Philosophical Psychology
Lecture 09: Interaction in Radical Interpretation

Stephen A. Butterfill
<s.butterfill@warwick.ac.uk>

December 7, 2016

title-slide

Radical Interpretation: The Question

'I want to know what it is about propositional thought–our beliefs, desires, intentions and speech–that makes it intelligible to others.' (Davidson, 1995 p. 14)

slide-3 'I want to know what it is about propositional thought—our beliefs, desires, intentions and speech—that makes it intelligible to others.'

(Davidson 1995, p. 14)

slide-4 The difference is just that whereas Davidson focusses on propositional thought, one theme of my talk will be that we need to consider a wider range of mental phenomena, including emotional processes, attention and goal-directed action.

slide-8 Let me start with the aim of an account of radical interpretation. An account of radical interpretation is an account of how you could in principle infer facts about actions and mental states from non-mental evidence. This is the fundamental project in theorising about social cognition (as I'll explain shortly [by invoking Marr's levels]).

slide-9 The evidence is evidence that you could possess in advance of knowing what any particular individual believes, desires, intends and so on.

slide-10 Here we must be careful. Donald Davidson and David Lewis have both used the term 'radical interpretation'. But Lewis is interested in how what I am calling 'The Evidence' determines facts about the mind. This is a *metaphysical* question. By contrast, Davidson is interested in the possibility of inferring The Mind from The Evidence. He is not assuming—and, I think, does not belive, that the Evidence metaphysically determines The Mind (although he's sometimes interpreted as so believing). Davidson's

project (and the part of Dennett's we're interested in) is epistemological, not metaphysical.

[NB: this is particularly important because we're talking about Dennett]

slide-5

The Intentional Stance

The 'Intentional Stance' (also 'Intentional Strategy') is Dennett's name for a procedure for ascribing desires, beliefs and other mental states to an individual without assuming any prior knowledge of what distinguishes her mind from others' minds.

slide-12 Dennett doesn't use the term 'radical interpretation' but his description of the Intentional Stance does provide a candidate account of radical interpretation.

'the intentional stance …

'first you decide to treat the object whose behavior is to be predicted as a rational agent; 'then you figure out what beliefs that agent ought to have , given its place in the world and its purpose. 'Then you figure out what desires it ought to have, on the same considerations, 'and finally you predict that this rational agent will act to further its goals in the light of its beliefs'

(Dennett 1987, p. 17) A couple of things about this are confusing.

slide-16 (Dennett 1987, p. 17)

A couple of things about this are confusing.

slide-17 First, what is the purpose of me?

slide-18 Second, where do these goals come from? Does Dennett mean desires here?

slide-19 This does make sense to me. Dennett derives this rule from the fundamental injunction to attribute all the beliefs the agent ought to have:

'one rule for attributing beliefs in the intentional strategy is this: attribute as beliefs all the truths relevant to the system's interests (or desires) that the system's experience to date has made available' (Dennett 1987, p. 18)

slide-20 But what about the desires? Which desires ought we to have? Here Dennett is less helpful.

I prefer gold rings to platinum ones; you prefer the converse. Does one of us have the wrong preferences?

‘We attribute the desires the system ought to have. That is the fundamental rule. It dictates, on a first pass, that we attribute the familiar list of highest, or most basic, desires to people: survival, absence of pain, food, comfort, procreation, entertainment.’ (Dennett 1987, p. 20)

He also suggests attributing desires for things we believe are good for us, or that will further other desires. From the context, I think Dennett’s idea is that, in essence, we all desire a small number of basic things but differ in our beliefs about how to get these.

slide-21 Put aside for the moment the question of whether this succeeds as an account of radical interpretation. I’ll come back to this. First I want to be sure we understand the point of giving an account of radical interpretation.

slide-12

Social Cognition vs Radical Interpretation

How would an account of the radical interpretation (which is about how discoveries about minds are in principle possible) further our understanding of actual social cognition?

slide-22 What is the relation between an account of radical interpretation and a theory of social cognition?

(This is a way of asking, How is radical interpretation relevant given that our topic is social cognition?)

An account of radical interpretation describes a route to knowledge starting from evidence that can be described without knowing anything about the particular actions, beliefs, desires and other mental states of any individual and ending with knowledge of these particulars.

A theory of social cognition is (in part) a theory of the processes by which we actually predict, understand, align with and shape each other.

What do these two things have to do with each other?

slide-23 Sometimes when you read Davidson and Dennett, it seems like an account of radical interpretation just is a theory of social cognition, a theory of the processes by which we predict, understand, align with and shape each other …

‘Do people actually use this strategy? Yes, all the time.’ (Dennett 1987, p. 21)

slide-24 ‘[a]ll understanding of the speech [and thoughts] of another involves radical interpretation’ (Davidson 1973, p. 125)

slide-25 But elsewhere Davidson says things which give the opposite impression.

