run or throw the ball, whether the receiver gets tackled or manages to score, it is still a football game. So there is a difference between saying that some theory is underdetermined by the data and saying that an interpretation is indeterminate. Being underdetermined by the

data just means that the implications of the theory go beyond the evidence for it. The theory entails claims about the world for which as yet there is no evidence one way or the other. This is as true of both the two sets of cases we have looked at, the theories about the moving object and the interpretations of what is happing on that field. In saying that the two interpretations of what is happening on that field are indeterminate, however, I am saying something different. An interpretation is indeterminate if so far as the interpretation is concerned the underlying facts being interpreted can be of various different sorts without being evidence against the interpretation. Suppose you have a complete physiological theory of how human bodies work. In order to be complete your theory will have as a consequence the proposition that under some circumstances the muscles in the running back's legs will cause him to move to the right rather than the left. If in the course of a football game these exact circumstances obtain for some running back and yet he moves to the left, your theory will be refuted, or at least have significant evidence against it. Like theory B in the radar example, it will need some revision. But the interpretation of these same events that says that this is a running back carrying the ball in a football game has no such problem. It is indeterminate as to which way the running back moves.

Of course to say that an interpretation is indeterminate is not to say that anything goes. It would be better to say that an interpretation specifies a range of possible facts, with things inside that range consistent with the interpretation, things outside inconsistent with it. If it is a football game, ball carriers can run to the left or to the right but they can't sit down and start working crossword puzzles. But it can still turn out that each of two interpretations of some specific set of events leave open underlying facts within some range and that the actual events at issue fall into that range for each interpretation. If that happens, then whatever exactly these facts turn out to be, they are consistent with both interpretations. That is what we are assuming for the two interpretations of what happens on that football field. That is why it is unobjectionable to assume that both the correct, football, interpretation and the incorrect, ritualized debate, interpretation can agree completely about the underlying facts on which each supervenes. Both interpretations can supervene on the same set of underlying facts because, as we can put it, their ranges of indeterminacy happen to overlap in such a way that the actual sequence of events on the field falls within both.

That won't always be the case with any two interpretations. Saying there is indeterminacy in interpretations doesn't mean that nothing falls outside the range of indeterminacy. If that were true then every interpretation would be consistent with every possible set of underlying facts. Suppose that Griffith, instead of interpreting what he saw on the field to be a ritualized debate tournament, had thought he was witnessing a horse race. Horse races are indeterminate in the same sense football games or debate tournaments are since jockeys can maneuver their horses in different ways, for instance. But for these two interpretations it is hard to see how the ranges of indeterminacy could overlap. There will be some underlying facts that are allowed by one interpretation and not by the other. Interpreting some set of events as a horse race, for instance, is not consistent with a complete lack of horses, though that is allowed by a football game interpretation. So if Griffith had thought he was witnessing a horse race rather than a football game, his mistake could have been traced to a mistake about this underlying fact.

4.

In the football example the indeterminacy arises from the fact that the events being interpreted involve groups of people and numerous choices on the part of those involved. At those places where a choice is possible for someone, each interpretation allows alternatives, each of which is consistent with the interpretation. That is why the underlying facts are neither nomologically, nor definitionally, connected to either interpretation. But it would be a mistake to think that indeterminacy only arises where the events being interpreted involve groups of people. This indeterminacy is equally characteristic of any explanation that appeals to an agent's reasons for doing whatever she did. Explanations of actions in terms of agents' reasons are also interpretative explanations in my sense. The football example is only a special case that happens to involve more than one person. To see this, consider cases where the agent needs to make a choice but can see no reason for choosing one way rather than another (so-called Buridan cases).

Suppose I am running some evening, being chased by some bad guys, and I come to a fork in the road. I can see no reason for going left rather than right or vice versa. Still, I need to keep running. I don't want to get caught. So I just make a choice and go, lets say, left. Clearly in this situation turning left is something I do intentionally but it seems false to say that I have a reason for doing it. I might neither have, nor think I have, reason to go left rather than right, though I have reason for continuing to run. And to do that I must go one way or the other. So I have a reason for choosing one or the other direction. But though I intentionally turn left, it is not true that I have a reason for turning left rather than right.

