# Interacting Mindreaders

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

December 14, 2011

**Abstract**

***

## 1. Mindreading: the normative project

The question is what constitutes evidence.

Mostly people have supposed that the evidence is available to pure observers — being able to interact with the targets of mindreading confers no in principle advantage.

In challenging this thesis we want to start with goal ascription, which is either a precursor to or a component of mindreading. Recent interest in non-propositional precursors to mindreading makes this appropriate.

## 2. Goal ascription

Purposive action is action directed to the realisation of one or more outcomes. Goal ascription is the process of identifying which outcomes others' purposive actions are directed to. Because goal ascription has received little attention, we first review some of the benefits that being able to identify others' goals brings.

It is a familiar idea that goal ascription enables one to learn from others' successes. For example, if you know or can guess that another agent's actions are directed to opening a nut, you may then be in a position to infer that the unfamiliar pattern of actions she is performing constitute a means to open nuts.[1] Or (in a different case) knowing the goal of another agent's actions

[1] *** (Horner & Whiten 2005)

may enable you to discover a new use for a tool.[2] A slightly less familiar idea is that goal ascription enables one to learn from others' failures as well as their successes. For example, suppose that while you are searching for some peanuts another agent attempts but fails to reach for a closed container. In some circumstances, if you know that the goal of the agent's action was to obtain the peanuts then you now have evidence as to where they might be.[3] This is one illustration of how goal ascription could in principle enable us to learn from others' failures.

Goal ascription enables one to predict and manipulate others' actions. ***if goal involved manipulating an object, you can predict where agent will go at some point; you can also manipulate the agents' action by revealling the location of the object. Eg. Liszowski pointing

Goal ascription is also instrumental for mindreading: knowing which outcomes an action is directed to may constrain hypotheses about what an agent intends as well as potentially providing information concerning what the agent knows, believes or desires. For example, if we know that the goal of an agent's action is to retrieve some peanuts, and if we also know where all the peanuts are, we may be able to infer that she does not know where the peanuts are, or that she falsely believes that some of the peanuts are over there.[4] (Of course this can also work the other way: information about an agent's beliefs or other mental states may support conclusions about her goals. Belief- and goal-ascriptions are mutually constraining.)

Despite the close connection between goal ascription and mindreading, goal ascription does not necessarily involve representing representations. To see why not we first need to be careful about the term 'goal'. *** [do the detour here]

The fact that goal ascription does not involve metarepresentation raises the possibility that goal ascription is possible independently of mindreading. Consider an agent who has no communicative skills and no metarepresentational abilities. Nothing prevents her from *representing* goals, of course. But could she actually ascribe goals? Is there any evidence which could be available to her and which would support goal ascriptions? Some philosophers have argued[5] or implied[6] that there could not be. On their view, the goal ascription and mindreading are interdependent in this sense: there is no evidence for hypotheses narrowly about goals, only evidence for more complex

[2] There is currently much

[3] Hare & Tomasello (2004) exploit this principle in testing chimpanzees' abilities to ascribe goals.

[4] Wimmer & Mayringer (1998) exploit this possibility in testing children's abilities to ascribe false beliefs.

[5] *Bennett

[6] *Davidson

hypotheses concerning both goals and mental states (such as beliefs). So, on this view, those without metarepresentational abilities are not in a position to know the goals of other agents' actions.

## 3. The problem of opaque means

While we lack a detailed theory of the evidential basis of goal ascription, it is certain that the evidence for goal ascription sometimes includes considerations about which ends actions are means to. Suppose an observer faces an action but cannot identify ends to which it could be a means. This may prevent her from recognizing the action's goal[7] by depriving her of evidence. To illustrate, contrast two cases of tool use. In one case, someone uses a reamer to juice a lime; in the other, someone else scores shag with a lame to prevent a loaf from cracking. Without communication, repetition or convention, an observer familiar with reamers but not lames may be able to identify the goal of the first action only. As this illustrates, ignorance about to which ends actions are means can be an obstacle to goal ascription.[8] Call this the problem of opaque means.