'The approach … I have outlined is not, I am sure it is clear, meant to throw any direct light on how in real life we come to understand each other' (Davidson 1980, p. 12)

How should we understand the relation between an account of radical interpretation and a theory of social cognition? (This is a way of asking, How is radical interpretation relevant given that our topic is social cognition?)

slide-26 One possibility is to appeal to David Marr's famous three-fold distinction bweteen levels of description of a system: the computational theory, the representations and algorithm, and the hardware implementation.

This is easy to understand in simple cases. To illustrate, consider a GPS locator. It receives information from four satellites and tells you where on Earth the device is.

There are three ways in which we can characterise this device.

slide-27 First, we can explain how in theory it is possible to infer the device's location from it receives from satellites. This involves a bit of maths: given time signals from four different satellites, you can work out what time it is and how far you are away from each of the satellites. Then, if you know where the satellites are and what shape the Earth is, you can work out where on Earth you are.

slide-28 The computational description tells us what the GPS locator does and what it is for. It also establishes the theoretical possibility of a GPS locator.

But merely having the computational description does not enable you to build a GPS locator, nor to understand how a particular GPS locator works. For that you also need to identify representations and alogrithms …

slide-29 At the level of representations and algorthms we specify how the GPS receiver represents the information it receives from the satellites (for example, it might in principle be a number, a vector or a time). We also specify the algorithm the device uses to compute the time and its location. The algorithm will be different from the computational theory: it is a procedure for discovering time and location. The algorithm may involve all kinds of shortcuts and approximations. And, unlike the computational theory, constraints on time, memory and other limited resources will be evident.

slide-30 So an account of the representations and algorithms tells us …

– How are the inputs and outputs represented, and how is the transformation accomplished?

slide-31 The final thing we need to understand the GPS locator is a description of the hardware in which the algorithm is implemented. It's only here that we discover whether the device is narrowly mechanical device, using cogs, say, or an electronic device, or some new kind of biological entity.

slide-32 The hardware implementation tells us how the representations and algorithms are represented physically.

slide-33 How is this relevant to my question? My question was, What is the relation between an account of radical interpretation and a theory of social cognition?

slide-34 I suggest that an account of radical interpretation is supposed to provide a computational description of social cognition; it tells us what social cognition is for and how, in the most abstract sense, it is possible.

slide-35 This is why Davidson says that a theory of radical interpretation isn't 'meant to throw any direct light on how in real life we come to understand each other'. What he means is that it isn't about the representations and algorithsm, nor about the hardware implementation–the neurophysiology in our case–of social cognition.

slide-36 But he's also right that all '[a]ll understanding of the speech [and thoughts] of another involves radical interpretation'. This follows trivially from the fact that a theory of radical interpretation is supposed to be a computational description of social cognition.

slide-37 Finally, the view I'm offering (a theory of radical interpretation is supposed to provide a computational description of social cognition) makes sense of a puzzle about Dennett's claim that people use the Intentional Strategy all the time. The puzzle is to understand how he could know this without doing some research. The answer, I think, is that Dennett's Intentional Stance, like any theory of radical interpretation, isn't a theory about how individuals understand each other; it is a theory about what it is to understand each other—that is, a computational description of social cognition.

slide-38 I've been arguing that

A theory of radical interpretation is supposed to provide a computational description of social cognition.

If this is right, then an account of radical interpretation makes a fundamental contribution to the study of social cognition.

slide-39 Simulation theory and theory theory are theories about REPRESENTATIONS AND ALGORITHMS, not about the COMPUTATIONAL DESCRIPTION. For this reason, you could accept Davidson's theory about radi-

cal interpretation while taking any view on simulation vs theory. The issues ([1]radical interpretation and [2] simluation vs theory) are only indirecty related.

slide-42 Debates over simulation vs theory theory are debates about alogrithms and representations. Whichever view is selected here, a computational description (theory of radical interpretation) is still needed. The computational description characterises minds and actions from the point of view of the simulation process (that is, as they are simulated) or from the point of view of the theory (that is, as the theory characterises them as being).

slide-44 Social cognition needs a theory of radical interpretation ...

... but we don't have one.

So far I've argued that we need a theory of radical interpretation in order to understand social cognition because a theory of radical interpretation provides a computational description of social cognition.

Now I want to argue that there is a problem: we don't have an adequate theory of radical interpretation. There are actually two main theories of radical interpretaion, Davidson's and Dennett's, so strictly I should offer objections to both. But here I'll just consider Dennett's theory.

objection_to_intentional_stance

An Objection to the Intentional Stance

Does the Intentional Stance actually describe how it would be possible, even in principle, to infer facts about minds and actions from evidence that can be described without knowing anything about the particular actions, beliefs, desires and other mental states of any individual?

slide-47 Suppose someone is faced with a choice between these two things. What will she choose? Let's apply the Intentional Stance to make a prediction.

The Intentional Stance tells us to attribute a desire for food. And to attribute beliefs the agent ought to have, which in this case are the beliefs that this stuff is poison and that this stuff is food.

