Mindreading and Joint Action: Philosophical Tools Notes for lectures, CEU Fall 2012-13

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

September 10, 2012

Abstract

***

Keywords: Mindreading, Joint Action, Action, Belief, Intention, Representation, Mental State

1. Mindreading and Tracking

1.1. Mindreading

Mindreading is the process of identifying mental states and actions as the mental states and actions of a particular subject on the basis, ultimately, of bodily movements and their absence, somewhat as reading is the process of identifying propositions on the basis of inscriptions (Apperly 2010, p. 4).

1.2. Tracking mental states

It is useful to contrast mindreading with tracking.

An ability to *track* mental states of a particular kind (e.g. beliefs) is ability that exists in part because exercising it brings benefits obtaining which depends on exploiting or influencing facts about others's mental states.

To track beliefs (say) is to exercise a belief-tracking ability.

To illustrate, suppose that Hannah is able to discern whether another's eyes are in view, that Hannah exercises this ability to escape detection while stealing from others, that Hannah's ability exists in part because it benefits her in this way, and that Hannah's escaping detection depends on exploiting

a fact about other's visual representations (namely that they usually cannot see Hannah's acts of theft when Hannah doesn't have their eyes in view). Then Hannah has an ability to track visual representatios. (This is not supposed to be a plausible, real-world example but only to illustrate what the definition requires.)

Sometimes tracking involves mindreading, but not always (as the story about Hannah and the eyes illustrates).

Tracking mental states without mindreading is probably quite common. Another example: preening (?).

2. That we do not understand what mindreading is

Empirical questions about mindreading include:

- When in development does mindreading first occur?
- What representations and processes make mindreading possible?
- Is mindreading automatic?
- Which animals are capable of mindreading?

Much progress has been made on these questions, and there is more still to make. I want to suggest that there is also an obstacle to progress. The obstacle is that we don't adequately understand what mindreading is.

Why think that we don't adequately understand what mindreading is? The strongest reason is this. Some apparently puzzling patterns in findings about mindreading can be resolved by thinking carefully about what mindreading is. But we'll only be in a position to evaluate this claim right at the end, when we have reflected on what mindreading is.

There are, though, some hints that we might not adequately understand what mindreading is. As we'll see, there are controversies concerning what mental states are, and what actions are. But when the topic is mindreading, these controversies are usually ignored and it is assumed that we all know what actions and mental states are. To better understand what mindreading is we will need to reflect on what actions and mental states are.

So my plan is to step back from empirical questions about mindreading and first focus on more narrowly philosophical issues about what mindreading is. Having done this, we'll come back to the empirical questions about mindreading. The philosophical part is valuable to the extent that it supports progress with questions about when, how and where mindreading occurs.

But you might still be sceptical that philosophy is really needed. Do we really not adequately understand what mindreading is? You probably shouldn't take my word for it. After all, not understanding things is what

I do for a living. So consider these questions [*more refined version in the plan for these lectures]:

- What evidence could in principle support the ascription of a particular belief to a given subject, and how does the evidence support the ascription?
- *Objectivity* Could a mindreader be able to identify beliefs despite not understanding what it is for a belief to be true or false?
- *Self-awareness* Does being a mindreader entail being able, sometimes, to identify one's own mental states and actions?
- Could there be mindreaders who can identify intentions and knowledge states but not beliefs?
- Does identifying an action necessarily involve representing an intention?

If we fully understood what mindreading was, we would be able to answer these questions in a principled way. The fact that we can't shows that we don't fully understand what mindreading is. And it suggests that we don't adequately understand it either.

To better understand what mindreading is we have to take a step back and ask what actions are and what mental states are.

3. What are mental states?

mental state =

subject [e.g. Ayesah]
+
attitude [e.g. desire]
+
content [e.g. that Ayesha eats ice cream]

The subject is just an object.

Explain attitude and content using 2 x 2:

The attitude is normally specified by its functional and normative roles, and these are usually explained in contrast with those of other attitudes. E.g. What distinguishes believing from supposing? These have related roles in guiding action. Velleman (*ref) suggests that believing differs from supposing in aiming at truth. We'll return to this idea later.

To specify the content we first need to identify something about its structure. Mental states are usually thought of as having propositional contents. But there is a variety of types of content that a mental state can have. For

	attitude		
	belief	desire	…
Ayesha eats ice cream	1	3	…
Frederique writes poetry	2	5	…
…	…	…	…

Table 1: Attitude versus content

instance, you can have an attitude towards a map-like structure, an image, an event-type, an object or a relation.

***examples (e.g. use navigation for attitudes towards maps?)

*Explain what propositions are (like numbers).