Some of the most plausibly unique aspects of human cognition depend on our abilities to recognise the goals of novel behaviours involving tools, and of communicative gestures. If goal ascription is based on entirely observation (so that the possibility of interaction is ignored), the problem of opaque means is likely to arise in both cases. We have just seen an illustration of how the problem of opaque means arises where tools are used to unfamiliar ends. Relatedly, it is also likely to arise where actions involve multiple steps that do not form a familiar sequence, can occur in various orders and can be interspersed among other activities; as in preparing spirit from grain, for example.

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, consider an experiment by ***ref whose two main conditions are depicted in figure *fig. The pictures in the figure stand for what participants, who were chimpanzees, saw. The question was whether participants would be able to work out which of two containers

[7] It is possible that some actions have more than one goal. To reduce parenthetical qualifications we shall write as if actions had only one goal. All of our key claims and arguments are consistent with the possibility of actions with more than one goal.

[8] This is not to say that no observer could ever identify the goal of any action she fails to recognise as a means to achieving that goal. Surely opaque means are not in every case an insurmountable obstacle to goal ascription. Our point is just that opaque means *sometimes* deprive observers of evidence and so prevent goal ascription.

concealed a reward. In the condition depicted in the left panel, participants saw a chimpanzee trying but failing to reach for the correct container. Participants had no problem getting the reward in this case, suggesting that they understood the goal of the failed reach. In the condition depicted in the right panel, a human pointed at the correct container. Participants did not reliably get the reward in this case, suggesting that they failed to understand the goal of the pointing action. This may be because of the problem of opaque means. One theoretically possible explanation of these findings is that the participants could identify to which end a failed reach might be a means, but not to which end a communicative gesture might be a means.[9] Whatever the truth about the chimpanzees' performance, this possibility illustrates how the problem of opaque means can be an obstacle to exploiting communicative gestures.

***FIGURE***

This, then, is the problem of opaque means: failures to identify to which ends actions are means can impair goal ascription. The problem is potentially a problem for interpreters given the standard, purely observational models of interpretation. It is not our intention to suggest that the problem of opaque means is a problem *for models of interpretation.* Rather, it is a potential problem *for interpreters.* The existence of this problem shows, we shall argue, that models of interpretation based on pure observation are less powerful than models taking into account the possibility of interaction. They are less powerful in this sense: some routes to knowledge of the goals of actions are available only where a model of interpretation takes the possibility of interaction into account. In what follows we shall suggest that abilities to engage in joint action with others provides a route to knowledge of the goals of other agents' actions which avoids the problem of opaque means. This is one way in which moving away from a purely observational model of interpretation yields a richer evidential base for ascriptions.

## 4. Joint action

Our overall aim is to defend this claim: there are routes to knowledge of the goals of others' actions which are closed to interpreters who merely observe their targets but open to interpreters capable of interacting with their targets. As a preliminary to defending this claim, we need to specify the types of interaction which are relevant. We shall focus on joint action.

***Much disagreement about what it is. Necessary condition is distributive goal. If people want to disagree, this is only terminological.

[9] *quote Moll & Tomasello

## 5. Your-goal-is-my-goal: a route to knowledge

If an interpreter is able to interact with her targets, if she is not limited to merely observing them, how might this provide her with a route to knowledge of the goals of their actions? The intuitive idea we started with hinges on joint action, a particular form of interaction in which the actions of two or more agents are directed to a single goal (see section 4 on the preceding page). Our intuitive idea was this: if an interpreter is engaged in joint action with her target, it's 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 in joint action with her target, then she already knows what the goal of her target's actions are.

Of course this intuitive idea won't work as it stands. For the inference it captures relies on the premise that the interpreter and her target are engaged in joint action. But for the interpreter to know this premise—for her to know that they are engaged in joint action—it seems she must already know which goal her target's actions are directed to. Worse, on many accounts of joint action the truth of the premise entails that the interpreter knows the goal of her target's actions.[10] Apparently, then, engaging in joint action presupposes, and therefore cannot be a source of, knowledge of others' goals.

Fortunately there is a way around this. For there are various cues which signal that one agent is prepared to engage in joint action with another. Seeing a new parent struggling to get his twin pram on to the bus, you grab the front wheels and make eye contact, raising your eyebrows and smiling. In this way you signal both that you are disposed to help and that are about to engage in joint action with the stranger. Since the stranger is fully committed to getting his pram onto the bus, he knows what the goals of his own actions will be. This makes enables him to infer the goal of your immanent actions: your goal is his goal, to get the pram onto the bus.