So not all intentional actions are done for reasons. Explanations of actions in terms of the agent's reasons do not cover everything agents do intentionally. There are some choices one makes, and sometimes in fact must make, where one doesn't oneself think one has a reason to choose one way rather than another. And Buridan cases of this sort are common. Most cases of doing things for reasons 'contain' intentional actions of this sort. When I turn down the road to the left I am of course doing <u>something</u> for a reason. I am running away from those bad guys. But that would also have been my reason had I turned down the road to the right. My action of 'running away from the bad guys' itself involves other intentional actions some of which, like turning down the road on the left rather than the one on the right, involve choices between different things which are, relative to my goal, equally 'reasonable'. And in all such situations the choices of each of those things are typically not done for reasons. Many of the so-called 'basic actions' by which one performs the (less basic) actions which one performs for reasons are still intentional actions.[12] But there are often numerous possibilities and for the most part the choice of one of these rather than another is not something one does for a reason, like the choice to turn left rather than right.

So explanations of actions in terms of the agent's reasons are frequently indeterminate in the same way an interpretation of those events on that field as a football game is indeterminate, and for the same reason. In each case the explanation (or interpretation) is consistent with various choices on the part of the agents involved. Within some range these choices can go in quite different ways and still be consistent with the interpretation in question.

All this argues that the same conclusions about the possibility of an alternative interpretation supervening on the same set of underlying facts can be drawn for explanations of actions generally that we saw followed for the interpretation of the football game. Suppose Andy Griffith sees me running, turning left at the fork in the road, and interprets what he sees as just another jogger, out to get some exercise. He would have misinterpreted what he saw. What I am really doing is running away from those bad guys. So here again we would have two interpretations, each of which is (or at least could be) consistent with the underlying facts on which its elements supervene, because each is indeterminate with respect to numerous open choices the agent in question can make, including the choice of whether to turn left or right at that fork in the road. This means that, as before, the difference between the correct and the incorrect interpretation may not be found in any of the underlying facts on which the elements of these two interpretations supervene. In particular it may not be found in some causal connection which one interpretation uses or presupposes and the other does not. The difference between the two interpretations is not 'causal' in this way since each interpretation might supervene on exactly the same set of underlying facts, including facts about the causal connections among the various events involved, such as the muscle contractions in my legs that propel me to the left rather than the right when I arrive at that fork in the road.

5.

An obvious question remains. The correct and incorrect interpretations are both consistent with the same set of underlying facts and yet one is correct and the other not. How can that be?

The answer, I suggest, is that what the correct interpretation includes, and the mistaken one misses, is the actual point or purpose of what the agent or agents are doing. Consider the running example again. Even if all my movements, even all my thoughts, would be the same whether I were merely out jogging or trying to escape some bad guys, the point of what I am doing would be quite different in the two cases. I am not merely trying to get some exercise; I am trying to save my skin. Perhaps when I first encounter those bad guys I reason that prudence is the better part of valor and decide to run away, heading with my usual opening sprint down the road I ordinarily take, in exactly the way I have begun my evening run every day for months. A few blocks along, just as I come to a fork in the road, I pass Andy Griffith, who thinks I am out for my usual evening run. But he is mistaken, even though he is correct about all the facts about my leg movements, speed, direction, and so on. That is, the elements of his incorrect interpretation of what I am doing supervene on exactly the same set of underlying facts as do the elements of the correct interpretation, which is that I am running away from those bad guys. (They are exactly the elements that our space alien, had we enlisted her at this point, would have observed.)

Nor can we say that because my conscious decision would be different in each case there must be a difference in the underlying facts for the two interpretations. Even leaving aside the fact that I need not have made any conscious decision, the purpose of what I am doing is not always the same as the explicit decision I come to, or even my belief I have about what I am doing. Akrasia and self deception are always possible. Not only all my physical movements but even my conscious reasoning and resulting decision,

in fact even my own belief as to what I was doing, might be the same whether I was just jogging or was actually running away.[13] The difference would be that my real purpose was to get away, no matter what I or anyone else thought I was doing. The possibility of weakness and self-deception shows that both the correct and incorrect interpretation of what the agent is really doing are consistent with any explicit reasoning or choice the agent makes.

