DOI: 10.1111/phis.12155

ORIGINAL ARTICLE

Philosophical Issues
A Supplement to NOÛS

Belief, desire and the prediction of behaviour*

José L. Zalabardo

University College London

Correspondence
José L. Zalabardo, University College London.
Email: j.zalabardo@ucl.ac.uk

1 | MEANING AND USE

Sometimes philosophers believe, rightly or wrongly, that a region of (declarative) discourse is in need of their help—that vindicating its standing requires the kind of treatment that a philosopher might be able to offer. There is a wide variety of circumstances leading to this situation. In many cases, a perceived tension between naturalism and the ostensible subject matter of the discourse is an important factor,[1] but there could in principle be lines of reasoning leading to the conclusion that a discourse needs philosophical help in which naturalism is not involved. I am going to discuss the situation focusing on linguistic items—sentences and terms, but parallel problems arise for their mental correlates—beliefs and concepts, and I'll be interested only in proposals for the linguistic version of the task that don't presuppose that the mental version has already been accomplished.

There are several general approaches that a philosopher can adopt to discharge this kind of task. One that has attracted some attention in recent years seeks to explicate a discourse in terms of features of the practice in which the discourse is embedded—of what speakers do with the expressions in the discourse. Its central insight is the Wittgensteinian thought that the semantic properties of linguistic expressions are determined by how they are used.[2] There are several ways of articulating this insight. The one I want to focus on proposes to vindicate the discourse by identifying features of the practice in which it is embedded that are necessary and sufficient for generating its semantic properties: a discourse will have the semantic properties of the target discourse if and only if it is embedded in a practice in which these features are present. I'm going to refer to features of a linguistic practice that can play this role with respect to its discourse as *pragmatic grounds*,[3] and to the project of vindicating

*I'm grateful to Joshua Gert, Javier González de Prado Salas, Lisa Miracchi, Johannes Roessler and Matthew Simpson for comments on this material. I've also benefitted from conversations or correspondence with Renée Baillargeon, Gergely Csibra, Luke Fenton-Glynn and Daniel Rothschild. I've presented versions of this material at a conference on Applied Neo-Pragmatism at the College of William and Mary, a conference on Expressivisms, Knowledge and Truth at UCL, a workshop on Points of View at the University of La Laguna, and the 24th World Congress of Philosophy in Beijing. I am grateful to these audiences.

a discourse by specifying its pragmatic grounds as the *pragmatist approach*.[4] Here's how the approach is characterised by Robert Brandom, one of its leading advocates:

> *[…] to describe what the discursive practitioners who deploy the vocabulary in question* do*, the practices they engage in, or the abilities they exercise, in virtue of which they count as using that vocabulary. The idea is to formulate […] necessary and sufficient conditions for doing what one needs to be doing in order thereby to be saying what can be said using the vocabulary […]. (Brandom, 2013: 85–86)*

The pragmatist approach aspires to provide an alternative to a standard conception of how to explicate the meaning of linguistic expressions. On this conception, meaning is to be explained in terms of relations that linguistic expressions bear to items in the world—the individuals, properties, etc. that terms denote or the states of affairs that sentences represent as obtaining. Call this approach *referentialism*. If the pragmatist approach is going to provide an alternative to the referentialist approach to a discourse, its pragmatic grounds will have to be specified without appealing to the language-world relations invoked by referentialism.

There are two kinds of question that need to be asked with respect to the pragmatist approach. First, we need to ask whether the specification of pragmatic grounds for a discourse, if achieved, would succeed in providing the kind of vindication that we seek. Second, for any given discourse, we need to ask whether it is possible to specify pragmatic grounds for it—whether we can provide a non-trivial specification of features of the practice in which the discourse is embedded that are necessary and sufficient for the discourse to have the semantic properties it has. The second kind of question has an intrinsic interest that's independent of the aspiration to use pragmatic grounds in a vindication of the discourse.

Here I'm not going to be concerned with the first kind of question. My goal will be to address a question of the second kind, about the availability of pragmatic grounds, in a specific case—with respect to discourse ascribing propositional attitudes. My question is, then, whether it is possible to provide a satisfactory specification of features of the linguistic practice in which our propositional-attitude discourse is embedded that are necessary and sufficient for the semantic properties of the discourse. If these features can be specified, an arbitrary discourse will have the same semantic properties as our propositional-attitude discourse if and only if it is embedded in a linguistic practice in which these features are present. My goal in this paper is to explore a specific feature of the practice of ascribing propositional attitudes that pragmatists might be inclined to include in its pragmatic grounds—the role that they play in the prediction of behaviour. I'm going to consider how this role can yield a necessary condition for a discourse to have the character of propositional-attitude ascription. I will close by briefly considering other factors that the pragmatist might want to include in the pragmatic grounds of propositional-attitude ascriptions.

2 | BELIEF, DESIRE, BEHAVIOUR

I'm going to concentrate on two attitudes—belief and desire, although I won't have a lot to say here about the contrast between these and other cognitive and conative attitudes. This involves a huge simplification of the complex network of concepts that we employ to characterise our mental life. Our ascriptions of beliefs and desires in real life are inseparable from our ascriptions of other mental phenomena, including emotions, tendencies, moods, attention levels, etc. My hope is that by concentrating on belief and desire ascriptions we will be able to isolate features of their pragmatic grounds that will

still play an important role in a more sophisticated account of our ascriptions of mental states that includes the whole range of mental phenomena.

Belief and desire ascriptions have, in general terms, a very simple syntax. They are, or appear to be, relational sentences, in which two binary predicates connect the subject to which the attitude is being attributed with the possible state of affairs that is being put forward as the content of the attitude.[5] The content can be represented by a that clause, but other constructions are also possible, as in "she desires to be left alone" or "he believes the witness's testimony". I will represent them as having the structure "aBp" and "bDq", with "a" and "b" standing for the subjects to which the attitudes are ascribed and "p" and "q" for the possible states of affairs postulated as the contents of the ascribed attitudes.