Our suggestion, then, is that the following inference characterises a route to knowledge of others' goals:

1. We are about to engage in some joint action[11] or other

2. I am not about to change which goal my actions will be directed to.

Therefore:

3. The goal of your actions will be my goal.

[10] *ref

[11] *What notion of joint action is needed here? Any will do as long as it involves distributive goals.

Call this the *your-goal-is-my-goal inference*. To say that this inference characterises a route to knowledge implies two things. First, in some cases it is possible to know the premises, 1–2, without already knowing the conclusion, 3. Second, in some cases knowing the two premises puts one in a position to know the conclusion. We shall consider these points in turn

Is it ever possible to know the premises without first knowing the conclusion? Consider the first premise. Sometimes in the right contexts you can see in others' facial expressions, engaging gestures or synchronized bodily movements that they are about to engage in joint action with you. Exploiting these indicators does not typically depend on knowing the particular contents of any of their beliefs, desires or goals. Expressions, gestures and movements can naturally indicate imminent jointness in much the way that can also naturally indicate emotions.[12] Of course these indicators provide no guarantee that others are about to helpfully engage in joint action with you. But they are sufficiently reliable to ground knowledge in some cases. So knowing the first two premises of the above inference does not require already knowing which particular goals others have.

On this point not everything needs to rest on indicators of imminent jointness. It is sometimes possible to know that others are about to engage in joint action with you even without relying on such indicators. Thanks to widespread dispositions to act jointly, in some situations it is reasonable to take for granted that others are disposed to be helpful and will act jointly with you. For example, this is often so for children surrounded by family or familiar adults who are having difficulty with a task. And in some cultures parents with struggling small children on public transport (say) can reasonably take for granted that those around them will act jointly with them when the need is clear.

Does knowing the premises of the your-goal-is-my-goal inference sometimes put one in a position to know the conclusion? If you know that others are about to engage in joint action with you, it is sometimes reasonable to infer that the outcome to which the joint action will be directed is the very outcome you are currently concerned with. Given the account of joint action in Section 6 above, this means that the others each have an individual goal whose fulfilment requires this outcome to occur. So knowing the outcome to which a joint action will be directed puts one in a position to know something about which goals the others have. Of course the inference this rests on is not deductive and will only be correct when certain background conditions are met. These background conditions include the others having largely true beliefs concerning which outcome you are concerned with; I return to this point below.

[12] Ideas along these lines are suggested by the discussion of *emergent coordination* in Knoblich, Butterfill & Sebanz (2010).

# 6. The problem of false belief [move to later]

We have just seen that failure to identify to which ends actions are means can impair goal ascription; this was the problem of opaque means. Another problem affecting goal ascription given purely observational models of interpretation arises from the interdependence of beliefs and goals. To illustrate, imagine sitting at a table on which two closed opaque boxes each contain an object; one contains an owl, the other a cat. If the goal of your action is to retrieve the owl, and you believe that the owl is in the north-most box, then (unless things are going very badly) you will reach into the north-most box. But of course if you had believed instead that the owl was in the other box, then, in acting on the same goal, you would have reached into the other box. Now consider someone observing your actions. If she is has good reason to believe, falsely, that you know the owl is in the south-most box, then she may be justified in supposing, incorrectly, that when you reach into the north-most box the the goal of your action is to retrieve the cat. As this illustrates, differences in belief between observers and protagonists can be an obstacle to goal ascription. Call this the problem of false belief.

There is disagreement over whether it is possible to knowledgably identify the goals of an agent's actions without also ascribing some beliefs to that agent. Bennett (1976, pp.48–50) and Davidson (1984) both appear to hold that this is impossible, that identifying goals cannot be done independently of ascribing beliefs. By contrast, Gergely, Nadasky, Csibra & Biro (1995), *Meltzoff?, Baillargeon, Scott & He (2010) and Woodward (1998) (among many others) appear to assume the opposite, that it is possible to identify goals without even being able to ascribe beliefs. For what it is worth, we tentatively favour this latter position.[13] However, this debate is not directly relevant to our concerns here. Both sides can agree that differences in belief between observer and protagonist are sometimes an obstacle to goal ascription. This is all that the problem of false belief requires.

