Behavior and Philosophy, *32*, 359-375 (2004) © 2004 Cambridge Center for Behavioral Studies

CAN WE INTERPRET IRRATIONAL BEHAVIOR?

Lisa Bortolotti
University of Manchester

ABSTRACT: According to some theories of interpretation, it is difficult to explain and predict irrational behavior in intentional terms because irrational behavior does not support the ascription of intentional states with determinate content. In this paper I challenge this claim by offering a general diagnosis of those cases in which behavior, rational or not, resists interpretation. I argue that indeterminacy of ascription and paralysis of interpretation ensue when the interpreter lacks relevant information about the system to be interpreted and about the environment in which the system is embedded. Moreover, the heuristics of interpretation that guide the ascription of beliefs can be limited in scope. In the end I suggest that by giving up the idea of a necessary rationality constraint on the ascription of intentional states we can develop a new framework for a more psychologically realistic account of interpretation.
Key words: interpretation, rationality, rationalization, heuristics, intelligibility

Introduction

Consider the following case. Someone says to you: "There are flies in my head." You are not sure of what this person is trying to tell you. You might think he is attempting to use an idiomatic expression you are not familiar with, or to metaphorically allude to a state of confusion or dizziness. Another possibility is that this person believes that (literally) there are flies in his head. In a case such as this, in which you lack information about the man who is uttering the sentence, you might not be sure about the belief content to be ascribed to the speaker and leave it indeterminate.

According to an influential view in philosophy of mind, there is a necessary connection between a system's rationality and our capacity to interpret the behavior of that system in terms of intentional states, that is, in terms of states that have a representational content. For instance, we might think that the shop manager has raised prices to increase profits. We describe the manager's behavior in terms his beliefs and desires. He *wants* to increase the profits and he *believes* that by raising the prices profits will increase. His belief and his desire are intentional states, that is, states with content that represent how the world is (belief) and how the agent wants the world to be (desire).

Philosophers like Davidson and Dennett have argued that we cannot explain or predict irrational behavior in intentional terms. This claim is part of a theory of

AUTHOR'S NOTE: Please address all correspondence to Lisa Bortolotti, CSEP/IMLAB, School of Law, Williamson Building, Oxford Road, University of Manchester, Manchester M13 9PL, United Kingdom. Email: Lisa.Bortolotti@manchester.ac.uk

interpretation. Interpretation starts from the observation of a system's behavior and ends with the ascription of intentional states to that system that make it possible to explain what the system does and predict what it will do in the future. I take it that to explain an instance of behavior is to refer to the most salient events that have brought about that behavior. When the behavior is a case of intentional action or the formation of a belief, we might look for an explanation of it that refers to the causes of the action or the belief that are *reasons* why the agent performed that action or formed that belief. According to most philosophers, a reason is the combination of those intentional states (beliefs, desires, emotions, etc.) that motivates a system to behave as it does.

Even though there is no explicit mention of intentional states in the explanation of an action, they might still play an important role, or so folk psychology says. What explains the fact that Stewart put on a hat before stepping outside? Today it is windy. In its elliptical form, this is an intentional explanation. I explain Stewart's action by referring to what Stewart believes (he believes that it is windy outside) and to what he desires (he wants to keep warm, or he wants to avoid catching a cold). If he didn't believe that outside it is windy or he didn't care about the possible effects of the temperature or the wind outside, he probably wouldn't have put the hat on.

Notice that you don't need to subscribe to any metaphysical view about the mind, or the nature of beliefs and desires, in order to talk about our folk-psychological practices. Whether you think that beliefs are inner states that causally affect behavior, or just useful fictions, you can agree that in everyday life we do tend to explain the actions and reactions of the people around us by reference to what they think and what they want. The question I am interested in is whether we can do that successfully in case of irrational behavior.

We seem to characterize irrational behavior in intentional terms, by ascribing beliefs and desires to people who, say, make reasoning mistakes or are weak-willed. Thanks to such ascriptions, we can explain why those people do what they do and we can predict what they will do next. Philosophers who believe that intentionality and rationality are necessarily linked can concede that there are some cases of intentional and yet irrational behavior by claiming that the interpretation and prediction of irrational behavior are not *impossible*, but *difficult*.

In this paper I am going to challenge this weaker thesis, which is grounded on the idea that when we engage in folk psychology we lack the resources for dealing successfully with irrational behavior. First, I shall explain what I mean by "rationality" and what it is for interpretation to be difficult. Interpretation becomes difficult if we cannot ascribe to a system intentional states with determinate content. In those cases, the interpreter cannot offer a unique satisfactory explanation or prediction of the behavior of the observed system. If the indeterminacy is pervasive, a paralysis of interpretation might ensue.

Then, I shall consider some scenarios of everyday interpretation. By reflecting on the features of these scenarios one comes to realize that the difficulties in interpreting or predicting a system's behavior are not necessarily due to the irrationality manifested by that system. Often, the limitations in the interpreter's

knowledge of the surrounding environment or of the system to be interpreted are responsible for indeterminate ascriptions.