The pragmatist's task can be characterised in the following terms. Consider all possible linguistic practices involving a discourse with this syntax—assertoric sentences of the forms "aBp" and "bDq", in which "a" and "b" stand for human beings and possibly other individuals, and "p" and "q" stand for possible states of affairs. Call these *BD practices*. The ultimate goal of the pragmatist approach to belief and desire ascriptions is to specify conditions on a BD practice that will be necessary and sufficient for asserting "aBp" (*B-pairing* a with p, as we will put it sometimes) to count as ascribing to a the belief that p obtains, and for asserting "bDq" (*D-pairing* b with q) to count as ascribing to b the desire that q obtains. I'm going to focus here on an aspect of our practice of ascribing beliefs and desires that can be expected to play a central role in the conditions that the pragmatist seeks to specify. I'm going to concentrate on the claim that a BD practice won't count as producing belief and desire ascriptions unless it exhibits this feature.

The feature I have in mind is the role that belief and desire ascriptions play in a familiar strategy for predicting behaviour, as in:

> *Lorna wants to drink water and believes that there's water in the fridge, so I predict that she will open the fridge.*

I'm going to focus on the idea that involvement in this predictive strategy is essential to belief and desire ascriptions—that a BD practice produces belief and desire ascriptions only if it drives the application of this predictive strategy.

My main job in what follows will be to provide a characterisation of the predictive strategy for which this claim can be plausibly made—of the strategy that we deploy when we use belief and desire ascriptions to predict behaviour. I'm going to take as my starting point the characterisation of this predictive strategy that Daniel Dennett has developed under the label *intentional stance*. I'm going to argue that some aspects of Dennett's characterisation are inadmissible for our purposes, but that his central idea can be deployed as the core of a different characterisation that satisfies our requirements.

3 | THE INTENTIONAL STANCE

The idea of explicating belief and desire ascriptions in terms of the role they play in a strategy for predicting behaviour has been championed by Daniel Dennett. He refers to the predictive strategy that he employs for this purpose as the *intentional stance*:[6]

> *Here is how it works: first you decide to treat the object whose behaviour 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. A little practical reasoning from the chosen set of beliefs and desires will in many—but not all—instances yield a decision about what the agent ought to do; that is what you predict the agent* will *do. (Dennett, 1987: 17)*

I think the strategy can be usefully characterised as involving two components—one concerning which beliefs and desires to ascribe and another concerning how to use the beliefs and desires thus ascribed for predicting behaviour:

> *IS1 Ascribe to the agent the beliefs and desires it ought to have.*
>
> *IS2 Predict that the agent will act to satisfy the desires you've ascribed to it in light of the beliefs you've ascribed to it.*[7]

Presumably the idea is that the ascriptions produced under IS1 are open to revision in light of their predictive success when used according to IS2, and the goal will be to strike a balance between ascribing the beliefs and desires the agent ought to have and ascribing those that result in accurate predictions.

Dennett doesn't use the intentional stance to provide a pragmatist account of belief and desire ascriptions. He deploys it instead to provide a referentialist account. It might not be obvious that Dennett's proposal has this character, since he doesn't locate the states of affairs that determine the truth conditions of belief and desire ascriptions where other referentialists have typically looked for them, but it is clear that his goal is to identify the states of affairs that play this role:

> all there is *to really being a true believer is being a system whose behavior is reliably predictable by the intentional strategy, and hence* all there is *to really and truly believing that* p *(for any proposition* p*) is being an intentional system for which* p *occurs as a belief in the best (most predictive) interpretation. (Dennett, 1987: 29)*

As I explained in Section 1, this is not how the pragmatist will want to use the predictive role of belief and desire ascriptions. Dennett uses our behaviour-predicting strategy to provide a criterion for when someone (an interpretee) counts as believing or desiring something. The pragmatist will want to use it to provide a criterion for when someone (an interpreter) counts as ascribing beliefs and desires—for when a BD practice achieves this. This difference has important consequences: Dennett's characterisation of the intentional stance is inadmissible as it stands in the context of the pragmatist project.

The problem concerns IS1—the proposed strategy for selecting the desires and beliefs we should ascribe. If we formulate it as a condition on an arbitrary BD practice, IS1 looks like this:

> *B-pair a with p just in case a ought to believe p*
>
> *D-pair b with q just in case b ought to desire q*

The problem is that whether the participants in a BD practice satisfy this condition depends on whether the occasions on which they B-pair a with p or D-pair b with q are the occasions on which they believe that a ought to believe p or that b ought to desire q. Hence the formulation of the condition depends crucially on facts concerning the participants' beliefs about who ought to believe/desire what. But this amounts to treating as given facts about when they count as believing that someone believes/desires something. And taking these facts as given presupposes that the kind of vindication that we are seeking for discourse attributing beliefs and desires has already been achieved in the mental realm—that we already know when someone counts as making mental attributions of belief and desire. If our goal

is an account of the discourse that will also explicate its mental correlates, IS1 cannot be used in a pragmatist account of discourse ascribing beliefs and desires.

IS2 doesn't suffer from these difficulties, and I propose to use it as the starting point for the project of including the behaviour-predicting role of belief and desire ascriptions in their pragmatic grounds. It will help to spell this out a little. First, the notion of acting to satisfy your desires "in light of your beliefs". I propose to unpack this as acting in a way that would be conducive to the satisfaction of your desires if your beliefs were true. When we apply IS2, we predict that the agent will display behaviour that would be causally efficacious in the satisfaction of the desires we've ascribed to it if the beliefs we've ascribed to it were true.[8] If we ascribe false beliefs, a behaviour might satisfy this condition even if in actuality it isn't causally efficacious in the satisfaction of the desires we've ascribed.