We then assume rationality in order to predict an action: she will choose the food, not the poison.

Now what happens when we observe that our prediction is wrong and that she chooses the poison which looks like food?

It's just here that her behaviour is informative, on the Intentional Stance. The failure of our prediction tells us that we are wrong about something.

slide-48 But what are we wrong about? The belief, the desire or the rationality?

Of course we cannot hope to work out what the error is on the basis of just this observation. But we can sometimes distinguish false belief (this poison is food) from inappropriate desire (to eat something poisonous) from a faulty link between beliefs-desire pairs to actions.

slide-49 What we want from a theory of interpretation is an explanation of how the available evidence ever enables us to distinguish between different kinds of errors. And as far as I can tell Dennett has not attempted to supply such a theory.

This illustrates a more general problem for Dennett.

slide-50 On the intentional strategy,

What enables an interpreter to distinguish your actions, beliefs, desires, feelings and other mental states from anybody else's?

On the intentional strategy as Dennett describes it, it basically comes down to two things:

- your 'place in the world' (location?)

- your biological needs

slide-51 So I think it is clearly a mistake to say that the Intentional Strategy is an account of radical interpretation. Minimally, an account of radical interpretation needs to explain how the interpreter can identify the distinctive actions, emotions, beliefs, desires and other mental states of an individual. The Intentional Strategy appears fundamentally unsuited to this purpose because the Intentional Strategy involves making very little use of evidence.

By itself, this is not a deep objection to Dennett. You might say, his theory is fine as far as it goes; it was only meant to cover a couple of basic principles, leaving much of the project for future development.

But what exercises me is that we don't have an account of radical interpretation. We have yet to explain how an interpreter can identify the distinctive actions, emotions, beliefs, desires and other mental states of another individual.

unit_041

Davidson's Theory of Radical Interpretation

What is Davidson's account of radical interpretation?

slide-54 I'm still not completely happy that I've properly explained the evidence which is the starting point. Perhaps this is a better way to put it: the evidence is evidence that you could possess in advance of knowing what any particular individual believes, desires, intends and so on.

I've just been saying that Dennett's account of the Intentional Strategy isn't much use as an accounnt of radical interpretation because it doesn't explain what sort of evidence would be useful in inferring facts about minds. Fortunately Davidson's theory is more illuminating …

slide-55 Evidence:

At time t, Ayesha comes to hold 'Sta piovendo' true because p

slide-56 Ayesha is just the person we are trying to interpret; she is the target of radical interpretation. We want to what her sentences mean and what she believes.

slide-57 Holding true is an attitude Ayesha has to a sentence. To hold a sentence true is to have a belief. So the evidence we are starting with is really evidence about beliefs. But, importantly, can know that Ayesha holds a sentence true without knowing what the sentence means and so without knowing what Ayesha believes.

So evidence of this kind is evidence that is in principle available to a radical interpreter at the start.

slide-58 The 'p' picks out a proposition; it may be that Ayesha holds this sentence true because it is raining, because her eyes are open, and because there are splashes in the puddle on the roof outside her window.

slide-59 Note that we are interested not in which sentences Ayesha holds true but in what causes here to change her beliefs—what causes her to come to hold a sentence true.

slide-60 At any particular time there will be many propositions p such that Ayesha comes to hold S true because p. As already mentioned, Ayesha comes to hold the sentence 'It is raining' true because it is raining, because she has her eyes open and because there are splashes in the puddle on the roof outside her window.

There is, then, no hope of inferring what Ayesha believes from a single change in Ayesha's holding a sentence true.

slide-61 But we can consider many different events of Ayesha coming to hold a sentence true at different times. Consider, for instance, that on one occasion Ayesha comes to hold the sentence 'It is raining' true because she has her eyes open. On another occasion, she comes to hold this sentence true while cycling through an intense storm; on this occasion, the fact that she has her eyes open plays no role in her coming to hold the sentence 'It is raining' true.

slide-62 The evidence confirms or falsifies a generalisation of the form, Ayesha comes to hold S true because p.

The hope is this: if we have enough evidence, we will find that the only generalisation supported by all the evidence is this one: Ayesha comes to hold the sentence 'It is raining' true because it is raining.

slide-63 This is unrealistic, of course. However Davidson's theory doesn't strictly require this because sentences are things with structure. They contain elements, the words, which reoccur in different sentences. Davidson exploits this in making his theory of radical interpretation much more sophisticated than the simplified version I am describing, and avoiding the implausible notion that we would observe Ayesha coming to hold each sentence true millions of times. But we can ignore the complication as it won't be central to our interests in social cognition (it would be more relevant for philosophy of language).

slide-64 So, idealising and simplifying, we have lots of evidence which, for each of many sentences like 'It's raining', supports a unique generalisation about why Ayesha comes to hold that sentence true. For example, the only generalisation supported by all the evidence for the sentence 'It's raining' is the one that says Ayesha holds this sentence true because it's raining.

slide-65 We already saw that this assumption is required by Dennett's Intentional Strategy. The assumption allows us to draw a conclusion about meaning:

slide-67 For our purposes you could replace this by 'S means that p', if you believed in meanings. (There are good reasons for appealing to truth conditions but they aren't central on this course.)

slide-68 … so when Ayesha comes to hold S true, she comes to believe that p

slide-69 Why is Davidson's a better theory?