In the attempt to offer a diagnosis of the problems that the interpreter might encounter, I shall provide a general framework for a better understanding of interpretation. I shall focus not on the relation between the rationality and the intentionality of a system's behavior, but on the background knowledge of the interpreter and the context-dependent features of the interpretive scenario. We do not have at this stage a detailed psychological explanation of how interpretation works, nor evidence that might suggest which principles we follow when we interpret the behavior of the creatures around us. So, these issues are almost entirely left to philosophical speculation. Still, we can aim at providing a framework that will encourage further empirical research and that is more compatible with our experience as interpreters than traditional accounts of interpretation.

Rationality and the Difficulties of Interpretation

Remember the man who claims that there are flies in his head. Our difficulty in understanding the man's utterance might be taken to be evidence for the claim that the ascription of radically implausible beliefs (from the point of view of the interpreter) is problematic. We need to ask *why* it is so difficult to arrive at a determinate ascription in cases like this.

If you believe that cases of irrational behavior are always problematic because our attempts to describe, explain, and predict behavior in folk-psychological terms necessarily rest on an assumption of rationality, then you are in good company. This view of folk psychology as tied to a general assumption of rationality is widely supported in the philosophical literature:

> If we are intelligibly to attribute attitudes and beliefs, or usefully to describe motions as behavior, then we are committed to finding, in the pattern of behavior, belief, and desire, a large degree of rationality and consistency. (Davidson, 1980, p. 237)

> It is clear that the scope for indeterminacy diminishes when we require [. . .] that interpretation should meet various standards of plausibility, rationality and consistency. (Child, 1994, p. 70)

> Propositional attitudes have their proper home in explanations of a special sort: explanations in which things are made intelligible by being revealed to be, or to approximate to being, as they rationally ought to be. (McDowell, 1985, p. 389)

> When we are not [rational], the cases defy description in ordinary terms of belief and desire. (Dennett, 1987, p. 87)

As I anticipated, there are two ways in which one might impose a rationality constraint on the interpretation of behavior. The *weak* necessary claim is that where there is no rationality there is no ascription of states with determinate

content. This claim seems to be supported by cases in which behavior that departs significantly from norms of rationality eludes the interpreter's attempt to offer an explanation of it in intentional terms. The *strong* necessary claim holds that behavior that departs from norms of rationality to a significant extent cannot be intentionally characterized at all. An example of a strategy that stems from the endorsement of the strong necessary claim is the original version of Dennett's intentional system theory. According to Dennett, one cannot adopt the intentional stance to understand and predict the behavior of a system that departs from norms of rationality. Depending on the aims of the interpreter and on the complexity of the system's behavior, the physical or the design stance might be adopted instead.

But one might not be convinced that the difficulty in interpreting the behavior of the man who says that there are flies in his head is due to his irrationality. It might very well be due to the interpreter's lack of familiarity with the speaker and with the specifics of the situation. Maybe the speaker's psychiatrist or one of his close friends would be better placed than a stranger to understand his *prima facie* puzzling utterance. Let's suppose the speaker suffers from a delusion due to brain damage (some patients affected by schizophrenia or in the first stages of dementia report that they feel insects moving inside their heads, and some drug users suffer from the delusion of insects crawling under their skin and consequently engage in obsessive scratching). Given his medical condition, the sentence "There are flies in my head" should be interpreted literally and is not as perplexing as we might think at first, because it might be uttered as a consequence of the speaker having abnormal perceptual experiences.

Compare the ascription of a belief to a brain-damaged person suffering from delusions with the ascription of beliefs to a young child or a dog, when these individuals are not familiar to us. The difficulty in ascribing to them beliefs with determinate content is not necessarily due to their departing from norms of rationality. Important factors are our lack of experience in recognizing their behavioral responses and the differences between our psychological make-up and theirs. This is a trivial point. Isn't it easier for me to describe, explain, and predict in intentional terms the behavior of a person who shares my interests, opinions, and desires than that of a person who doesn't? But clearly some of these differences between me as interpreter and the person to be interpreted have nothing to do with the satisfaction of normative standards of rationality.

A clarification needs to be made at this point. What do we mean by "rationality" in this context? We need to distinguish the adherence to a schema for practical reasoning from the conformity to other standards of rational thought and action. The appeal of the idea that intentional explanation rests on an assumption of rationality probably derives from an equivocation between these two notions. The adherence to a schema for practical reasoning is what rationalizes behavior. There is a sense in which folk-psychological explanations are *rationalizations*, that is, they are straight-forward applications of the schema for practical reasoning. Think about the case of Stewart, who puts on his hat before facing a cold wind.

The schema applies to his action:

1. *Belief*: Stewart believes that it is windy outside.
2. *Desire*: He wants to be protected from the wind.
So,
3. *Action*: He puts on his hat before going outside.