*Also explain different types of propositions (Russellian, Fregean &c)

*illustrate limits of different kinds of content (compare with different kinds of number)

4. The origin of the attitudes

Take an attitude like belief or desire. Suppose someone offers a partial characterisation of the attitude. For instance, suppose they say that belief aims at truth whereas desire aims at satisfaction. What is this partial characterisation answerable to? On what grounds should we accept or reject it?

We might treat claims about the attitudes as merely terminological stipulations, so that the only requirement is coherence. This serves only to push back the question further. What are we attempting to capture in characterising an attitude?

Another possibility is to think of claims about the attitudes as answerable to ordinary thinking about mental states. While I doubt we can escape ordinary thinking entirely, I think we should be cautious in appealing to it for two reasons. One is that we don't actually know very much about how people ordinarily think about mental states. The other is that ordinary thinking about mental states may not be right, or even consistent.

The approach to the attitudes I prefer is modelling. This is going to take a while to explain but the idea is simple. Decision theory provides us with a model capable of explaining, within limits, the preferences that agents manifest. The model involves subjective probabilities and desirabilities, which roughly resemble belief and desire in some ways (Jeffrey 1983, p. 59, Davidson 1985). So the model potentially provides two things. One is a fairly precise characterisation of belief-like and desire-like attitudes. The other is an explanation of when postulating them is justified. Justification for postulating these states in a particular case depends on how well the model explains

the agents' preferences.

4.1. Actions, outcomes and conditions

That was a bit abstract. Let's get into details. (What follows is based on Jeffrey 1983; it's not my own work.) I'm going to go very slowly at the start. This will be a bit painful if you're already familiar with decision theory, but it's worth it because we will use the basic ideas more than once. (This will be important again in the context of joint action.)

Imagine we are deciding between two *actions*, cycling to the seminar or catching the bus. Let's also suppose that we are only interested in two types of *outcome* these actions could have, staying dry versus not staying dry and getting exercise versus not getting exercise. Among the various possible outcomes, our preference ranking is:

1. getting exercise and staying dry
2. not getting exercise and staying dry
3. getting exercise and not staying dry

Suppose we know that cycling will result in the third outcome, 3, whereas getting the bus will result in the second outcome, 2. Then we should get the bus. In this situation, actions guarantee outcomes so how we act should depend just on which outcome we prefer.

Very often actions are not so simply linked to outcomes. We often don't know whether we will get wet if we cycle. Whether we get wet depends on further *conditions*, such as whether it rains or whether there is flooding. In general, which outcome occurs depends both on the actions we choose and on the conditions we encounter.

	condition	
action	no flooding	flooding
cycle	get exercise and stay dry	get exercise and get wet
take bus	get no exercise and stay dry	get no exercise and stay dry

Table 2: Outcomes depend on actions and conditions

Usually we are not certain about all the conditions relevant to a decision. Instead we have to form a view about their probability. For example, we might know that there is a fair chance of flooding without knowing outright that there is flooding. This means we don't know for sure which outcome

cycling will result in. It might result in our most preferred outcome, getting exercise and staying dry; but it might also result in our least preferred outcome, getting exercise and getting wet.

For this reason, in deciding what to do we should ideally take into account both our preferences concerning the outcomes and the probabilities of the conditions obtaining. This could be done as follows.

We first consider the probabilities (see table 3). The top left cell represents the probability of no flooding if we choose to cycle. Of course, the probability of flooding is independent of whether we cycle or take the bus. But the approach is flexible enough to accommodate cases where our actions can influence the probability of the conditions ocurring.

	condition	
action	no flooding	flooding
cycle	probability of no flooding if we cycle: 0.3	probability of flooding if we cycle: 0.7
take bus	probability of no flooding if we get the bus: 0.3	probability of flooding if we get the bus: 0.7

Table 3: Probabilities of four conditions obtaining

We then assign weights to the outcomes that reflect how desirable they are in relation to each other (see table 4).

	condition	
action	no flooding	flooding
cycle	[get exercise and stay dry] desirability of outcome: 3	[get exercise and get wet] desirability of outcome: -1
take bus	[get no exercise and stay dry] desirability of outcome: 1	[get no exercise and stay dry] desirability of outcome: 1

Table 4: Desirabilities of four outcomes

Finally we multiply the probability and desirability matrices (see table 5 on the following page). Adding the rows then gives expected utilities for each action. The idea is that we should perform the act with the greatest expected utility. In this case, cycling gets just 0.2 whereas taking the bus gets 1, so we should take the bus.