Because it specifies the evidence from which radical interpretation can start, and because it allows us to connect multiple events and so at least have a chance of distinguishing different errors someone might make.

slide-70 How can it be elaborated?

slide-71 — exploit sentence structure

slide-72 — include desire

slide-73 What are its limits?

slide-74 — no use for wordless targets

slide-75 — bold assumption about evidence

Can we really know which events are events of Ayesha coming to hold a particular sentence true in advance of knowing anything about her mind and actions? On the face of it, holding true will involve an intentional action!

unit_061

Objections to Davidson's Theory of Radical Interpretation

There are some compelling objections to Davidson's account of radical interpretation which motivate searching for ways to improve it.

slide-77 Dennett: his account is fine for targets of interpretation who lack words, but it offers no way of exploiting evidence about linguistic behaviours in his account of radical interpretation.

Davidson: his view has the converse weakness. (Why is this a weakness? Our inability to use words in communicating with an alien species would not necessarily prevent us from coming to know much about their minds and actions.)

An adequate theory of radical interpretation ought to avoid both weaknesses: it should characterise inferences for targets of interpretation without words, and it should characterise the additional complexities for radical interpretation entailed by the use of words.

slide-78 What is the point of being there with someone while it's happening to her? Being there with someone often enables you to know and to regulate—and even to share, sometimes—what she's feeling. …

Why is this an objection to the claim that Davidson's accounnt of Radical Interpretation is a fully adequate computational description of social cogntion in humans? Let me explain …

slide-79 1. On Radical Interpretation (and the Intentional Stance), the outputs of social cognition are (i) propositional attitude ascriptions and (ii) action predictions.

2. Emotions unfold …

3. … and this is not comprehensible as a series of changes in propositional attitudes.

So: 4. Understanding the way emotions unfold is not a matter of ascribing propositional attitudes or predicting actions.

But: 5. Humans do sometimes understand the way anothers' emotions are unfolding.

So: 6. Radical Interpretation (and the Intentional Stance) is not a fully adequate computational description of human social cognition.

slide-85 On Davidson's account of radical interpretation, we can think of its upshot as an assignment of propositions to sentences. The propositions give the truth conditions, or meanings, of the sentences and so enable us to identify the target's beliefs and other propositional attitudes.

Because the evidence Davidson considers is attitudes towards whole sentences, it turns out that, for any assignment of propositions to sentences, there are ways of generating an alternative assigment of propositions to sentences which is exactly as well supported by the evidence as the original assignment of propositions to sentences is.

So on Davidson's account of radical interpretation, there is no possibility of uniquely determining the truth conditions, or meanings, of sentences.

This is analogous to having a computational theory of a GPS device on which there is just no possibility of the device distinguishing between its being here and its being at the same point on the opposite side of the earth.

Let me illustrate how the indeterminacy arises …

slide-86 [Use Shoemaker's shadows.]

slide-87 Incidentally, a similar objection involving indeterminacy arises for Dennett.

What does the objection tell us? If I gave you a computational theory of the GPS device that suffered from indeterminacy, you would rightly reject that theory because the device can, as a matter of fact, determine which side of the planet it is on. But should we take the same attitude towards Davidson's theory. He says not …

slide-88 'It makes no sense, on this approach, to complain that a theory comes up with the right truth conditions time after time, but has the logical form (or deep structure) wrong. We should take the same view of reference.' (Davidson 1984b, p. 223)

But pointing …

slide-89 The evidence Davidson starts from is changes in the attitude of holding a sentence true. To be detectable, such changes must involve the target of interpretation uttering a sentence. How are such events represented at the outset of radical interpretation?

slide-90 If they are represented merely as sequences of joint displacements, bodily configurations and sounds, then we need an account of how it is determined which events are changes in the attitude of holding a particular sentence true.

slide-91 If, on the other horn, they are represented as intentional actions, then we are presupposing some insight into the contents of the target of radical interpretation's intentions. We know that she has an intention to express a particular attitude towards a particular sentence.

slide-92 'a radical interpreter is not, at the beginning of his study, informed about any of the basic propositional attitudes of his subject.' (Davidson 1984a, p. 17)

'The important limitation is that [the radical interpreter] doesn't know in detail the contents of any of the propositional attitudes of the person to be interpreted: she doesn't know what he intends, believes, wants or means by what he says.' (Davidson 1994, p.)

slide-93 Davidson might take this horn of the dilemma and just insist that his view of what an account of radical interpretation aims to achieve is a bit less exciting than I have been suggesting. It doesn't start from no insight into what someone intends, just relatively little.