In fact, though, the reliability of this kind of prediction depends on an implicit reference to the range of behaviours available to the agent. Lorna can be expected to open the fridge only if (a) this is something that she is able to do (e.g. she hasn't locked herself out of the house) and (b) even if there were water in the fridge there would be no better way of satisfying her desire available to her (e.g. she doesn't have a bottle of water in her hand). This suggests that IS2 can be reformulated as:

> *IS2* Predict that the agent will display a behaviour that would be most conducive (no less so than any other behaviour available to it) to bringing about the states of affairs that are the contents of the desires you've ascribed to it if the states of affairs that are the contents of the beliefs you've ascribed to it obtained.*[9]

The strategy can be usefully reformulated in terms of conditional probabilities. If G is the content of a desire we've attributed to the agent, S is the content of a belief we've attributed to the agent, and B is a behaviour available to the agent, let's refer to the value of p(G|B&S) as B's *degree of G-efficacy modulo S*.[10] And let's say that B is *G-optimal modulo S* just in case no other behaviour available to the agent has a higher degree of G-efficacy modulo S. Then we can reformulate IS2* in the following terms:

> *If you have ascribed to the agent a desire with content G and a belief with content S, predict that it will display a behaviour that's G-optimal modulo S.*

Notice that a behaviour prediction using this model won't follow directly from your ascription of beliefs and desires to the agent. You will need additional information about the world in order to determine which of the behaviours at the agent's disposal would be most conducive to satisfying those desires if those beliefs were true. This includes information about which behaviours are available to the agent and about the causal efficacy of these behaviours with respect to the desires you've ascribed to it in situations in which the beliefs you've ascribed to it are true (or the probability of the contents of the desires conditional on the contents of the beliefs and the behaviours). Even if we ascribe to the agent a desire for p and the belief that a certain behaviour would be most conducive to bringing about p,[11] the prediction of this behaviour will have to appeal to information to the effect that the behaviour in question is as a matter of fact available to the agent. Even if we ascribe to Lorna the desire to drink water and the belief that opening the fridge is (part of) the behaviour most conducive to her drinking water, the reliability of the prediction that she'll open the fridge rests on the assumption that she is actually right in thinking that she can make her way to the fridge (that the kitchen door is not locked, etc.).

Could we treat the involvement of our belief and desire ascriptions in the predictive strategy expressed by IS2* as part of their pragmatic grounds? This would require formulating IS2* as a condition on an arbitrary BD practice, and claiming that a BD practice produces belief and desire ascriptions only if it satisfies this condition.

For this purpose we need to think of IS2* in abstraction from the fact that the states of affairs on which the prediction is based are postulated as the contents of the agent's beliefs and desires. When considered in this way, the strategy will take the following form:

> *Predict that the agent will display a behaviour that would be most conducive to bringing about certain states of affairs X if certain states of affairs Y obtained.*

Then we can say that a BD practice will be involved in the application of this strategy when it uses the states of affairs with which it D-pairs the agent to fill the X argument place in this schema and the states of affairs with which it B-pairs the agent to fill the Y argument place. In other words, a BD practice will be involved in the application of the behaviour-predicting strategy expressed by IS2* just in case behaviour prediction follows this template:

> *Predict that the agent will display a behaviour that would be most conducive to bringing about the states of affairs with which you have D-paired it if the states of affairs with which you've B-paired it obtained.*

Or:

> *If you have D-paired the agent with G and you have B-paired it with S, predict that it will display a behaviour that's G-optimal modulo S.*

I'm going to refer to this predictive strategy as the *minimal intentional (MI) strategy*. Our goal is to find a way of treating the role of belief and desire ascriptions in the prediction of behaviour as part of their pragmatic grounds—of what gives them the character of belief and desire ascriptions. I am proposing that we might achieve this if we treat involvement in the MI strategy as a necessary condition for a BD practice to produce belief and desire ascriptions.

4 | INDETERMINACY AND CHARITY

A familiar feature of the predictive strategy described by IS2 is that any given behaviour can be predicted by innumerable belief-desire combinations. Specifically, for any behaviour and any desire, the behaviour can be predicted by ascribing the desire, so long as we are allowed to make adjustments in the beliefs we ascribe—the behaviour would be most conducive to bringing about satisfaction of the desire if the beliefs were true. Lorna's opening the fridge door can be predicted by ascribing to her the desire to drink water and the belief that there's water in the fridge, but also by ascribing to her the desire to build the Death Star and the belief that the blueprint is kept in the fridge.

This is not a problem in the presence of IS1, since in normal circumstances it is more likely that Lorna ought to desire to drink water and believe that there's water in the fridge than that she ought to desire to build the Death Star and believe that the blueprint is kept in the fridge. In general, as Dennett points out, IS1 has the consequence that we end up ascribing mostly true beliefs: "An implication of the intentional strategy, then, is that true believers mainly believe truths" (Dennett, 1987: 19). This is,

he explains, "an elaboration and further specification of the fundamental rule: attribute those beliefs the system *ought to have*" (Dennett, 1987: 19–20). However, once IS1 is discarded, as I am proposing, we are left with a behaviour-predicting strategy that shows no preference for the ascription of true belief. This is the situation with the MI strategy. A BD practice can be involved in the application of the MI strategy even if it is indifferent to the truth value of the states of affairs with which it B-pairs agents.

If the pragmatist thinks that our preference for ascribing true belief is an essential feature of our practice, she will want this preference to be a consequence of her account of the pragmatic grounds of belief and desire ascriptions. She will want to maintain that a BD practice shouldn't count as producing belief and desire ascriptions unless it exhibits this preference in some form. As we've just seen, she won't secure this result by treating involvement in the application of the MI strategy as a necessary condition for a BD practice to produce belief and desire ascriptions. She could simply stipulate that in order to produce belief and desire ascriptions, a BD practice that's involved in the application of the MI strategy must also B-pair agents with obtaining states of affairs whenever possible. However, it would be preferable if she could find a criterion that renders our preference for ascribing true beliefs intelligible. I'm going to argue that the pragmatist can achieve this by taking a closer look at the way in which our belief and desire ascriptions are involved in the prediction of behaviour.