Stewart's action is rational because it makes sense given Stewart's set of beliefs and desires. In other words, (3) is a reasonable conclusion given (1) and (2). Stewart's intentional states (1) and (2) rationalize action (3). But to put on a hat is not necessarily the best action available to Stewart in the circumstances and might fail to satisfy other criteria. For instance, the hat might fly away if the wind is too strong or it might not be warm enough to protect Stewart from the cold wind. The question now is not whether (3) follows from (1) and (2), but whether (3) is *the rational* thing to do. This second question appeals to rationality as optimization, which is at home in theories of economics (e.g., Popper's principle of rationality) and in the psychological literature on human reasoning (see Kahneman & Slovic, 1982).

Does the explanation of Stewart's behavior in terms of his beliefs and desires (and possibly other intentional states) presuppose that Stewart is conforming or responding to some standards of rationality? Do we need to assume that the beliefs and desires Stewart entertains are those he *ought to* have, or that his action conforms to what he *ought to* do? No. All we need is that he has a reason for his action, that is, there is a belief-and-desire pair that motivates him to act in that way. Whether Stewart's beliefs, desires, and actions are rational is beside the point if all we want to know is whether or not his behavior can be explained in folk-psychological terms.

Pettit and Smith (1990) make an interesting distinction that can be mapped onto my distinction between (*a*) rationalization and (*b*) rationality as optimization. They distinguish the *intentional conception*, according to which agents do things for reasons (and reasons are given by their beliefs and desires) from the *deliberative conception*, according to which agents do what they consider desirable given certain values or standards. The distinction between intentional and deliberative in this context is not identical to the distinction between rationalization and satisfaction of normative standards because the deliberative conception also requires that agents be aware of the standards they meet when they are being rational and that they be responsive to them. But both dichotomies—intentional versus deliberative conception and rationalization versus optimization—seem to be grounded on the basic idea that considerations about the extent to which intentional states, options, or actions satisfy certain standards are not relevant to whether or not agents' behavior qualifies as intentional.

While the intentional conception says that every action issues from a set of beliefs and desires that rationalize it, the deliberative conception holds that somewhere in the process leading to action there is normally the belief that the

option chosen has a property which provides some justification for choosing it (Pettit & Smith, 1990, p. 566).

In the attempt to identify the mark of intentional action, some philosophers argue for the view that there cannot be intentionality without rationality, but their evidence only supports the claim that intentional behavior can be rationalized. Child (1994) characterizes what rationality is for Davidson and seems to endorse the view that the term "rationality" refers to the more demanding notion of satisfaction of some normative standards. Rationality is about what reasons for action are *good* reasons.

> Rationality includes everything relevant to saying what constitutes a good argument, a valid inference, a rational plan, or a good reason for acting, and everything relevant to the application of those notions to the particular case; it includes practical rationality, not just theoretical rationality; it involves inductive as well as deductive rationality; and considerations about rationality are relevant too, in assessing how a belief and desire must cause an action that they explain. (Child, 1994, pp. 57-58)

The question is, then, why should we endorse the view that rationality as optimization is a constraint on the ascription of intentional states? Clearly, there can be arguments that are not good arguments, plans that are not rational, and inferences that are not valid. To claim that these arguments, plans, and inferences are not instances of intentional behavior seems overwhelmingly implausible. It might be much more appealing to claim that there is a limit to how irrational intentional behavior can be. Is a very bad argument still an argument? Is a very irrational plan still a plan?

One might argue that the distinction between rationality as optimization and adherence to a schema for practical reasoning is misleading. One might think that where rationality in its demanding form breaks down, it is also difficult or impossible to rationalize behavior. Think about a case in which the violation of a normative standard for beliefs makes it impossible for the potential believer to adhere to a schema for practical reasoning. Suppose that Mark has many inconsistent beliefs, that he believes that *p* and not-*p*, that *q* and not-*q*, and so on. He has beliefs and desires, but the fact that inconsistency is widespread in his belief system could make it impossible for me to explain and predict his behavior. If I attempt to explain the fact that Mark cut the tree by saying that he believed that the tree was dead, you can challenge my explanation by reminding me that he also believes that the tree is not dead. If I predict that Mark will go to see the latest Spielberg movie because he believes that Spielberg is the best living director, you will doubt my prediction on the basis that Mark assented to the claim that Spielberg is not the best living director. How can I offer an intentional characterization—or a rationalization—of the behavior of a largely inconsistent believer?

I don't think that the project of interpreting people like Mark is doomed. At any one time Mark will endorse either *p* or not-*p* and the interpreter will make an attribution on the basis of Mark's behavior at that stage. Why did Mark cut down

the tree? Because he believed it was dead. Why was he so upset afterwards? Because at that time he also believed it was alive. Did he change his mind between starting to cut the tree and becoming upset? No, but one belief was more operative at the beginning and the other became operative later.

Only if Mark is explicitly inconsistent and endorses the conjunction of two contradictory statements will the interpreter have to wait for other clues in his behavior before attempting an ascription or be resigned to ascribing a pair of inconsistent beliefs. If Mark's adherence to a schema for practical reasoning is totally compromised, then Mark will not count as an intentional system, but the reason for that will be his failure to satisfy the conditions for intentionality and not his failure to satisfy the standard of consistency.

