Interacting Mindreaders

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

December 1, 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.

Goal ascription enables one to predict and manipulate others' actions, and to learn from their failures and 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. This is one illustration of how goal ascription could in principle enable us to learn from others' failures.

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.[1]

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[2] or implied[3] 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 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. 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?[4]

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

---

[1] Of course this can also work the other way: belief- and goal-ascriptions are mutually constraining.

[2] *Bennett

[3] *Davidson

[4] 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).

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.[5] 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 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.[6]

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.[7] 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

---

[5] *refs

[6] *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.)

Bennett (1976, pp.48–50) suggests that a theory of goals ascription has to be developed together with a theory of (proto-)belief ascription: 'An animal's behaviour does not show what it registers unless we know what it seeks; but how can we learn what it seeks before we know what it registers?' [p. 48]

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.'

[7] *ref? Dennett?

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. *)

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.[8] 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:

[8] 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).

($E_{3'}$) 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

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.

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

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.

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.