5 | THE ONTOGENESIS OF THE INTENTIONAL STANCE

We know that humans can apply the MI strategy at least from the age of 4. This is the age at which children reliably pass classical displacement tests for false-belief understanding, and passing these tests requires applying the MI strategy. In a famous instance of this kind of test, children listen to a narration illustrated with dolls. They are told that Sally has a basket and Anne has a box. Sally puts a marble in her basket and leaves the scene, leaving the basket, with the marble, behind. In Sally's absence Anne moves the marble from the basket to the box. Then Sally returns to the scene. Where will Sally look for her marble? (Baron-Cohen, Leslie, & Frith, 1985).[12] Children under the age of 4 tend to give the wrong answer: Sally will look for her marble in the box, where the marble was placed in her absence. However, children from the age of 4 reliably give the correct answer: Sally will look for her marble in her basket. On the most natural description of how they achieve this, these children are applying the MI strategy: predicting that Sally will look in the basket because this behaviour would be most conducive to satisfying her desire to retrieve the marble if her (false) belief that the marble is in the basket were true. We know, then, that the ability to implement the MI strategy is present by age 4.

In fact, recent research strongly suggests that this ability is present at a much earlier age—probably from the second year of life. This is not manifested in elicited-response tests, like the Sally-Anne test, until the age of 4, but much younger children reliably pass spontaneous-response tests that involve the application of the MI strategy. A typical experiment of this kind, using the violation-of-expectation (VOE) method, is presented in Onishi and Baillargeon (2005). The experiment involves a scene with a green box and a yellow box. An agent hides a toy inside the green box and later, in her absence, the toy is moved from the green box to the yellow box. The agent returns and reaches either inside the yellow box or inside the green box. 15-month-olds reliably looked longer when the agent reached inside the yellow box than when she reached inside the green box, showing that they expected (predicted) that she would reach inside the green box. The infants' prediction is naturally described as following the MI template: they predict that the agent will reach inside the green box because doing so would be the behaviour most conducive to satisfying her desire to obtain the toy if her belief that the toy is in the green box were true.

I'd like to emphasize that by describing the subjects who succeed in these tasks as applying the MI strategy, I'm not providing a full account of how their behaviour predictions are produced. Subjects who use the MI strategy to predict that Sally will look for her marble in the basket will single out this behaviour as one that would be most conducive to satisfying Sally's desire if her belief were true. Hence in order to apply the strategy, subjects first need to decide what Sally desires and what she believes. The MI strategy does not dictate how to achieve this. It needs to be supplemented with procedures for deciding which beliefs and desires to ascribe, even if the ascriptions produced in this way are subsequently assessed and revised in light of their behaviour-predicting success under the MI strategy.[13]

Several developmental psychologists have suggested that prior to the stage at which children manifest their mastery of the MI strategy with their successful performance in false-belief tests, they employ a more rudimentary strategy for predicting behaviour that shares some features with the MI strategy. I'm going to focus here on the account of this early behaviour-predicting strategy developed by Gergely Csibra, György Gergely and others, who refer to it as the *teleological stance*.

In a series of VOE experiments, these researchers have demonstrated the use of this more rudimentary behaviour-predicting strategy from the age of 9–12 months. In one of their experiments, 12-month-old infants were shown on a screen a habituation sequence consisting of a small circle and a big circle separated by an obstacle—a rectangular figure, with the small circle jumping over the obstacle and approaching the big circle (Gergely, Nádasdy, Csibra, & Bíró, 1995). The subjects were then presented with two kinds of test events with the small and big circles placed as before, but with the obstacle removed. In the first, the small circle followed exactly the same trajectory as in the habituation sequence, even though the obstacle was no longer there. In the second, the small circle approached the big circle following a straight trajectory. The subjects consistently displayed longer dishabituation times (i.e. looked longer) with the first test event, in which the small circle "jumps over" a non-existent obstacle, than with the second test event, in which the small circle takes the most direct route to the big circle. This was taken to show that the infants expected the small circle to take the most direct route, and were surprised when it didn't.

It is natural to describe the situation as a case of behaviour prediction: the infants predict that the small circle will take the straight route. It is also natural to say that this prediction is reached through the attribution of a goal—the goal of approaching the big circle through the shortest route. The infants predict that when the obstacle is removed the small circle will follow the straight trajectory towards the big circle because this is the behaviour most conducive to achievement of this goal. In these respects, the strategy employed by the infants resembles the MI strategy employed by older children and adults.

However, Csibra and Gergely (1998: 256) contend that there are two major respects in which the two strategies differ from each other. They characterise our procedure as involving two separate aspects, which we could label *predictive* and *ascriptive*. The predictive aspect consists in the way in which we select the behaviour we predict as a function of certain states of affairs. We predict that the agent will display a behaviour that would be most conducive to bringing about certain states of affairs if certain other states of affairs obtained. We predict, for example, that Sally will display a behaviour that would be most conducive to getting the marble if the marble was in the basket. The ascriptive aspect consists in ascribing to the agent representational mental states with these states of affairs as their contents, playing a causal role in the production of the behaviour. In the Sally-Anne example, we ascribe to Sally mental states that represent her obtaining the marble and the marble being in the basket, and we accord to these states a causal role in the production of the predicted behaviour.

According to Csibra and Gergely, the 12-month-olds' predictive routine differs from ours in both these respects. On the one hand, their predictive strategy appears to employ a function from states of affairs to behaviour that is formally simpler than the one we employ—a unary function in place of our

binary function. We predict that the agent will display a behaviour that would be most conducive to bringing about A if B obtained. The 12-month-olds, by contrast, appear to predict that the agent will display a behaviour that will be *as a matter of fact*, i.e. by their lights, most conducive to bringing about A.[14] On the other hand, according to Csibra and Gergely, in the 12-month-olds the ascriptive aspect is simply missing. They don't ascribe to the agent a mental state representing the relevant goal as a desired state of affairs. I'll have something to say about the ascriptive aspect in the final section, but for now I want to concentrate on the predictive aspect of the situation.

Csibra and Gergely argue persuasively that the 12-month-olds really employ the simpler predictive strategy, rather than a limited version of ours, in which the second argument place is restricted to obtaining states of affairs. They refer to it as the *teleological strategy*. It can be formulated in the following terms:

> *Predict that the agent will display a behaviour that is most conducive to bringing about the goal you've ascribed.*

If we define the degree of *G-efficacy* of a behaviour B as the value of p(G|B), and say that a behaviour is *G-optimal* if no other behaviour available to the subject has a higher degree of G-efficacy, the strategy can be reformulated as follows:

> *If you've ascribed to the agent goal G, predict that it will display a G-optimal behaviour.*

Notice that subjects who fail the Sally-Anne test give the answer required by the teleological strategy. Looking in the box is the behaviour that is as a matter of fact most conducive to bringing about the goal of finding the marble.