The two sets of conditions are not completely unrelated, and this explains why people have often conflated them in the literature on belief ascription. In most cases, however, it is helpful to keep the two notions separate to offer a more convincing account of the capacities involved in having beliefs, capacities that cannot be sufficient for the satisfaction of most standards of rationality.

The Interpreter's Toolbox

In the previous section I suggested that when difficulties in interpretation emerge and the interpreter does not seem able to ascribe to the system beliefs with determinate content, rationality might not be the central explanatory notion. The problem does not necessarily lie in the system's failure to behave rationally. All that is required is that the states that the interpreter is tempted to ascribe to the system as beliefs are, to some extent, explanatory of some of the system's other beliefs and actions by having the right causal relations with the rest of the system's manifest behavior. I am going to argue in this section that most of the difficulties we encounter when we fail to ascribe states with determinate content is due to the fact that the behavior of the system is outside the scope of the application of the heuristics that *guide* (as opposed to *constrain*) the interpretive exercise. Considerations about rationality might still have an important role to play at the level of what I call the "heuristics of interpretation," together with considerations about the plausibility of the system's beliefs and their similarity to the interpreter's beliefs. However, considerations of rationality, plausibility, or similarity don't have to determine *whether* the behavior can be described in intentional terms. There need be nothing incoherent, or even perplexing, in the case of intentional systems that behave irrationally.

Whether we ascribe intentional attitudes with determinate content presumably depends on several factors. It is a precondition for ascribing intentional states that the interpreter has the capacity to master folk-psychological concepts. Think about the results of the experiments on the false-belief task. Children younger than three years of age fail to ascribe false beliefs to the character of an incomplete fictional story and, as a consequence, they fail to predict what the character will do in the rest of the story (Wimmer & Perner, 1983). It seems that intentional explanation and prediction require some understanding of what beliefs are and how they

work—they require a basic grasp of the concept of belief. By reflecting on the data gathered in the false-belief task one might conclude that humans are not born competent folk psychologists but they acquire some mind-reading capacities in the course of their normal development.

There are contingent aspects of the context of interpretation that might affect the process of ascription. As I already suggested, our attributions are more likely to be determinate in content and to support specific predictions if the system whose behavior we are observing is well known to us, belongs to our socio-cultural environment, or shares interests with us (considerations of similarity or familiarity). For instance, I have been told that in Greece one is supposed to shake one's head to mean "yes" and nod to mean "no." Such a difference in non-verbal communication, if not acknowledged, could make everyday conversation between English- and Greek-speaking people very confusing and significantly undermine mutual understanding.

Another factor in the account of how to provide attributions of intentional attitudes is the assumption of rationality. Here we need to be able to distinguish two claims. One is that our ascriptions are *more likely* to be determinate if the system whose behavior we are attempting to explain is rational (*weak contingent claim*). In this case we are not introducing a constitutive aspect of minded beings but a useful heuristic for the interaction with *some* minded beings, those beings whose behavior is likely to conform to certain standards of rationality, e.g., well-educated adult humans with no reasoning impairment.

Some authors would take a more radical position and say that if the system's behavior is not rational enough (where the threshold needs to be specified) our ascriptions are going to be indeterminate (*weak necessary claim*) or impossible (*strong necessary claim*). I concede that considerations about how likely it is for a system's behavior to satisfy standards of rationality can be a valuable guide to the interpreter, together with considerations of similarity or familiarity, relevant information about the environment, and the appeal to folk-psychological generalizations. Experience teaches us that rationality is not always the key to ascribing intentional states with determinate content. Sometimes we interpret effortlessly the behavior of systems that fall short of some standards of rationality and, conversely, find it difficult to interpret systems whose behavior meets such standards. Rationality might not always be the key to good interpretation.

Let's consider how interpretation works in the following three scenarios:

Scenario 1

I enter the kitchen and see that Karen is kicking the fridge repeatedly. I remember I noticed yesterday that the door doesn't shut properly. I realize that Karen is trying to shut the fridge and ascribe to her the belief that if she kicks the fridge hard enough it will stay shut.

Karen's behavior might not be the best available solution to her present problem, but it is perfectly intelligible to me given what I know about her (that she usually does not kick objects randomly), what I know about the fridge (that it

needs fixing), and what I can observe. If she did kick any object around her when frustrated, then her behavior would not necessarily be intentional and could not be explained in terms of her desire to keep the fridge shut. Considerations that might be relevant to my ascription of intentional states to Karen are the following: I would probably engage in the same behavior if I wanted to shut the fridge; people might try drastic methods to shut doors when they cannot do so otherwise; etc. The former is an appeal to what I would do if I were in Karen's shoes, while the latter is a generalization about people's behavior. Other considerations of the same kind could easily apply.