I think this response would be unsatisfactory given our interests in social cognition. The hardest part is surely to understand the step from *no* insight into others' minds and actions to *some* such insight. If Davidson's account of radical interpretation is really just about the step from *some* insight into others' minds and actions to a bit more insight, it might be interesting but it isn't the theory we were looking for. It won't after all provide us with a computational theory of social cognition.

slide-94 Can Davidson instead take the first horn of the dilemma and say that the changes in attitude are represented merely as sequences of joint displacements, bodily configurations and sounds?

He might insist that as long as there is some account of how we get from the joint displacement to the changes in attitude towards the sentence, there is no problem.

So on this horn, Davidson's radical interpretation project is the computational theory of social cognition which we are looking for. But it is in-

complete because it doesn't include an account of the transition from joint displacements, bodily configurations and their effects (such as sounds) to changes in attitude towards the truth of sentences.

On the face of it, this doesn't sound like there's going to be an objection to Davidson here. But things get very interesting when we consider what is currently known about how this transition is made. Essentially, we will see that the capacities involved in getting from joint displacements, bodily configurations and their effects (such as sounds) to goal-directed actions provides a significant form of social cognition in its own right.

[This line of objection leads beautifully into speech perception and the motor theory!]

slide-95 I think all of these objections arise from a single source. Davidson's account of radical interpretation starts and ends with linguistic expressions of changes in attitudes towards whole sentences. It doesn't consider simple object-directed actions like reaching for a mug or catching a ball, and it doesn't consider nonlinguistic communicative activities like pointing; nor does it consider expressions of emotion like some smiles and grimaces.

slide-96 Social cognition needs a theory of radical interpretation …

… but we don't have one.

slide-98 So what next? We have a choice.

slide-99 I take the view that social cognition involves multiple systems which operate on quite different principles. So because I think of an account of radical interpretation as providing a computational description of a system, I have to reject the idea that there is just one account of radical interpretation. Instead we need multiple accounts of radical interpretation for the multiple systems involved in social cognition.

In thinking about how to break it down, there are two possibilities.

slide-100 One possibility would be to consider whether we can perceptually experience mental states. This appears to involve taking a position substantially at odds with Davidson's account of radical interpretation insofar as the computational theory will look quite different for perception … the evidence that we start from will not be linguistic expressions of changes in attitude towards sentences, and the sorts of inference we make will probably not have the causal character of the inferences Davidson makes use of in his account of radical interpretation.

slide-101 A different possibility would be to try to continue developing Davidson's account of radical interpretation by thinking about this ques-

tion: How do humans make the transition from bodily configurations, joint displacements and their effects (e.g. sounds) to goal-directed actions?

[I think that these are not strictly speaking different alterantives. Categorical perception of expressions of emotion is a form of goal ascription, just as speech perception is. So the Teleological Stance provides the computational description common to the most fundamental forms of social cognition.]

teleological_stance_v2

The Teleological Stance [recap]

The Teleological Stance (Gergeley and Csibra , 1995) provides a computational theory of pure goal ascription. Pure goal ascription is the process of identifying goals to which anothers' actions are directed independently of any knowledge, or beliefs about, the intentions or other mental states of an agent.

slide-103 In the lecture on behaviour reading last week, I offered a brief survey of mechanisms that could be involved in getting from joint displacements and bodily configurations to larger, more abstract bits of behaviour grouped into units in ways that reflect structures of action.

slide-104 An account of pure goal ascription is an account of how you could in principle infer facts about the goals to which actions are directed from facts about joint displacements, bodily configurations and their effects (e.g. sounds). Such an account is a computational theory of pure goal ascription.

slide-107 'an action can be explained by a goal state if, and only if, it is seen as the most justifiable action towards that goal state that is available within the constraints of reality' (Csibra and Gergely 1998, p. 255)

slide-113 We start with the assumption that we know the event is an action.

slide-116 Why normally? Because of the 'seen as'.

slide-121 Any objections?

I have an objection. Consider a case in which I perform an action directed to the outcome of pouring some hot tea into a mug. Could this pattern of inference imply that the outcome be the goal of my action? Only if it also implies that moving my elbow is a goal of my action as well. And pouring some liquid. And moving air in a certain way. And ...

How can we avoid this objection?

slide-123 Doesn't this conflict with the aim of explaining *pure* behaviour reading? Not if desirable is understood as something objective. [explain]

slide-124 Now we are almost done, I think.

slide-126 OK, I think this is reasonably true to the quote. So we've understood the claim. But is it true?

slide-127 How good is the agent at optimising the selection of means to her goals? And how good is the observer at identifying the optimality of means in relation to outcomes? **For optimally correct goal ascription, we want there to be a match between (i) how well the agent can optimise her choice of means and (i) how well the observer can detect such optimality**. Failing such a match, the inference will not result in correct goal ascription.