The MI strategy is clearly more powerful than the teleological strategy. Everything that can be predicted using the teleological strategy can also be predicted by the MI strategy. A behaviour that is most conducive to bringing about A is a behaviour that would be most conducive to bringing about A if B obtained, for any obtaining B. And so long as the second argument place in the MI predictive function is restricted to obtaining states affairs, the MI strategy achieves no more than the teleological strategy. The MI strategy comes into its own when this restriction is lifted. Then the MI strategy can produce predictions for which the teleological strategy offers no match. But clearly there is a price to be paid for this increased power in terms of computational resources. We need to go from considering the actual causal efficacy of available behaviours with respect to a goal state to considering their hypothetical causal efficacy with respect to a goal state if things differed from actuality in certain respects.

These considerations motivate an intriguing hypothesis about our adult predictive practice. On this account, the teleological strategy remains our default procedure for predicting behaviour. We continue to use it whenever there is no advantage to switching to the more involved MI strategy. This is reserved for hard cases in which the teleological strategy can't produce the right predictions. In these hard cases, and only in these, we call upon the more sophisticated resources of the MI strategy.[15] And when we do, we aim to minimise the difference between the actual situation (by our lights) and the conception of the situation that we ascribe to the subject to generate our prediction. This would make sense on the assumption that the further from actuality a situation is located, the harder it would be to compute the relative causal efficacy in that situation of the available behaviours with respect to the ascribed goal. In this way, we reserve the more computationally demanding procedure for cases in which it is really needed, reverting to the more economical approach everywhere else (Csibra & Gergely, 1998: 258; Gert, 2012: 117). Let's refer to this overall approach to behaviour prediction, using the teleological strategy as a default and deploying the MI strategy only in the hard cases, as the *hybrid policy*.[16]

I've motivated my hypothesis that we follow the hybrid policy in our adult behaviour predictions using the developmental story presented by Csibra and Gergely. However the developmental story can only lend fairly weak support to the hypothesis. Even if things develop exactly as Csibra and Gergely claim before the MI strategy makes its appearance, once it is in place it might entirely displace the teleological strategy. In addition, the existence of a period before the MI strategy appears, in which only a more rudimentary predictive strategy is available, is coming under increasing pressure as new empirical results continue to bring forward the age by which infants can employ the MI strategy.[17] And even if a pre-MI strategy existed, it might not have the features that Csibra and Gergely ascribe to it.[18]

However, we have reason to believe that the teleological strategy continues to play some role in our behaviour prediction after the MI strategy has made its appearance. It seems increasingly likely that we acquire the ability to apply the MI strategy in the second year of life at the latest, and yet the performance of 3-year-olds in elicited-response false-belief tests seems to suggest that they are still inclined to predict behaviour along the lines of the teleological strategy. This inclination might well be due to the processing difficulties involved in applying the MI strategy, but what matters for our purposes is that when the MI strategy can't be applied, what children offer instead is the performance that you would expect from subjects who are applying the teleological strategy.[19]

Does this inclination disappear by the age of 4? There's some empirical evidence suggesting that the inclination to default to the teleological strategy survives into adulthood. I think this is a natural way to interpret the phenomena studied in the literature on the so-called curse of knowledge.[20] *The curse of knowledge* is the label used by Susan Birch and Paul Bloom to refer to "a tendency to be biased by one's own knowledge when attempting to appreciate a more naive or uninformed perspective" (Birch & Bloom, 2004: 256).[21] Several studies show a tendency in adults as well as children to overestimate the knowledge of others in a variety of contexts.[22]

As Birch and Bloom point out, young children's difficulties with traditional false-belief tasks can be described as manifestations of the curse-of-knowledge bias. Children who say that Sally will look for the marble in the box are using their own knowledge to predict the actions of uninformed Sally. I have followed Csibra and Gergely in describing these cases as applications of the teleological strategy, and in general we can expect applications of the teleological strategy to produce instances of the curse-of-knowledge bias. Hence if as adults we continue to use the teleological strategy as our default procedure for predicting behaviour, as I'm suggesting, we should expect the persistence of the bias into adulthood, and this is precisely what we find. In fact, with ingenious modifications of the displacement-task template, Birch and Bloom have shown an underlying tendency in adults to give the wrong answer in false-belief tests (Birch & Bloom, 2007). Their results are open to interpretations under which they wouldn't support my case, but so long as these authors are right in seeing their results as manifestations of the course-of-knowledge bias, they can be taken to lend some support to the hypothesis that adults follow the hybrid policy for predicting behaviour, using the teleological strategy as a default and shifting to the MI strategy only in cases that the teleological strategy can't handle well.

6 | THE HYBRID POLICY AS PRAGMATIC GROUNDS

In Section 3 I suggested that the pragmatist could try to treat involvement in the application of the MI strategy as part of the pragmatic grounds of belief and desire ascriptions. In Section 4 I argued that this wouldn't result in treating our preference for the ascription of true beliefs as an essential feature of our practice of ascribing beliefs and desires. In Section 5 I have presented some empirical evidence for

the claim that our behaviour predictions based on the ascription of beliefs and desires don't follow in general the MI strategy, but the hybrid policy, in which the use of the MI strategy is restricted to cases in which the teleological default doesn't generate the right predictions. What I want to argue now is that the pragmatist could succeed in treating our preference for ascribing true beliefs as essential to the practice by including in its pragmatic grounds involvement in the application, not of the MI strategy, but of the hybrid policy.

On this proposal, a BD practice would generate belief and desire ascriptions only if it was involved in the application of the hybrid policy. To see what form this involvement would take, consider how an interpreter would proceed if she employed the hybrid policy to predict an agent's behaviour. She would first try to identify a state of affairs G such that the behaviour that the agent goes on to produce can be described as G-optimal (most conducive to bringing about G). If this didn't work, and only then, she would try to identify a state of affairs G and a non-obtaining state of affairs S such that the behaviour that the agent goes on to produce can be described as G-optimal modulo S (most conducive to bringing about G if S obtained). Notice that S would have to be non-obtaining (by the interpreter's lights) for the manoeuvre to make a difference, since, if S obtained, a behaviour would be G-optimal modulo S just in case it is G-optimal.