If the purpose of the action is what determines whether an interpretation is correct, we can see at least one reason why interpretations are indeterminate. I have been arguing that interpretations are indeterminate, whether they are about the actions of individual agents or about events that involve cooperation among several agents. They supervene on the underlying facts but they allow ranges of facts, rather than specifying specific underlying facts at every point, as determinate theories do. That is why there is no nomological connection between the underlying facts and the supervening, interpretive claim. Different interpretations can be perfectly consistent with the same set of underlying facts. And in the cases we have considered so far this is apparently because both interpretations allow open choices for the agent or agents in question. If this is right it tells us how the sort of indeterminacy I am claiming for these interpretations is possible, what it consists in so to speak, at least in these cases. But it does not explain why these, or any, interpretations have this feature. The answer to that question, I think, is to be found in the same thing that makes one interpretation correct and another one mistaken. The difference is that the correct interpretation includes, and incorrect ones miss, the actual purpose (or purposes) of the action or actions. Purposes necessarily involve the sort of indeterminacy we have seen in the interpretations we have looked at. It may be easiest to see this, and to see that having further open choices is not essential, if we shift for a moment from actions that have purposes to objects that do.

So consider the large rock that rests at the corner of my friend Steve's driveway, just where it meets the road. This rock has a purpose. Steve's house is on a hillside and his driveway is rather steep. It crosses a ditch (via a culvert) as soon as it leaves the road and then immediately makes a left turn downhill. The driveway is also narrow enough that if you drive in then, when you want to leave, you have to back out, since it is very difficult to turn around. The purpose of the rock is to keep people who are backing out of the driveway from accidentally going off into the ditch.

There are a few things to notice here before explaining how indeterminacy enters into this story. First, obviously this rock has this purpose only because Steve has a purpose for it. In this it differs from what biologists sometimes call 'functions', which are a result (roughly) of the evolutionary history of the ancestors of the thing that has the function.[14] Such functions can be discovered but they are not assigned. Even non-sentient things such as flower petals can have them. Rocks can't have functions of that sort, not being organisms or parts of organisms[15]. But they can have purposes, because people can have purposes for them.

Purposes involve indeterminacy in at least three ways. First, there are plenty of features of that rock that have nothing directly to do with its purpose of keeping people from driving into that ditch when backing out of Steve's driveway. To serve its purpose of course it must have some color, for instance. But within limits, it probably doesn't matter what color the rock is. Similarly for size and shape. In general when objects have purposes those purposes are served by specific features of the objects, such as, in the case

of this rock, the fact that it might do some damage to one's car to hit it. But objects always have plenty of other features than the ones which serve the purpose in question, and (perhaps within some limits) those can be anything at all and the purpose will still be served.

Second, nothing about a thing's purpose by itself specifies how it is going to be achieved. Even for that rock, its purpose might be achieved in more than one way. People might see it in their mirrors and turn slightly to miss it. Or they might hit it with a tire and change course slightly. If either happens the rock will have served its purpose. But, so far as this purpose is concerned, it doesn't matter which happens. Nothing about a thing's purpose requires that it be achieved in a specific way.[16]

But, third, the fact that something has a purpose in no way insures that this purpose will actually be achieved, or even that it can be. The fact that the purpose of that rock is to keep people from driving into the ditch is perfectly consistent with its having no effect whatsoever on the people backing out of Steve's driveway. Imagine that it is in fact a very small rock that no one even notices. Its purpose could still be to keep people from driving into the ditch.

All these sources of indeterminacy are, I think, consequences of the fact that purposes are 'intentional states' in something at least very like the way beliefs and desires are, which is why inanimate things such as rocks can have purposes only if someone has a purpose for them. But, given that, the rock still has a purpose and that purpose creates an 'intensional' context in the sentences in which it is referred to, just as any other intentional state does. For example, the purpose of that rock is to keep people from going into the ditch when backing out of Steve's driveway. It is also true as a matter of fact that keeping people from going into the ditch when backing out of Steve's driveway is a saving of the amount of gasoline needed to hitch them up to Steve's pickup to pull them out. But it doesn't follow that the purpose of that rock is to save this gasoline, though of course it might have been.