Note that such considerations are not obviously *rationality* assumptions. In this scenario there is no need for any substantive notion of rationality to appear among my assumptions as an interpreter. The question of whether or not it is rational for Karen to kick the fridge is not relevant to my attempt to describe her behavior in intentional terms. As an interpreter, the interesting question for me is whether or not I can explain Karen's behavior by ascribing her intentional states, such as the desire to shut the fridge. Whether kicking the fridge is the best way to satisfy that desire in those circumstances seems to be a completely different question, and a more complex one. An assessment of the rationality of Karen's behavior depends on the details of the situation, the alternatives Karen might have, the time frame in which she needs to find a solution, the long-term effectiveness of her proposed solution, and so on.

Scenario 2

Yesterday I visited my friend Dan at the hospital. He had just undergone some surgery and he was still under the influence of a general anesthetic. I asked him how he was feeling and we talked for ten minutes. During our conversation he seemed fine and gave me a fairly detailed account of what the doctor had told him earlier. Today I go back to the hospital to check if he is still fine. He is, but he has no recollection whatsoever of the conversation we had yesterday. Actually, he seems not to realize I had been there at all. He tells me that he thought he saw me in a dream. I ascribe him the belief that I didn't go to see him yesterday.

Given that I have no reason to think that Dan is being insincere, I interpret what he says literally. I know he usually has a very good memory, but I also know he was under an anesthetic when he saw me and this helps me make sense of the fact that he cannot remember our conversation. I might even appeal to the consideration that in the same conditions I would behave pretty strangely too, or that people might not remember what happens to them when they are under the influence of an anesthetic. In this case, even more transparently than in the case we considered before, these assumptions derive from my previous experience of interactions with other agents or from things I have learned. They certainly do not depend on whether the behavior in question is itself rational or whether it is rational for agents to engage in it.

My opponent could disagree on this last point. Maybe the interpreter does not rely on the system being rational, but he might rely on the system *not* being

irrational. The fact that Dan was under anesthetic is, after all, an *excuse* for the fact that he forgot something he would not have forgotten otherwise and that he consequently has a false belief. In ascribing to him the belief that I didn't go to see him yesterday, I implicitly rely on the fact that it was not irrational for him to forget my visit given his conditions. In Davidsonian terms, we are confronted with an *explicable error*.

I can see the appeal of this reply, but I am still not persuaded that *rationality* has a role to play. In the case just described, all that I am relying on is that it is *normal* for Dan to be forgetting yesterday's events. This notion comes from my background knowledge about the effects of a general anesthetic, not from the assumption that Dan's behavior is rational. After all, it would be misleading to claim that if we remember what we have experienced in our recent past then we are rational. Our ability to remember past events is a sign that some of our cognitive capacities, those related to memory, are functioning well, rather than a sign of rationality. In the example, Dan is *excused* because his memory was not functioning well under the influence of an anesthetic, but this does not make his behavior more or less rational.

Let me turn to the last scenario. This is supposed to make us realize that, in some cases, interpretation could not happen unless the interpreter worked under the assumption that the system is not behaving rationally.

Scenario 3

While walking in the city centre, I am stopped by a man. He asks me whether I have noticed the dog resting on the steps of a nearby Catholic church. I shake my head and ask him the reason for this question. The man tells me that he saw the dog looking at him seriously and raising his front paw in a gesture of salute, but he wasn't sure whether the salute was addressed to him or to me. When I say that I don't think the dog was saluting me, the man looks very excited and mutters something about having received a revelation (this example is adapted from the case reported by Leeser & O'Donohue, 1999).

I realize that the man's behavior is very unusual, even though I have no previous information about his mental health. The fact that he seems to derive a completely unjustified conclusion from the observation of a dog might not prevent me from ascribing him the belief that something significant happened to him, even though it is mysterious to me why he thinks the events he reports are significant. However, if I knew that the man suffers from schizophrenia and that in the past he has had other "significant experiences" involving dogs, I could ascribe him a more definite set of beliefs and attitudes. My knowledge that certain rational explanations of his behavior become less likely given his condition would make it *easier* for me to offer a tentative interpretation of his behavior. I could use what I know about the behavior of schizophrenic subjects to make sense of what he said and did.

In scenarios 1 and 2 there is no difficulty of interpretation. In scenario 3 the interpretation of the behavior of the agent might result in an indeterminate

ascription (or in the thought that the agent is not expressing a belief, but, say, making fun of the interpreter). What is the diagnosis of this apparent difficulty? The difficulty seems to disappear, as I suggested, if more information about the cognitive situation and history of the system to be interpreted is available to the interpreter. How does this suggestion relate to the view that intentional characterization requires an assumption of rationality?

Discussion of the Scenarios

What some philosophers call rationality is just the basic capacity to act in accordance with one's beliefs in order to satisfy one's desires. A system satisfies this basic notion when it adheres to a schema for practical reasoning, when it forms beliefs and desires, and when it acts for reasons. This kind of behavior supports the notion of rationalization as I have described above and is legitimately regarded as intentional. An intentional system does and says things for reasons, has beliefs and desires, and acts on its beliefs in order to satisfy its desires. Karen kicks the fridge because she wants to shut its door. The man I met in the street is excited because he thinks the fact that the dog saluted him was part of a revelation.