Now, in order for a BD practice to be involved in the application of the hybrid policy, its practitioners need to predict the behaviour of agents along the lines of the hybrid policy, and use "D" and "B" to pair agents with the states of affairs on which their interpretative hypotheses are based: they'll D-pair an agent with a state of affairs G when they predict that the agent will display G-optimal behaviour, either in absolute terms or modulo some non-obtaining state of affairs, and they'll B-pair an agent with a (non-obtaining) state of affairs S when they predict that the agent will display H-optimal behaviour modulo S, for some state of affairs H.

If the use of "B" and "D" is regulated in this way, agents will be B-paired with non-obtaining states of affairs only in exceptional circumstances. Lorna's trip to the fridge can be predicted using the teleological strategy, as a behaviour as a matter of fact most conducive to bringing about the goal of drinking some water. Hence Lorna will be D-paired with the state of affairs consisting in her drinking water by default, with no need to B-pair her with a non-obtaining state of affairs. To be sure, her behaviour could also be predicted using the MI strategy by D-pairing her with the state of affairs of her building the Death Star, which would require B-pairing her with the non-obtaining state of affairs of the fridge containing the blueprint, but this prediction is trumped, under the hybrid policy, by the successful teleological default.

If we treated the involvement of a BD practice in the application of the hybrid policy as a necessary condition for its B- and D-pairings to count as belief and desire ascriptions, you would count as ascribing beliefs and desires only if your ascriptions were regulated in this way. And, as we've just seen, it follows from this that false belief would be ascribed only in exceptional circumstances—when the agent's behaviour can't be successfully predicted without taking this step. If regulation according to the hybrid policy were an essential feature of belief and desire ascription, our preference for ascribing true belief (or rather, our reluctance to ascribe false belief) would be rendered intelligible.

Notice that it's a feature of this approach that the regulation of belief and desire ascriptions by the hybrid policy never calls for the ascription of true belief. This suggests a picture of interpretation that focuses on the ways that the agent's conception of the situation differs from the interpreter's. Interpretation, on this approach, starts from the background assumption of a shared conception of the environment between interpreter and agent—or, from the interpreter's point of view, the assumption that the agent sees things as they really are. This background assumption is overridden only when doing so leads to better predictions of the agent's behaviour.[23]

7 | OTHER CONDITIONS

My main claim has been that the pragmatist can include the behaviour-predicting role of belief and desire ascriptions in their pragmatic grounds by treating involvement in the application of the hybrid policy as a necessary condition for a BD practice to count as producing belief and desire ascriptions. I have also argued that this approach succeeds in treating our preference for the ascription of true belief as an essential feature of the concept. In this section I want to review briefly other conditions that the pragmatist might want to include in the pragmatic grounds of belief and desire ascriptions.

We can think of the hybrid policy as a procedure for *indexing* subjects with states of affairs, and the leading thought of the approach I'm exploring is that this indexing is what the ascription of beliefs and desires fundamentally consists in. The thought that the ascription of propositional attitudes has this indexing character is familiar from the literature on the measurement analogy (Churchland, 1979: 105; Stalnaker, 1984: 8–11; Matthews, 2007), and the position I'm exploring should be seen as a (pragmatist) version of this approach, in which agents are indexed with the states of affairs on which interpreters base their predictions of the agents' behaviour according to the hybrid policy.

However, there are features of our practice that we might regard as essential to it but are not covered by the link with our predictive strategy. One is the thought that beliefs and desires cause the behaviour they help to predict. One might hold that a BD practice that doesn't acknowledge this causal dimension of its indexings of subjects with states of affairs should not count as ascribing beliefs and desires. I want to suggest that the pragmatist might be able to accommodate this feature of the concepts of belief and desire with the demand that in order to count as belief and desire ascriptions, B- and D-pairings should be treated as incorporating a difference-making claim (Woodward, 2011).[24] Suppose that you predict that the agent will display the behaviour that would be most conducive to bringing about q if p obtained, but you think that the agent would display this behaviour even if it wasn't the behaviour that would be most conducive to bringing about q if p obtained. We might want to say, as a first stab, that in order for your B- and D-pairings to count as belief and desire ascriptions, in these circumstances you shouldn't B-pair the agent with p and D-pair it with q. If the causal dimension of belief and desire ascription could be incorporated into our account of the pragmatic grounds of the discourse in this or some other way, the pragmatist would have the resources for accommodating the thought that belief and desire ascriptions can be used to explain, as well as predict, an agent's behaviour, without coming into conflict with the idea that intentional explanation is a species of causal explanation (Davidson, 1963).

Another aspect of our conception of the practice that the pragmatist might want to treat as essential is the thought that beliefs and desires represent things as being a certain way. One way in which one might try to construe this intuition is in terms of counterfactual connections between belief, desire and behaviour. The thought is that believing p and desiring q requires not only displaying the behaviour that would be most conducive to bringing about q if p obtained. In addition, it has to be the case that for any state of affairs x, if you desired x you would display the behaviour that would be most conducive to bringing about x if p obtained, and for any state of affairs y, if you believed y, you would display the behaviour that would be most conducive to bringing about q if y obtained. One might argue that this thought grasps the representational character of belief and desire.[25] Then the pragmatist might extract from it a necessary condition for a BD practice to produce belief and desire ascriptions:

> *If you B-pair the agent with p, then for every state of affairs x, if you D-paired the agent with x, you would predict that it would display a behaviour that would be most conducive to bringing about x if p obtained.*

> *If you D-pair the agent with q, then for every state of affairs x, if you B-paired the agent with x, you would predict that it would display a behaviour that would be most conducive to bringing about q if x obtained.*