If we return now to actions and the events that constitute actions, we shouldn't be surprised that we find these same sorts of indeterminacy. Human actions are events that have purposes supplied by the agents of the actions.[17] So for actions, or at least for most of them, figuring out what the purpose of the action is or was is essential to figuring out what the action is or was. To do that is to give what I am calling an interpretative explanation.

6.

Purposes, I am claiming, explain actions but always involve indeterminacy. More than one purpose or set of purposes is always consistent with the actual underlying facts about the objects or events on which the purpose supervenes. That is the fundamental reason why the underlying facts for any interpretation need not entail or even be nomologically connected to just one interpretation. But the question remains, what makes one assignment of purposes correct and another not if it is not the underlying facts being interpreted?

I think the straightforward answer is simply that the action or actions being interpreted really do have the purposes assigned by the correct interpretation. To see what this comes to it will help to distinguish two different sorts of questions. There is a difference between asking how we know what purposes <u>are</u> (at all, so to speak) and

asking how we know when people really have certain purposes. These can get conflated if we think that figuring out when others really have a certain purpose must be in the end a matter of reducing purposes to their constitutive elements and then doing an investigation of when the actions of others possess those elements.

But this is not how it is. We should distinguish between saying that some concept 'applies to' some data and saying that a concept is 'based on' certain data. Consider some theoretical entity, such as a electron. How do we know that there is any such thing as a electron? The answer is that electrons are hypothesized by empirically very well established physical theory. And the evidence for the theory is also evidence for electrons, in fact this is all the evidence for electrons that there is. Electrons have precise, detailed roles in explanations of lots of physical phenomena, including electricity, chemical bonds, and many others. They contribute to these explanations, that is, the idea that there are electrons is empirically applicable. But at the same time electrons are only known to exist because of their place in these explanations. So the idea of an electron, besides being applicable to phenomena, is also based on exactly the same phenomena in the sense that electrons are essentially theoretical entities.

If different theories which do not use the concept of an electron are found to do a better job explaining the same phenomena, that will be taken to show that we were wrong to think that there ever were any such things as electrons. Or we might in that case end up saying that there turned out to be several different sorts of electrons, perhaps, or that, besides electrons, there were other particles that were previously thought to be electrons but were not. The point is that there is no other reason to think that there are electrons at all beside their usefulness in the theories in which they appear. If that usefulness turns out to have been illusory, so will electrons.

The fate imagined here for the notion of an electron is in essence the fate predicted by eliminative materialism for all intentionalistic and purposive concepts. The idea is that once the underlying neurophysiological mechanisms are understood, the much cruder 'folk' theoretical concepts such as purpose and intention used in commonsense explanations of action will be seen to have been illusory. This claim presupposes that these concepts, like that of an electron, are not only applicable to the phenomena with which they deal but also that they are based on these phenomena in the sense that we have no further reason to think the things they refer to actually exist beyond that provided by whatever evidence supports the theories in which they appear. But not all concepts that are applicable to the phenomena they explain are like this. In particular the idea of the purpose of an action is not like this. Our grounds for having these concepts, that is, for thinking that they apply at all, are not based on the phenomena to which they apply but arise independently of the explanations in which they are used.

Think about our space alien again. She suspends judgment as to whether the complex organisms she encounters on this planet have any intentional states, including any aims or purposes. But that doesn't mean she needs to suspend judgment about whether she herself has goals or intentional states, and she needn't suspend judgment about this even if she goes on to accept the version of solipsism that actually denies there are any other minds than her own. Such a solipsist would simply deny the applicability of intentional or purposive predicates to others than herself. This is, or at least seems to be, a coherent position. (It is for instance the position dualists seem forced into by 'other minds skepticism.') It is not incoherent even if it seems very implausible. The thought

would seem to have to be that one knows directly from one's own case what e.g. purposes and intentions are, while for others one only knows for sure about the various movements and sounds their bodies produce.

There are two things to notice about this position. First, this position is very implausible. Once one agrees that there are such things as intentional states, purposes, and the like, the evidence that others have these states is overwhelming. Someone who refused to make use of purposive explanations of others, including, importantly, purposive explanations of what they are doing in making sounds come from their mouths, would find it virtually impossible to make any sense at all of human activities. She would thus find it impossible to engage in any sort of distinctively human interactions. At the same time, regarding these activities as purposive, intentionally contentful, and the like would completely solve this problem.[18] So, while coherent, this form of solipsism would seem to be profoundly unempirical. Once one sees that it is possible for something to have purposes and other intentional states, the evidence is overwhelmingly in favor of the claim that others do indeed have them.