But I don't think this is an objection to the Teleological Stance as a computational theory of pure goal ascription. It is rather a detail which concerns the next level, the level of representations and algorithms. The computational theory imposes demands at the next level.

slide-128 So this is the teleological stance, a computational description of goal ascription.

Although this is rarely noted, I think the Teleological Stance takes us beyond Dennett's intentional stance because it allows us to distinguish between people on the basis of what they do. You reach for the red box; your goal is to retrieve the food. I reach for the blue box, so my goal is to retrieve the poison.

But there is a problem for the Teleological Stance ...

slide-129 The Teleological Stance is a starting point for a theory of radical interpretation ...

... but it's limited.

teleological_stance_limits

Limits of The Teleological Stance

The Teleological Stance is subject to at leasttwo limits. One is the Problem of Opaque Means: ignorance about to which ends actions are means can impair goal ascription. Another is the Problem of False Belief: where agents act on false beliefs, the teleological stance generates systematically incorrect goal ascriptions.

slide-132 Return to my two boxes. Consider what the Teleological Stance says about someone who reaches for the blue box, which contains posion: The goal of her action is to retrieve the poision.

This ascription will be incorrect when the agent has false beliefs about the contents of the box mental states of another individual.

Now you might say, 'Of course this is a limit. How could it not be? After all, the whole point of the Teleological Stance is that it doesn't involve ascribing mental states. And how could you correctly identify the goals of actions based on false beliefs if you aren't ascribing mental states?'

To repeat: how could you correctly identify the goals of actions based on false beliefs if you aren't ascribing mental states?

I'm so glad you asked …

slide-133 The problem of opaque means: failures to identify to which ends actions are means can impair goal ascription.

slide-134 While wriggling the pram, it looks a lot like she's trying to throw the baby out, or as if she's attacking the bus. More prosaically, it's also hard to tell whether her goal is to extract the pram from the bus or to get it on.

slide-135 The case of Byrne's Rwandan mountain gorilla's preparing stinging nettles is another good case; it might well be hard for an unskilled observer to recognise to which end these actions are means.

slide-136 The use by another of an unfamiliar tool to achieve something. For example, maybe she has a novel-to-you tool for hulling rice which involves throwing it into the air; not recognizing this tool's function you are puzzled by her action and unable to identify it's goal.

slide-137 OK, so opaque means impair goal ascription. But why do they do so? Reflection on the teleological stance already gives us the answer …

slide-138 Why do opaque means impair goal ascription? (Csibra and Gergely 1998, p. 255)

To make this inference, you have to know which outcomes an action is a means of realising. Where the problem of opaque means arises, this is exactly what you don't know.

So the problem of opaque means prevents you from using the teleological stance to identify the goals of an action.

slide-139 The Problem of Opaque Means arises for referential communication.

slide-140 The problem of opaque means: failures to identify to which ends actions are means can impair goal ascription.

slide-141 The problem of opaque means also affects communicative actions because these characteristically have goals which the actions are means to realising only because others recognise them as means to realising those goals (a Gricean circle).

To illustrate, you have to imagine that you didn't understand pointing. We can take a step towards this by imagining landing on a planet where people point to things with their shoulders rather than their fingers, and where the shoulders are turned to a location 35 degrees westwards of the object. It might take a while to figure out that some shoulder movements are pointing gestures.

slide-142 The Teleological Stance is a starting point for a theory of radical interpretation …

… but it's limited.

your_goal_is_my_goal

Your goal is my goal

If an interpreter is able to interact with her targets, if she is not limited to merely observing them, how might this enable her to exploit a route to knowledge of the goals of their actions? The answer hinges on interactions involving collective goals.

slide-148 An outcome is a *collective goal* of two or more actions involving multiple agents just if the actions are directed to this goal and this is not, or not just, a matter of each action being individually directed to that goal.

slide-151 In what follows I'm going to rely on two assumptions: (1) joint actions involve collective goals (2) you can identify something as a joint action without ascribing any mental states.

slide-152 Here is an intuitive idea that doesn't quite work: if an interpreter is engaged in an interaction with her target that involves a collective goal, it may be easy for the interpreter to know what the goal of her target's actions is because this goal is the goal of her own actions. So if she knows the goal of her own actions and she knows that she is engaged with her target in an interaction involving a collective goal, then she already knows what the goal of her target's actions are.

Roughly speaking, the mindreader can reason about her target thus: your goal is my goal.

slide-153 Of course this intuitive idea is no use it stands. For the inference it captures relies on the premise that the interpreter and her target are engaged in actions with a collective goal. But for the mindreader to know this premise it seems she must already know which goal her target's actions are directed to.

slide-154 Fortunately there is a way around this. For there are various cues which signal that one agent is prepared to engage in some joint action or other with another, and joint actions involve collective goals. Seeing you struggling to get your twin pram onto a bus and noticing you have the haggard look of a new parent, a passing stranger grabs the front wheels and makes eye contact with you, raising her eyebrows and smiling. (The noise of the street rules out talking.) In this way she signals that she is about to act jointly with you. Since you are fully committed to getting your pram onto the bus, you know what the sole goal of your own actions will be. But you also know that the stranger will engage in joint action with you, which means that, taken together, her actions and your actions will have a collective goal. This may enable you to infer the goal of the stranger's imminent actions: her goal is your goal, to get the pram onto the bus.

slide-155

1. You are about to attempt to engage in some joint action[1] or other with me.
2. I am not about to change the single goal to which my actions will be directed.