It is possible to rationalize their behavior, but this does not mean that their behavior meets any normative standards of rationality, save that of having been done for a reason rather than being the result of a twitch or spasm. Kicking the fridge might not be the best way to deal with Karen's problem, and the man I met is hardly justified (at least from our point of view) in thinking that the dog did something meaningful. That is, instances of behavior that might fail to meet some standard of rationality can nonetheless be regarded as intentional.

My suggestion is that we should view the practice of interpretation as guided by fallible (but generally successful) heuristics. By reflecting on the practice of interpretation we can sketch a very simple framework. When we interpret others we use our background knowledge constituted by some folk-psychological generalizations and by what we know about the system to be interpreted and the surrounding environment. In some contexts we also adopt the heuristic of *similarity* and the heuristic of *rationality*. That is, we tend to assume that people who are in some respects similar to us behave as we do, and we also assume that some human adults are consistent, reason competently, and so on. The heuristics won't apply to all of the instances of behavior that we might be interested in interpreting, and even when they do apply, they will be applied in different ways and to a different extent depending on the situation and our knowledge of the individual system. For instance, we might realize that in traumatic situations reasoning competently is not at all easy, or that even people who are similar to us in their opinions and social behavior might act very strangely when heavily intoxicated. In cases like these the heuristics of rationality and similarity will be adapted—or not adopted at all. We know that the traumatized person is not likely to reason at his best and that the intoxicated person might react very differently from us.

I refer to the interpreter's assumptions as flexible heuristics rather than constraints because they are supposed to guide the interpreter and help him to ascribe intentional states with determinate content, but they are not criteria for deciding whether one system has beliefs or not. There is nothing particularly original in the idea that considerations of similarity and rationality are of help. In the debate about how we read other minds, where the theory-theorists are opposed to the simulation-theorists, these considerations are very familiar.

In my account of the way in which we ascribe beliefs, these considerations, similarity and rationality, do not need to be seen as incompatible and do not necessarily exhaust the considerations that the interpreter might find useful in the process of making an attribution. To what extent these heuristics contribute to the practice of interpretation can vary. In cases of radical interpretation the interpreter might rely heavily on them because the knowledge of the system and its surrounding environment might be very limited, sometimes consisting only in what the interpreter can observe directly. In cases of more familiar attributions (attributions to members of the same linguistic community, friends, etc.) the knowledge of the system and the environment is more extensive and the heuristics will be less indispensable in the circumstances. In scenarios 1 and 2 the actors were familiar to the interpreter and, in normal circumstances, their behavior could be easily explained and predicted. Recall that some adjustments were necessary in scenario 2, where the actor, Dan, fails to remember an event that happened just the day before because he was affected by a general anesthetic. In scenario 3 the speaker is a member of the same linguistic community as the interpreter but his behavior is still perplexing and difficult to make sense of. My hypothesis was that his behavior, i.e., the fact that he ascribes special significance to the apparent gesture made by a dog in front of a church, could be understood better by people who knew that he is a person who has schizophrenia.

A Davidsonian might be persuaded by my distinction between the notions of rationalization and rationality as optimization yet resist my rejection of a necessary constraint on interpretation. Even if one agrees that our everyday practice of interpretation is guided by fallible heuristics rather than a general constraint, one might still want to defend the view that there is a kind of constitutive constraint on interpretation. Maybe the constitutive constraint should not be that a system's behavior must satisfy certain standards of rationality—instead one might believe that the correct interpretation of a system's behavior is the one that makes best possible sense of that system's total life and conduct. Let me say first that I am more sympathetic to an overall intelligibility constraint than to a rationality constraint. However, we might not need either of those.

The overall intelligibility constraint would not be of much use in the actual practice of interpretation because interpreters have to deal with temporally bounded instances of a system's behavior, they are in conditions of partial ignorance, and they lack a general overview of the system's interactions with the given environment. One might recognize that actual interpreters need to rely on their background knowledge and heuristics yet claim that these considerations put no pressure on the suggestion that there is a constitutive constraint that is not

necessarily reflected in the practice of interpretation. The overall intelligibility constraint would still be operative, not in the account of the empirical project of interpretation but in the account of ideal interpretation, and the ideal interpreter would be the useful fiction that allows us to gain a better understanding of what we ought to do when we ascribe beliefs.

I see the force of this objection. However, I also think that if we did take such a position seriously we would lose interest in the theory of interpretation. Child (1994) identifies the core thesis of *interpretationism* with the idea that we can gain an understanding of propositional attitudes by reflecting on how we interpret a subject's behavior. Interpretationism is appealing because it attempts to clarify difficult theoretical issues such as the nature of intentional states and the conditions for being a believer by analyzing our actual practice of interpretation. If we endorse the constitutive version of the overall intelligibility constraint, the powerful link between the theoretical issues and our task as interpreters is broken. The conditions of belief attributions are no longer in step with the evidential basis on which we actually attribute beliefs in real time in the real world. The conditions of belief attributions will depend on the constraints applicable to an idealization of our actual practice of interpretation that does not necessarily reflect what we do when we ascribe beliefs. This seems to undermine the basic motivation for the interpretationist view.