I've gone very slowly over some familiar ideas because I want us to attend to the basics. There are three basic elements: actions, conditions and outcomes. Which outcome occurs depends on two things: the action chosen

	condition	
action	no flooding	flooding
cycle	0.3 [probability] ∗ 3 [desirability] = 0.9	0.7 ∗ -1 = -0.7
take bus	0.3 ∗ 1 = 0.3	0.7 ∗ 1 = 0.7

Table 5: Multiplying probabilities by desirabilities

and the conditions that obtain. Desirabilities attach to outcomes. Probabilities attach to conditions. Expected utilities attach to actions. Expected utilities can be derived from desirabilities and probabilities. One procedure for choosing an action is to compute expected utilities (in the way illustrated) and then perform the action with the highest expected utility.

4.2. Reversing direction

For simplicity I have so far been speaking as if we wanted to introduce a procedure for deciding how to act. But of course that isn't our aim at all. Our ultimate aim is to be able to justify claims about attitudes—or at least to understand what sort of considerations might provide justification. We still have a way to go.

So far we have seen how subjective probabilities and desirabilities determine expected utilities for actions. If we knew the expected utilities of the actions available to an agent, we could make a prediction about how she will act—about whether she will cycle or take the bus, say. But to work out an agent's expected utilities we would need to know the probabilities she assigns to relevant conditions and how desirable she finds the various outcomes. How could we know this? **What is the evidential basis for ascribing an agent subjective probabilities and desirabilities and how does the evidence support the ascriptions?**

Suppose we knew these things:

1. The agent will always perform an action with at least as much expected utility as any other actions available to her.

2. The agent has just two actions available to her: she can either cycle or get the bus.

3. The agent is getting the bus.

Then we also know that the agent assigns the same, or higher, expected utility to getting the bus than to cycling. So given some assumptions, an agent's actions reveal her expected utilities.

But this doesn't reveal much about the agent's preferences or probabilities. After all, expected utility is a function of both. The agent might be getting the bus because she does not particularly desire exercise. Or the agent might be getting the bus *despite* particularly desiring exercise because she assigns a high probability to flooding (and a low desirability to staying dry).

On the face of it then, an agent's actions might reveal at most her expected utilities (given assumptions listed above) and leave us blind to her subjective probabilities and desirabilities.

If we knew the agent's subjective probabilities then, given some assumptions, we could work out her desirabilities from her expected utilities. Someone who thinks flooding is likely but nevertheless cycles when she could have taken the bus must desire exercise more than staying dry.

***HERE *** intuitive explanation of Ramsey's criterion.

4.3. Ramsey's insight

Ramsey's criterion:

> 'Suppose that A and B are consequences [outcomes] between which the agent is not indifferent, and that N is an ethically neutral condition [i.e. the agent is indifferent between N and not N]. Then N has probability 1/2 if and only if the agent is indifferent between the following two gambles.
>
> B if N, A if not
> A if N, B if not' (Jeffrey 1983, p. 47)

4.4. What is subjective desirability? What is subjective probability?

So far I have also not said what the desirabilities and probabilities are. You may have some intuitions about what these represent. In particular, you may notice that they are intuitively related to desires and beliefs. But strictly speaking so far we should treat as unresolved the issue of what these desirabilities and probabilities are. Our aim is to use the desirabilities and probabilities to elucidate attitudes like belief and desire. So it would clearly be a mistake to rely on those attitudes in saying what desirabilities and probabilities are.

4.5. Towards a model

We want a model to explain actions.

Here's the idea in outline. We want a method that will enable us to assign subjective desirabilities and subjective probabilities to an agent by observing

some of their actions. And we then want to be able to *predict* their actions using the assigned subjective desirabilities and subjective probabilities.

4.6. A gap

***So far we have a theory of how attitudes relate to action but we do not have a theory about how they are acquired (a theory of belief fixation).

> ‘A theory of mind needs a story about mental processes, not just a story about mental states. … the logical behaviourism of Wittgenstein and Ryle had, as far as I can tell, no theory of thinking at all (except, maybe, the silly theory that thinking is talking to oneself). I do find that shocking. How could they have expected to get it right about belief and the like without getting it right about belief fixation and the like?’ (Fodor 1998, p. 9–10)

> ‘modern philosophers … have no theory of thought to speak of. I do think this is appalling; how can you seriously hope for a good account of belief if you have no account of belief fixation?’ (Fodor 1987, p. 147)

References

Apperly, I. A. (2010). *Mindreaders: The Cognitive Basis of “Theory of Mind”*. Hove: Psychology Press.

Davidson, D. (1985). A new basis for decision theory. *Theory and Decision*, *18*, 87–98.

Fodor, J. (1987). *Psychosemantics*. Cambridge, Mass.: MIT Press.

Fodor, J. (1998). *Concepts*. Oxford: Clarendon.

Jeffrey, R. C. (1983). *The Logic of Decision, second edition*. Chicago: University of Chicago Press.