Therefore:

3. A goal of your actions will be my goal, the goal I now envisage that my actions will be directed to.

slide-156 For example, because you have made eye contact with me while I was in the middle of attempting to do something)

slide-158 I claim (i) you could know the premises without already knowing the conclusion, and (ii) knowing the premises could put you in a position to know the conclusion. So the inference is a route to knowledge.

It describes how interacting interpreters might come to know facts about the goals of others' actions.

The teleological stance is one route to knowlegde of other's goals, and this is another.

slide-160 So the Teleological Stance and Your-Goal-Is-My-Goal provide computational descriptions of complementary processes of goal ascription. Each is limited, but in different ways than the other.

[1] We leave open the issue of how joint action is to be characterised subject only to the requirement that all joint actions must involve collective goals. Attempts to characterise joint action in ways relevant to explaining development include Tollefsen (2005), Carpenter (2009), Pacherie (2011) and Butterfill (2012).

slide-163 The two routes to knowledge are complementary: one demands knowledge of means-ends relations and so is no good when the means are opaque to the interpreter; the other places different demands on the interpreter.

slide-164 I claim that your-goal-is-my-goal enables you to avoid the problem of opaque means. But how does it work?

Earlier I mentioned three examples. Let's see how your-goal-is-my-goal enables you to avoid the problem of opaque means in each of these three cases.

slide-166 Actually I already did this one in introducing the inference.

slide-167 We saw earlier that the problem of opaque means may impair goal ascription where actions involve novel uses for tools. How could your-goal-is-my-goal mitigate the problem in such cases? Imagine we are interacting with a young child, Ayesha, and want her to understand how a new tool is used. It is difficult to convey this to her directly. So we first get her interested in achieving an outcome that would require the new tool, knowing that she will perform actions directed to achieving this outcome. We then signal to Ayesha that we will act jointly with her. Now she is in a position to know what the goal of our action will be when we deploy the tool. She is able to identify this goal despite being unable to recognize it as an end to which our tool-using action is a means. She is able to identify this goal because she knows that this is her goal and that we were attempting to engage in joint action with her. This is one illustration of how interacting interpreters have at their disposal ways of identifying the goals of actions involving novel uses of tools which are unavailable to interpreters who can only observe.

slide-168 As this example indicates, exploiting your-goal-is-my-goal can shift the burden of identifying goals from a mindreader to her target. In the example Ayesha is the focal mindreader and we are her target; but her success in identifying the goal of our actions depends on this, that our willingness to act jointly with her is based on *our* knowledge of the goals of *her* actions. In purely observational mindreading, the target's beliefs about the goals of the mindreader's actions are not normally relevant (except, of course, when the mindreader is ascribing such beliefs). But interacting mindreaders who rely on your-goal-is-my-goal thereby rely on their targets' having correctly identified the goals of their actions. Of course this is sometimes a reason not to rely on your-goal-is-my-goal. But where the target understands relevant means-ends relations, such as actions involving novel tools, the your-goal-is-my-goal route to knowledge of others' goals may sometimes be the only option.

slide-169 This is a bit of a problem …

slide-170 Interaction makes available routes to knowledge of the goals of actions.

Simple forms of interaction enable correct goal ascription in situations where it would not otherwise be possible. So if we are giving a computational description of social cognition, we must consider the interpreter as not merely observing her target but also potentially interacting with her.

I also want to make an observation which leads to a conjecture. Simple forms of interaction simultaneously give rise to a need to identify false beliefs and a means of distinguishing actions that are guided by such beliefs. This suggests a conjecture: the ability to track false beliefs may be a consequence of the ability to engage in simple forms of social interaction.

slide-175 This leaves open many outstanding questions.

Are there further ways in which interaction matters in radical interpretation? (Or is it merely a facilitator?)

How does the theory sketched here contribute to problems of radical interpretation? — E.g. Indeterminacy resolved?

ISSUE: Is there a deep theoretical obstacle to specifying canonical theory of mind? Davidson's claims about normativity (norms of the observer ...)?

ASIDE: Extend minimal theory of mind to include preferences, emotions

How can we test theories of radical interpretation? — Signature limits (already mentioned for minimal theory of mind)

nonreferential_communication

Interaction and Nonlinguistic Referential Communication

We've seen that capacities for interaction make available a route to knowlegde of others' goals which avoids the problem of opaque means, characterised by the 'your-goal-is-my-goal' inference. This suggests a solution to a challenge about nonlinguistic referential communciation.