Another aspect of our conception of the practice that one might expect the pragmatist to consider for possible inclusion in its pragmatic grounds is the procedure that we follow for selecting our hypotheses as to which beliefs and desires to ascribe. On the picture I have presented, belief and desire ascription result from a process in which ascribers first select a state of affairs A and predict that the subject will display behaviour that is most conducive to bringing about A or, if this doesn't produce the right results, they select states of affairs A and B and predict that the subject will display behaviour that would be most conducive to bringing about A if B obtained. As I mentioned in Section 5, this characterisation of the process says nothing about how the ascriber is supposed to select hypotheses as to which states of affairs to try for the role of A in the teleological template or A and B in the intentional template.[26] Notice that this silence doesn't imply that the ascriber won't have to employ some procedure for achieving this. All it means is that which procedure is used is not treated as essential to the practice—that different ascribers employing different heuristics to select their B- and D-pairings can all count as ascribing beliefs and desires, so long as the pragmatic grounds of the discourse are satisfied. I believe this is the right result. Different ascribers might employ different ascription heuristics in different circumstances, and this diversity should not in principle affect the status of their B- and D-pairings as belief and desire ascriptions. If this is right, then it follows that the pragmatist should refrain from building a commitment to a specific account of the character of mindreading into the pragmatic grounds of belief and desire ascriptions.[27] The pragmatic grounds of the discourse should not dictate how we ought to read minds. The pragmatist should only be claiming that however we achieve this, our 'readings' acquire the status of belief and desire ascriptions only as a result of their employment in the prediction of behaviour in accordance with the hybrid policy.

8 | CONCLUSION

I have explored the idea of including the role that belief and desire ascriptions play in the prediction of behaviour in the pragmatic grounds of the discourse. I have made a case for the claim that our behaviour prediction based on belief and desire ascriptions follows what I've called the hybrid policy, treating the teleological strategy as a default and switching to the (minimal) intentional strategy only in cases that the teleological strategy can't handle well. I have proposed including the behaviour-predicting role of belief and desire ascriptions in their pragmatic grounds by treating involvement in the application of the hybrid policy as a necessary condition for a practice to count as producing belief and desire ascriptions. I have argued that this approach succeeds in treating our preference for our ascription of true belief as an essential feature of the concept. I have briefly reviewed some other factors that one might want to include in the pragmatic grounds of belief and desire ascriptions.

ENDNOTES

[1] Huw Price's *placement problems* fall in this category (Price, 2004; 2013: 6–8).

[2] "For a *large* class of cases—though not for all—in which we employ the word 'meaning' it can be explained thus: the meaning of a word is its use in the language" (Wittgenstein 1953/2001: §43, translation slightly amended).

[3] I'm borrowing the term from Huw Price.

[4] I've applied this approach to truth discourse in Zalabardo (2016).

[5] Possible states of affairs might be too coarse-grained to serve as relata in propositional-attitude ascriptions, in light of the intensional character of the notion. I'm not going to consider this complication here.

[6] Other philosophers have exploited the link with behaviour in their account of propositional attitudes, without emphasizing the predictive aspect. See Ramsey (1927), Stalnaker (1984), and Whyte (1990).

[7] As far as I can see, the ascription of rationality to the agent plays no role in the strategy beyond the assertion that the agent will act to further its goals in the light of its beliefs.

[8] I prefer to speak of behaviour, rather than action, to emphasize the continuity between the intentional stance and other strategies we might adopt for predicting the behaviour of a complex system. Very crudely, behaviour becomes action as a result of its link with the beliefs and desires with which we connect it in our applications of the intentional stance.

[9] In real-life behaviour prediction, the selection of the behaviour we predict is likely to involve an element of cost-benefit analysis. I won't consider here this complication.

[10] This magnitude cannot be used to compare the efficacy of behaviour with respect to different goals, as it hasn't been calibrated to the probability of the goals.

[11] This is not the typical scenario. Normally we base our behaviour prediction on the ascription of beliefs about features of the subject's environment, not about the causal efficacy of possible behaviours.

[12] The experimental model was first introduced by Wimmer and Perner (1983). For a review see Wellman, Cross, and Watson (2001).

[13] See the final paragraph of Section 7, below.

[14] Many researchers have endorsed the idea that motivational concepts are mastered before informational concepts, phylogenetically as well as ontogenetically. Thus in their seminal study on mindreading in chimpanzees, David Premack and Guy Woodruff write: "Of all possible guesses, we find the most compelling one to be that inferences about motivation will precede those about knowledge, both across species and across developmental stages. Not even the chimpanzee will fail tests that require him to impute *wants, purposes*, or *affective attitudes* to another individual, but he may fail when required to impute states of knowledge" (Premack & Woodruff, 1978: 526). Henry Wellman has defended a similar view specifically with respect to human infants: "I suggest that children younger than three, say two-year-olds, fail to understand belief-desire psychology. They utilize instead, a simple desire psychology" (Wellman, 1991: 19).

[15] I think that this two-procedure hypothesis is in principle compatible with one-system accounts of adult mindreading (Carruthers, 2017). On this picture the same system that produces intentional predictions would default to the pared-down teleological operation when the full resources of the system are not required.

[16] Ian Apperly and Stephen Butterfill have defended "the existence of two types of system for belief reasoning: one that is cognitively efficient but limited and inflexible and another that is flexible but demanding of general cognitive resources" (Apperly & Butterfill, 2009: 956). However, they don't think that the teleological strategy offers a plausible way of achieving cognitive efficiency (Apperly & Butterfill, 2009: 961).

[17] See Scott and Baillargeon (2017) for a review of recent results.

[18] On one alternative (Baillargeon, Scott, & He, 2010; Onishi & Baillargeon, 2005), infant behaviour prediction is based, not on the subject's own conception of the situation, but on *true* beliefs ascribed to the agent, which might not include all of the subject's beliefs. Renée Baillargeon (personal communication) has now abandoned the view that the emergence of the MI strategy is preceded by a period in which infants use a more rudimentary strategy for behaviour prediction. On another alternative, infant behaviour prediction is based on the subject's conception of the situation as well as her conception of which goals should be pursued (Perner & Roessler, 2010; Roessler & Perner, 2013). Perner and Roessler use for this view the label *teleological account*, and refer to Csibra and Gergely's picture as the *hybrid account*. As they explain (Perner & Roessler, 2010: 227), their approach also differs from Csibra and Gergely's in that they are more concerned with the explanation of actions that with the prediction of behaviour.