Of course, it is open to the Davidsonian to have a double account. That is, one can insist that there is a constitutive constraint that operates at the level of ideal interpretation and at the same time allow that other considerations (e.g., the interpreter's background knowledge and the heuristics he might use) play a major role in the real world. However, before resigning to the divorce between actual interpretation and the inquiry into the nature of the mental, it would be interesting to see whether or not the considerations that guide the interpreter in his daily practice have something interesting to tell us. They might offer us an insight into the nature of belief states and the conditions for being a believer. This alternative route needs developing but is worth pursuing as it would prove faithful to the original motivation of interpretationism, which did not seem to allow for there being a gap between the principles that govern the practice of interpretation and the necessary constraints on belief ascription.

Towards a New Account of Interpretation

The interpreter starts from the observation of what subjects say and do and how they interact with their environment. On the basis of these observations and some background knowledge, he ascribes beliefs. The most controversial point in the theory of interpretation is the specification of the background knowledge that the interpreter needs to make sense of what the believer says and does. It does seem right to presuppose that the interpreter needs to be guided by some assumptions in the task of ascribing beliefs to the systems around him. The problem is how to characterize these assumptions and how to identify their source. As we know, the received view is that the interpreter's assumptions about the

behavior of the systems around her reflect a general constitutive link between intentional behavior and rationality. This link has been advocated by Quine (1960), Davidson (1982), Dennett (1979), and Cherniak (1986).

However, the inadequacy of the rationality constraint motivates us to attempt a revision of this picture and hint at the direction that this project of revision might take. When the interpreter identifies a system as intentional he does so on the basis of the system's behavior. If the system asserts that *p*, claims to have some evidence for *p*, defends the claim that *p* when challenged, and sometimes acts on *p*, then the system manifests the pattern of behavior of a believer with respect to *p* and the interpreter will ascribe to the system the belief that *p*. But notice that there is no assumption that *it is rational* for the system to believe that *p*. The system might not have *good* evidence for *p* or might neglect some evidence against *p*, but these considerations alone should not constitute a reason for the interpreter to doubt that the system believes that *p*.

The core idea of a theory of belief ascription is that an interpreter can make sense of another system's behavior by:

- observing the behavior of the system
- observing the environment in which the system is embedded
- establishing causal connections among himself, the system, and the environment
- making hypotheses about what the system believes on the basis of his expectations about what systems of that kind, or that individual in particular, would believe in that environment

The formation of the hypotheses in the last instance is grounded on the interpreter's background knowledge, which includes some folk-psychological generalizations, some information about the system, the kind of system, and the individual, and some information about the environment. A folk-psychological generalization is something like: "people whose bodily tissue is damaged experience pain," "creatures avoid what they are afraid of," and so on. The information about the system that might turn out to be relevant includes data about the system's psychological make-up and stage of development, its belonging to some linguistic or cultural group or community, and its background beliefs and opinions. Of course, this information won't always be available to the interpreter. The same applies to the information about the environment, which might include information about other agents in that environment that are not being the direct focus of interpretation, or contextual factors that have a causal impact on the system and its behavior.

Think about the classic scenario of radical interpretation. I see a man pointing at a hopping rabbit and uttering a sentence in a language unknown to me. Given what I can observe about the man and the environment (that there is a rabbit, that the man seems to want to refer to the rabbit with his utterance, etc.) and given my assumption that the man is sensitive to events in his environment, my working hypothesis will be that the man is uttering something like: "Look, there is a

rabbit," "What a beautiful rabbit!," or "Dinner!" Only a better acquaintance with the language that the man is speaking, or a better acquaintance with the man's intentions and beliefs about rabbits, could make my ascription more determinate. Even though I do not know much at all about the man, just by observing him I can assume that he is a normal adult human being with a complex network of beliefs and desires. If I knew that the man found rabbits particularly cute, I would guess that he might have meant to say: "What a beautiful rabbit!" If it had come to my attention, instead, that the man finds rabbits delicious to eat, I would ascribe him the belief that the rabbit would make a fine dinner and predict that he will try to catch the rabbit.

In interpreting the man's behavior I am not necessarily relying on any rationality assumption. I might think that the man who finds rabbits cute is likely to exclaim "What a beautiful rabbit!" because his appreciation of rabbits gives him a reason to do so. But if there is something normative here is just the relation between his appreciation of rabbits and his exclamation, the fact that the former is a reason for the latter.