The second thing to notice here is that the coherence of this form of solipsism presupposes that concepts such as 'purpose' can be known from, and applied to, oneself independently of their applicability to others.[19] If it is coherent to suppose that one might oneself be the only purposive agent in existence, then the concept of a purposive agent is not dependent on having a place in theories explaining the behavior of other people. So if this is right the knowledge of what purposive agency is, and the knowledge that there are purposive agents at all, could survive the discovery that a complete explanation of the behavior of others, by neurophysiology for instance, had no place for the concept of purposive agency. But of course neurophysiology (eventually, when complete, etc.) is supposed to explain the behavior of everyone, oneself included, not just of 'others'. So the form of solipsism that holds that I know from my own case that I myself have purposes, intentional states and the like, but I don't know whether anyone else has such states, cannot be allowed by eliminativists of the sort mentioned at the beginning of this paper. An eliminativist will have to hold that the form of solipsism we are considering here is in fact not coherent since it doesn't apply neurophysiological explanations to everyone it should apply to, oneself as well as others. It tries to make an exception for the first person case. An eliminativist will have to hold that observing oneself no more reveals, or provides grounds for hypothesizing, purposes or intentions than observing others does.

This is the heart of the issue. When I 'learn from my own case' that, say, my purpose in running down the street is to get away from those bad guys, does that mean that I somehow 'observe' myself internally and then on some grounds or other attribute such a purpose to myself, more or less in the same way someone *else* who is observing or thinking about me might? An eliminativist, who holds that purposive concepts are (supposed to be) based on the evidence they (try to) explain, will have to answer yes to this question. For an eliminativist, purposes are simply crude or defective empirically based concepts. But the answer has to be no. Having a purpose, acting with some purpose, is *itself* a 'state' of an agent, frequently perfectly conscious, different from merely *attributing* such a state to someone, even oneself. And given the sort of state it is, there has to be an element of self-awareness involved. In the normal situation at least, someone who is acting with some purpose must by the very fact of acting on it, realize

what this purpose is. Being an agent, acting with some purpose, is itself a certain sort of 'mental' state. So 'having a purpose' is not a theoretical concept, like 'electron', that depends for its use (and one might say, existence) completely on some explanatory theory. Thinking that notions like purpose, intention, and the like are theoretical concepts analogous to 'electron' is similar to the mistake pointed out long ago by J. L. Austin of thinking that all language use is descriptive. Not all mental states have their content exhausted in being 'about' something else in the way beliefs and desires are about something else. In particular *doing something intentionally* is not like this.

The paradigm first person example of purposes is surely not just *describing* one's own purposes to oneself, that is, thinking *that* one has a purpose of some sort. It is actually having that purpose. Otherwise, if 'purpose' were a purely third person, explanatory concept the content of which was unavailable to whomever actually had the purpose just in virtue of having it, while we could interpret or think about our actions, it is hard to see how we could actually perform actions. It would be as if we were condemned to being merely internal observers of the motions and sounds we were making, forever trying to figure out what we ourselves were doing. This is not just comical; it is incoherent. If it were correct then it is hard to see how anyone could ever actually *do* anything at all, since just having a purpose, without also reflecting on it, wouldn't by itself be enough to let the agent know what her purpose actually was. So it is hard to see how she could actually be pursuing it. At best she would, like everyone else, have to try to figure out what she was doing by thinking about the movements and sounds she was making. But of course 'trying to figure out what I am doing' is itself a purposive activity. That is why this picture is incoherent. It has to presuppose that the internal observer is herself acting with some purpose of which she is aware, i.e. in *trying to figure out* the purposes of the motions and sounds she observes herself making. If this purposive activity, trying to figure out the purposes of the motions and sounds she is making, isn't something the content of which is available to her just by having it then presumably yet another observational level will be needed to try to figure out the purpose of this one. But obviously this does no good since exactly the same issue would arise in exactly the same way all over again.