*TODO link with inefficient actions as communicative. How do we fix the goal of an inefficient action? Consider the simple 'inefficient route' study as an example.

slide-177 Why is nonlinguistic referential communication promising? It promises to provide a link between basic forms of social cognition involving, say, goal ascription, and more sophisticated forms of mindreading involving the ascription of propositional attitudes. To see why, consider an inference ...

slide-178 1. At time t, Ayesha comes to hold ‘Questo è pericoloso’ true because p, while she is pointing to the dog

slide-180 2. Something about the dog (and not, say, it’s Shoemaker shadow) explains Ayesha’s change in attitude towards ‘Questo è pericoloso’.

If only we had a theory of interpretation that extended to nonlinguistic referential communication …

slide-181 Earlier I suggested that we need something analogous to the Teleological Stance for nonlinguistic refernetial communication.

I also observed that it seems not straightforward to provide such an inference. But thinking about the your-goal-is-my-goal strategy suggests an alternative approach …

slide-182 Nonlinguistic communicative gestures are associated with their targets. A chimpanzee will follow a pointing finger to a container (see Moll and Tomasello 2007, p. 6), children can see where an arrow is pointing (compare Leekam et al. 2010), and even prelinguistic six month olds associate some words with objects (Tincoff and Jusczyk 1999, 2011). (There is debate on when infants are first able to rapidly associate words with novel objects (Werker et al. 1998; Friedrich and Friederici 2011).)

slide-184 Object choice task: The chimpanzee follows the point but doesn’t open the container the experimenter is pointing to.

slide-185 Object choice task: [replica not arrow: Leekam et al 2010, p. 117] The 2- and 3-year-olds can associate the replica with the corresponding container but don’t make use of the association when there is a neutral face.

And they don’t follow the arrow when the face is absent (although they will follow a pointing finger).

slide-186 So what would be sufficient?

slide-191 Csibra’s ‘two stances’:

Teleological and referential action interpretation ‘rely on different kinds of action understanding’

These are initially two distinct ‘action interpretation systems’ and they come together later in development

(Csibra 2003, p. 456)

References

Butterfill, S. (2012). Joint action and development. *Philosophical Quarterly*, 62(246):23–47.

Carpenter, M. (2009). Just how joint is joint action in infancy? *Topics in Cognitive Science*, 1(2):380–392.

Csibra, G. (2003). Teleological and referential understanding of action in infancy. *Philosophical Transactions: Biological Sciences*, 358(1431):447–458.

Csibra, G. and Gergely, G. (1998). The teleological origins of mentalistic action explanations: A developmental hypothesis. *Developmental Science*, 1(2):255–259.

Davidson, D. (1977 [1984]b). Reality without reference. In *Inquiries into Truth and Interpretation*, pages 215–226. Oxford University Press, Oxford.

Davidson, D. (1980). Towards a unified theory of meaning and action. *Grazer Philosophische Studien*, 11:1–12.

Davidson, D. (1984a). *Expressing Evaluations: The Lindley Lectures (monograph)*. University of Kansas, Lawrence.

Davidson, D. ([1984] 1973). Radical interpretation. In *Inquiries into Truth and Interpretation*, pages 125–139. Oxford University Press, Oxford.

Davidson, D. (1994). Radical interpretation interpreted. *Philosophical Perspectives*, 8:121–128.

Davidson, D. (1995). Could there be a science of rationality? *International Journal of Philosophical Studies*, 3(1):1–16.

Dennett, D. (1987). *The Intentional Stance*. MIT Press, Cambridge, Mass.

Friedrich, M. and Friederici, A. D. (2011). Word learning in 6-Month-Olds: fast Encoding–Weak retention. *Journal of Cognitive Neuroscience*, 23(11):3228–3240.

Leekam, S. R., Solomon, T. L., and Teoh, Y. (2010). Adults' social cues facilitate young children's use of signs and symbols. *Developmental Science*, 13(1):108–119.

Moll, H. and Tomasello, M. (2007). Cooperation and human cognition: the vygotskian intelligence hypothesis. *Philosophical Transactions of the Royal Society B*, 362(1480):639–648.

Pacherie, E. (2011). Framing joint action. *Review of Philosophy and Psychology*, 2(2):173–192.

Tincoff, R. and Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-Month-Olds. *Psychological Science*, 10(2):172–175.

Tincoff, R. and Jusczyk, P. W. (2011). Six-Month-Olds comprehend words that refer to parts of the body. *Infancy*, 17(4):432–444.

Tollefsen, D. (2005). Let's pretend: Children and joint action. *Philosophy of the Social Sciences*, 35(75):74–97.

Werker, J. F., Cohen, L. B., Lloyd, V. L., Stager, C., and Casasola, M. (1998). Acquisition of Word-Object associations by 14-Month-Old infants. *Developmental Psychology*, 34(6):1289–1309.