I can also focus on the rabbit, and notice that it accelerates when the man points at it and he starts running after it. Though there is no utterance to which I can ascribe meaning in this case, I might wonder why the rabbit hops faster and hides in a bush. The reason might be that the rabbit identifies the excited man as a potential predator and wants to avoid being captured (or, more simply, runs away because it is scared). In this case I shall still appeal for my ascription to the knowledge of the environment and the assumption that there are causal relations among the actors within the same environment. I shall also appeal to what I know about rabbits. Rabbits are not likely to engage in sophisticated, abstract, or self-reflective thought, but they might be able to form simple beliefs about their surrounding environment, and potential prey and predators are likely to attract their attention. In this scenario we have two cases of radical interpretation, yet there is no need for the interpreter to assume a conceptual link between the intentionality of belief states and rationality.

Rather, the interpreter's expectations and assumptions could be seen as a set of inductive generalizations triggered by contextual features of the environment shared by the interpreter and the believer and by what the interpreter already knows about the believer. These generalizations might be of very heterogeneous nature and might partially concern the reasoning capacities of the believer, the similarity of the believer's opinions to those of the interpreter, or even the believer's capacity to follow certain norms for the formation, integration, and revision of beliefs. But in some contexts, as we have seen, it is the assumption that the believer will behave irrationally that leads to a more determinate and effortless attribution. We sometimes expect the speaker to be irrational: When we ask undergraduate students to solve the Wason selection task[1], we expect them to

[1] In this task subjects are asked to test a conditional rule, and most of them fail to solve the task because they do not appreciate the necessity to falsify the rule. For more details on the experiment see Watson & Johnson-Laird (1965).

commit a reasoning mistake. When we discuss politics with an obstinate friend we might expect her to defend a view that we find implausible. Often, it is the assumption that believers will not meet some standards of rationality that allows us to ascribe beliefs with determinate content or predict the intentional behavior of others.

Conclusion

I have discussed and challenged the following view about the relation between the intentionality exhibited by believers and rationality: Even though it is possible to ascribe beliefs to systems that behave irrationally, the belief states we ascribe to them might not turn out to be determinate in content, and this shows that there is some weak but necessary relation between intentionality and rationality.

I have suggested that there can be another way to explain the difficulties that the interpreter sometimes finds in ascribing beliefs with determinate content, both to systems that do not share the interpreter's language and to systems that seem to behave irrationally. In both circumstances the interpreter might have limited knowledge of the situation, for instance, of the psychological make-up of the system or of causal relations in the environment. Rationality does not necessarily have a major role to play, although considerations about the rationality of a system's behavior might be useful in some specific contexts. My account, if properly developed and corroborated by the psychological data, promises to overcome some of the criticism faced by theories of interpretation for postulating constraints that do not seem to operate in the actual practice of interpretation. Another advantage of my position is that systems that do not exhibit rational behavior (e.g., as ordinary people when they make reasoning mistakes or people affected by mental illness) can still be seen as systems with beliefs, desires, and reasons for action.

An issue I have not addressed is whether or not the suggested framework does what an interpretationist would like it to do—elucidate the nature of intentional states such as beliefs via the analysis of the practice of belief ascription. How can my account deliver insights into the nature of the mental if it does not point at any constitutive fact about believers? Maybe it cannot. But a more definite answer won't be available until more work will have been done on the analysis of our actual practices of belief ascription and on the description of the psychological mechanisms underlying them.

References

Cherniak, C. (1986). *Minimal rationality*. Cambridge, MA: MIT Press.

Child, W. (1994). *Causality, interpretation and the mind*. Oxford: Clarendon Press.

Davidson, D. (1980). Psychology as philosophy. In D. Davidson (Ed.), *Essays on actions and events* (pp. 229-238). Oxford, UK: Oxford University Press.

Davidson, D. (1982). Paradoxes of irrationality. In R. Wollheim (Ed.), *Philosophical essays on Freud* (pp. 289-305). London: Cambridge University Press.

Davidson, D. (1984). *Inquiries into truth and interpretation*. Oxford: Clarendon Press.

Dennett, D. C. (1979). Intentional systems. In D. C. Dennett (Ed.), *Brainstorms: philosophical essays on mind and psychology* (pp. 3-22). Montgomery, VT: Bradford Books.

Dennett, D. C. (1987). Making sense of ourselves. In D. C. Dennett (Ed.), *The intentional stance* (pp. 83-101). Cambridge, MA: MIT Press.

Kahneman, D., Slovic, P., et al. (Eds.). (1982). *Judgements under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.

Leeser, J., & O'Donohue, W. (1999). What is a delusion? Epistemological dimensions. *Journal of Abnormal Psychology*, *108*(4), 687-694.

McDowell, J. (1985). Functionalism and anomalous monism. In B. McLaughlin & E. Lepore (Eds.), *Actions and events* (pp. 387-398). Oxford: Blackwell.

Pettit, P., & Smith, M. (1990). Backgrounding desire. *The Philosophical Review*, *99(4)*, 565-592.

Quine, W. V. O. (1960). *Word and object*. Cambridge, MA: MIT Press.

Watson, P. C. & Johnson-Laird, P.N. (1965). *Psychology of reasoning: Structure and content*. London: Batsford.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*, 103-128.