In what follows we shall argue that abilities to engage in joint action provide an interpreter with a route to knowledge of the goals of other agents' actions which does not depend on her abilities to ascribe beliefs. This is not

[13] It is striking that, as far as we can tell, neither Bennett nor Davidson offers an argument for this claim. They do note that beliefs and goals make an interdependent contribution to observed action. But this by itself does not show that goal ascription cannot in some cases involve justifiably ignoring the possibility of differences in belief between interpreters and their targets. For instance, suppose that two people are sitting opposite each other at a low table which is sparsely populated with objects. The objects are all out in the open; manifestly, both can clearly see them. If one reaches to grasp one of these objects (the duck, say), must the other ascribe beliefs in order to knowledgeably identify the goal of her action? On the face of it, she need not. Even if she had no ability to ascribe beliefs, she might nevertheless be in a position to acquire knowledge of the goal of the other's action.

because abilities to engage in joint action provide a way to avoid the problem of false beliefs altogether. Rather, as we shall see, abilities to engage in joint action provide a way to shift the burden of resolving the problem of false belief from an interpreter to her target.

## 7. OLD

These simple facts about goal ascription raise many questions. Some concern mechanism, how in fact one subject is able to discover facts about which outcomes another agent's actions are directed to. Another set of questions focuses on the evolution of goal ascription and the costs and benefits of being able to ascribe goals and of being a potential target of goal ascription. Our concern here is not directly with any of these questions. Instead we shall focus on a more narrowly epistemic question. What evidence could support hypotheses about the outcomes to which actions are directed? And how would the evidence support the hypotheses?[14]

Of special interest is evidence available independently of any knowledge of mind or language. We want to know how it is possible to identify goals even without knowing what an agent believes or desires and even without understanding their communicative actions. Accordingly we will adopt the perspective of a goal ascriber who knows nothing about the mental states of her target agent that would distinguish this agent from any other. We will also stipulate that there is initially no common ground, shared culture or conventions. And we will stipulate that the goal ascriber is initially unable to understand any communicative actions.

There are two sorts of motivation for these restriction on the evidential basis. One is simply that developmental and comparative research indicates that goal ascription does appear to take place in such circumstances.[15] This makes it important to understand the evidence on which such ascriptions could be based. (Of course identifying evidence that could support such ascriptions would not all by itself enable us to explain how goal ascriptions are actually made, but identifying evidence is necessary if we are ever to explain the reliable success of mechanisms for goal ascription.) Another source of motivation is the conjecture that goal ascription is a prerequisite for the more sophisticated mindreading activities which reveal mental states and meanings. The coherence of this conjecture depends on the possibility of

[14] These questions are versions of those Davidson constructs a theory of interpretation to answer. While what follows draws on Davidson's insights, our aims here are more modest than his. For we are concerned only with a fraction of the problem of ascribing mental states and meanings; and, unlike Davidson, we are not concerned with larger claims about the nature of mind. See Davidson (1973, 1990); Lepore & Ludwig (2005).

[15] *refs

knowing something about which outcomes an agent's actions are directed to independently of knowing what she believes or desires and independently of understanding her communicative actions.[16]

So what evidence could support goal ascription by someone who knows nothing discriminating about her targets' mental states or communicative actions? Ordinary third-person goal ascription, simplified and idealized, works like this.[17] Faced with an action, the would-be goal ascriber first asks which outcomes this action could be a means to realising. She then considers which of these outcomes are potentially beneficial for, or desirable to, the agent. Any such outcomes are identified as goals to which the action is directed. So the fact that an action is a means to realising some outcome which is potentially beneficial or desirable is evidence for the conclusion that this outcome is one to which the action is directed. Schematically, the proposal is that:

**($E_1$)** Action $a$ is a means of realising outcome $G$.

and:

**($E_2$)** The occurrence of outcome $G$ is potentially beneficial for, or desirable to, the agent of $a$. (And there is no other outcome, $G'$, which action $a$ is a means of realising and which would be more beneficial for, or more desirable to, the agent of $a$.)

jointly constitute evidence for the conclusion that:

**(C)** $G$ is a goal to which action $a$ is directed.

This proposal might be extended in various ways. For instance, Southgate, Johnson and Csibra offer a 'principle of efficiency' according to which:

> 'goal attribution requires that agents expend the least possible amount of energy within their motor constraints to achieve a certain end.' (Southgate, Johnson & Csibra 2008, p. *)

[16] *Compare and contrast Davidson? He did think relational attitudes (holding true) are the foundation for interpretation. But he also thought that interpretation had to happen all at once.)