It follows that just having a purpose includes, typically at least, awareness of that purpose, automatically so to speak, without the need for further interpretation and without anything analogous to 'observation'. The idea that I have to interpret my own actions (or movements) in the way I interpret those of others is not coherent. It leads to a regress, since interpretation is itself a purposive activity. This is a way of saying that 'purposiveness' is not a theoretical concept, used in and dependant on explanations of behavior. It is indeed applicable to behavior but it is not based on behavior. It is a concept we bring to explanations of behavior from the fact of our own agency.

These last two points together give us at least the basic elements of an answer to the question of what makes an interpretation, an assignment of purposes, correct. The answer is that this is an empirical issue that turns, like any such issue, on the explanatory power of the interpretation proposed, but the essential concept involved, that of purposiveness, is not itself based on evidence in the way other theoretical concepts are. We understand purposiveness, acquire the very idea so to speak, from the fact that we are ourselves purposive agents. Once we have that idea, however, we can apply it like any theoretical concept to the events and behavior that confront us.

[1]The original version included lots more detail and was of course much funnier. It is available at

[2] She is probably a relative of the Martian described by Daniel Dennett in 'True Believers', in *The Intentional Stance* (Cambridge, MIT Press, 1987), pp. 13-42.

[3] We will return to this issue below.

[4] Though this is true of my version of the story, all these elements are in fact included in Griffith's actual routine.

[5] Wittgenstein, *Philosophical Investigations*, paragraph 200.

[6] This assumption will be defended below.

[7] This is the actual claim, but I will sometimes abbreviate this by saying that the two *interpretations* supervene on the same underlying facts.

[8] McLaughlin, Brian and Karen Bennett, "Supervenience", The Stanford Encyclopedia of Philosophy (Fall 2006 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2006/entries/supervenience/>.

[9] Quoted in Scott Sehon, *Teleological Realism* (Cambridge, MIT, 2005), p. 117.

[10]See for instance Simon Blackburn, 'Moral Realism' in *Morality and Moral Reasoning* ed. J. Casey, Methuen, London, 1973, reprinted in his *Essays In Quasi-Realism*, New York, Oxford University Press, 1993.

[11] On the general question of whether there can be supervenience without either logical or nomological 'reduction' see Sehon (2005), Chap. 8. Sehon gives several examples, based on making the supervening property a noncomputable number, where there is supervenience but no possibility of entailment of the supervening facts by the underlying facts no matter what extra scientific law is supposed to connect the two.

[12] A 'basic action' is something one does, such as raising ones arm, but not 'by means of' doing something else. To take the earlier example, one might signal one's boredom by means of rolling one's eyes back. So signaling boredom would not be a basic action but rolling ones eyes back presumably is.

[13] According to Wittgenstein, 'It is, of course, imaginable that two people belonging to a tribe unacquainted with games should sit at a chess-board and go through the moves of a game of chess; and even with all the appropriate mental accompaniments.' (*Philosophical Investigations*, #200)

[14] I am simplifying things here since there is another account of function which doesn't depend on evolution, roughly the 'causal role' account. Though using that account would complicate the argument here, so far as I can tell it makes no essential difference. See my *Reasons and Purposes* (Oxford, Oxford University Press, 2003), Chap. 1 for a fuller discussion of these two accounts.

[15] Even a rock could have a function in the 'causal role' sense though. A rock might function to keep moisture in the soil under it from evaporating for instance.

[16] Of course whoever gave the object its purpose might have believed or even intended that the purpose would be achieved in a specific way. But that is not strictly required for a thing's having a purpose.

[17] I don't intend this to be a definition. It seems to me to be true, but if there are actions of which it is not true, then obviously what I say here won't apply to them. Of course

'supplying' a purpose needn't be, in fact cannot be, an intentional action, since that would just lead to a regress.

[18] Dennett imagines a 'predicting contest' between a human using the ordinary, purposive interpretations and an outer space alien such as the one imagined above who knows all the underlying physically described facts but uses no purposive or intentional concepts. Without the intentional and purposive concepts, the outer space visitor of course losses badly. See Dennett (1987).

[19] This of course was famously questioned by Wittgenstein and his followers.