[19] Perner and Roessler have emphasized the relevance in this connection of the fact that the performance of 3-year-olds in these tasks, although poor, is far from random (Perner & Roessler, 2010: 200; Roessler & Perner, 2013: 44).

[20] Thanks to Alex Jackson for alerting me to the relevance of this phenomenon.

[21] They borrow the label from Camerer, Loewenstein, and Weber (1989).

[22] See Birch and Bloom (2004) and Ghrear, Birch, and Bernstein (2016) for references. For related results, see Mitchell, Robinson, Isaacs, and Nye (1996) and Epley, Keysar, Van Boven, and Gilovich (2004).

[23] We have considered situations in which we ascribe belief in states of affairs that we believe not to obtain. We also sometimes ascribe belief in states of affairs that we don't believe to obtain and we don't believe not to obtain. In addition, we sometimes ascribe ignorance on matters on which we hold beliefs. In a longer presentation of this material I refine the characterisation of the hybrid policy to make room for these possibilities.

[24] Perner and Roessler have highlighted the potential of difference-making conceptions of causation in this context, although their own concern is to invoke them to vindicate the causal efficacy of what they call *objective reasons* (Perner & Roessler, 2010: 208–09; Roessler & Perner, 2013: 36–7).

[25] I think this is congenial to the approach in Ramsey (1927).

[26] See Csibra, Bíró, Koós, and Gergely (2003: 129) and Csibra and Gergely (2007).

[27] See Carruthers and Smith (1996) for an overview of the debate.

REFERENCES

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, *116*, 953–970.

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

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a 'theory of mind'? *Cognition*, *21*, 37–46.

Birch, S. A. J., & Bloom, P. (2004). Understanding children's and adults' limitations in mental state reasoning. *Trends in Cognitive Sciences*, *8*, 255–260.

Birch, S. A. J., & Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science*, *18*, 382–386.

Brandom, R. (2013). Global anti-representationalism? In H. Price (Ed.), *Expressivism, pragmatism and representationalism*. Cambridge: Cambridge University Press.

Camerer, C., Loewenstein, G., & Weber, M. (1989). The curse of knowledge in economic settings: An experimental analysis. *Journal of Political Economy*, *97*, 1232–1254.

Carruthers, P. (2017). Mindreading in adults: Evaluating two-systems views. *Synthese*, *194*, 673–688.

Carruthers, P., & Smith, P. K. (Eds). (1996). *Theories of theories of mind*. Cambridge: Cambridge University Press.

Churchland, P. M. (1979). *Scientific realism and the plasticity of mind*. 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*, 111–133.

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

Csibra, G., & Gergely, G. (2007). 'Obsessed with goals': Functions and mechanisms of teleological interpretation of actions in humans. *Acta Psychologica*, *124*, 60–78.

Davidson, D. (1963). Actions, reasons, and causes. *The Journal of Philosophy*, *60*, 685–700.

Dennett, D. C. (1987). True believers. In *The intentional stance*. Cambridge: MIT Press.

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology*, *87*, 327–339.

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

Gert, J. (2012). *Normative bedrock: Response-dependence, rationality, and reasons*. Oxford: Oxford University Press.

Ghrear, S. E., Birch, S. A. J., & Bernstein, D. M. (2016). Outcome knowledge and false belief. *Frontiers in Psychology*, *7*

Matthews, R. J. (2007). *The measure of mind: Propositional attitudes and their attribution*. Oxford: Oxford University Press.

Mitchell, P. L., Robinson, E. J., Isaacs, J. E., & Nye, R. M. (1996). Contamination in reasoning about false belief: An instance of realist bias in adults but not children. *Cognition*, *59*, 1–21.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*, 255–258.

Perner, J., & Roessler, J. (2010). Teleology and causal reasoning in children's theory of mind. In J. Aguilar & A. Buckareff (Eds.), *Causing human action: New perspectives on the causal theory of action*. Cambridge: MIT Press.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*, 515–526.

Price, H. (2004). Naturalism without representationalism. In M. De Caro & D. Macarthur (Eds.), *Naturalism in question*. Cambridge: Harvard University Press.

Price, H. (2013). *Expressivism, pragmatism and representationalism*. Cambridge: Cambridge University Press.

Ramsey, F. (1927). Facts and propositions. *Proceedings of the Aristotelian Society*, *7*, 153–170.

Roessler, J., & Perner, J. (2013). Teleology: Belief as perspective. In S. Baron-Cohen, H. Tager-Flusberg, & M. V. Lombardo (Eds.), *Understanding other minds: Perspectives from developmental social neuroscience* (3rd ed). New York: Oxford University Press.

Scott, R. M., & Baillargeon, R. (2017). Early false-belief understanding. *Trends in Cognitive Sciences*, *21*, 237–249.

Stalnaker, R. C. (1984). *Inquiry*. Cambridge: MIT Press.

Wellman, H. M. (1991). From desires to beliefs: Acquisition of a theory of mind. In A. Whiten (Ed.), *Natural theories of mind: Evolution, development and simulation of everyday mindreading*. Oxford: Basil Blackwell.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, *72*, 655–684.

Whyte, J. T. (1990). Success semantics. *Analysis*, *50*, 149–157.

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

Wittgenstein, L. (2001). *Philosophical investigations* (3 ed.). G. E. M. Anscombe, Trans.) Oxford: Blackwell. (Original work published 1953)

Woodward, J. (2011). Causal perception and causal cognition. In J. Roessler, H. Lerman, & N. Eilan (Eds.), *Perception, causation, and objectivity*. Oxford: Oxford University Press.

Zalabardo, J. L. (2016). Empiricist pragmatism. *Philosophical Issues*, *26*, 441–461.

How to cite this article: Zalabardo JL. Belief, desire and the prediction of behaviour. *Philosophical Issues*. 2019;1–16. https://doi.org/10.1111/phis.12155