Dennett (1987, p. 17) 'Here is how it works: 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 your predict that this rational agent will act to further its goals in the light of its beliefs. A little practical reasoning from the chosen set of beliefs and desires will in many—but not in all—instances yield a decision about what the agent ought to do; that is what you predict the agent will do.'

[17] *ref? Dennett?

If this is a correct principle of goal attribution, we could extend the proposal above to incorporate it:

> **(E$_3$)** No alternative action, $a'$, is a means to realising outcome $G$ and would involve expending less energy than $a$.

Now the proposal is that (E$_1$) to (E$_3$) are jointly evidence for (C).

In at least some cases goal attribution is likely to be more complicated than this proposal allows. To illustrate, note that some agents may weigh the efficiency of alternative actions against their possible side effects and how reliable they would be as a means to realising an outcome. Where this is true, identifying the evidential basis for goal ascription may require a similar weighing of these factors in inferring backwards from actions to their goals.[18] Specifying exactly what should be weighed and how is beyond the scope of this paper, (and may also be something which varies between species of agent). We can mark the gap with an alternative to (E$_3$) which uses an unspecified notion of 'better' as a placeholder:

> **(E$_{3'}$)** No alternative action, $a'$, is a better means to realising outcome $G$.

This, then, is the standard approach to answering our question about goal attribution: (E$_1$), (E$_2$) and (E$_{3'}$) jointly constitute evidence for (C) given that these approximate conditions under which it would be rational to perform $a$ in order to realise $G$ and given that agents approximate to performing $a$ in order to realise $G$ rather than any other outcome under these conditions.

## References

Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences*, *14*(3), 110–118.

Bennett, J. (1976). *Linguistic Behaviour.* Cambridge: Cambridge University Press.

Csibra, G., Bíró, S., Koós, O., & Gergely, G. (2003). One-year-old infants use teleological representations of actions productively. *Cognitive Science*, *27*(1), 111–133.

[18] This is loosely related to what Csibra and Gergely call 'the principle of rational action'. As they formulate the principle, '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 & Gergely 1998, p. *; cf. Csibra, Bíró, Koós & Gergely 2003).

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

Davidson, D. (1974 [1984]). Belief and the basis of meaning. In *Inquiries into Truth and Interpretation* (pp. 155–170). Oxford: Oxford University Press.

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

Davidson, D. (1990). The structure and content of truth. *The Journal of Philosophy*, *87*(6), 279–328.

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

Gergely, G., Nadasky, Z., Csibra, G., & Biro, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, *56*, 165–193.

Hare, B. & Tomasello, M. (2004). Chimpanzees are more skilful in competitive than in cooperative cognitive tasks. *Animal Behaviour*, *68*(3), 571–581.

Horner, V. & Whiten, A. (2005). Causal knowledge and imitation/emulation switching in chimpanzees (pan troglodytes) and children (homo sapiens). *Animal Cognition*, *8*, 164–181.

Knoblich, G., Butterfill, S., & Sebanz, N. (2010). Psychological research on joint action: Theory and data. In B. Ross (Ed.), *Psychology of Learning and Motivation*, volume 51. Academic Press.

Lepore, E. & Ludwig, K. (2005). *Donald Davidson: Meaning, Truth, Language, and Reality.* Oxford University Press.

Southgate, V., Johnson, M. H., & Csibra, G. (2008). Infants attribute goals even to biomechanically impossible actions. *Cognition*, *107*(3), 1059–1069.

Wimmer, H. & Mayringer, H. (1998). False belief understanding in young children: Explanations do not develop before predictions. *International Journal of Behavioral Development*, *22*(2), 403–422.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*, 1–34.