Jago, Mark (2006) Belief and Bounded Rationality.

Belief and Bounded Rationality

Mark Jago

Draft: please don't circulate or cite.

Abstract

Predictive accounts of belief ascription, either following the principle of charity or Dennett's intentional stance, have proved popular recently. However, such accounts require us first to treat agents as perfectly rational agents and then revise this assumption as appropriate. I argue that such downwards revision is no easy task and that several proposed accounts are not satisfactory. I propose a way of characterising agent's belief states which shares Dennett's approach but avoids treating agents as perfectly rational, and develop a formal account in terms of *fan models*.

1 Introduction

Since Quine's *Word and Object* [Qui60], there has been more-or-less general agreement on the correct treatment of intentional attitudes. In a strict ontological sense, "the canonical scheme for us is the austere scheme" according to which there are "no propositional attitudes but only the physical constitution and behaviour of organisms" [Qui60, p. 221]. However, intentional idioms are "practically indispensable" [Qui60, p. 219].[1] There are, of course, disagreements within this general viewpoint. Dennett [Den87, pp. 342–343] divides the resulting accounts into those based on a *normative principle*, according to which we ascribe the attitudes an agent *ought* to have, given its circumstances, and those based on a *projective principle*, whereby one ascribes those attitudes that one would have oneself in those circumstances. In this paper, I want to consider the former group of accounts, which includes those based around Davidson's *principle of charity* [Dav85] and Dennett's own *intentional stance* [Den87]. In particular, I want to argue that Dennett's intentional stance has great difficulty in dealing with agents with bounded rationality—you, me, and everyone else.

[1] See also Sellars [Sel56].

I will assume, without much argument, that Dennett's motivation is more or less correct, but argue that the method he gives us for ascribing beliefs and desires to others cannot avoid attributing too many beliefs. I will then suggest another method, which shares Dennett's outlook but avoids this problem.

2 The Predictive Strategy

is intended as a way of bridging the gap between realist and interpretational accounts of intentional attitude attribution (or rather, of claiming that this is a deeply unhelpful dichotomy). Dennett holds that, "while belief is a perfectly objective phenomenon ... it can be discerned only from the point of view of one who adopts a certain *predictive strategy*, and its existence can be confirmed only by an assessment of the success of that strategy" [Den87, p. 15]. Here, Dennett is in agreement with Quine in that determining the truth of belief attributions could not be reduced to the existence some underlying physical phenomena:

> It will often happen also that there is just no saying whether to count an affirmation of a propositional attitude as true or false, even given full knowledge of its circumstances and purposes [Qui60, p. 218].

Dennett describes his approach as follows:

> first you decide to treat the object whose behavior is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs. A little practical reasoning from the chosen set of beliefs and desires will in most instances yield a decision about what the agent ought to do; that is what you predict the agent will do. [Den87, p. 17]

Let us call this method of ascribing beliefs the *predictive strategy*. A first objection is that it becomes hard to explain false belief. If we follow Dennett's claim that we should "attribute as beliefs all the truths relevant to the system's interests (or desires) that the system's experience to date has made available" [Den87, p. 18], then how can we explain where false beliefs come from?

Dennett's response is that "[t]he falsehood has to start somewhere" [Den87, p. 19]. Stich [Sti81] objects that there remain situations that cannot be explained in this way. One of his examples considers a newspaper vendor who on an occasion gives the wrong change. But what, as Dennett replies, is to be explained here? If this was a genuine mistake, perhaps caused by a temporary brain malfunction or miscalculation, then we would not expect to rationalise the mistake in terms of beliefs and desires [Den87, pp. 83–88]. It was an irrational mistake and so we should not search for rational reasons.

Dennett's response is acceptable in the case of an irrational error, such as the vendor's mistake. The vendor would more than likely be unable to explain why he made the mistake himself. However, the situation changes when we consider mistakes that happen because of bounded resources. Suppose a chess player could win a game by making a particular series of moves but, because he has limited time in which to think and can only think a certain number of moves ahead, does not make these moves and ends up losing. In a sense, he has made a mistake because he did not make the most rational moves (assuming, of course, that he wanted to win). The problem here is that the agent's experience—his knowledge of the rules of chess and the positions of the pieces on the board—makes available to him information about the winning strategy.[2] But we should not be tempted to say that the agent had no reason for acting as he did, i.e. with less than ideal rationality. If we pointed out the winning strategy to the agent after the game, he might claim that he could have discovered it himself, if only he had more time, or an ability to look more moves ahead. So there are reasons we can cite to explain cases of agents acting with less than perfect rationality. These reasons are not captured by Dennett's predictive strategy, which would predict that the agent chooses the winning strategy every time.

A related problem is that, in treating agents as perfect reasoners, it is difficult to distinguish those beliefs an agent has, from those it might come to believe through further reasoning. Dennett's partial response is to distinguish opinion, which is a classical on-off affair, from belief which may be a matter of degree, governed by Baysian rules [Den81, chapter 16]. Opinion is a matter of assent but nevertheless belief provides the basis for an agent's opinion. We should at the least be able to say, given that agent *a* has *these* beliefs, that it should be able to assent to this or that (or come to *this* opinion). Dennett agrees with de Sousa [dS71] that a Baysian-style

[2] We might be tempted to say that the information about the winning strategy was not *available* to the agent at all, since his bounded resources did not permit him to access the information in a useful way. I discuss a logic of information in [Jag06a].

theory of belief should be used “to explain (or at least predict statistically) the acts of assent we will make given our animal-level beliefs and desires.” Such beliefs “explain our proclivity to make these leaps of assent, to act, to bet on the truth of various sentences.” [Den81, p. 304].

The problem here is that one can only give one’s assent based on one’s beliefs if one can see how what one is assenting to is one of, or is supported by, one’s beliefs; otherwise, we are simply discussing an agent who guesses all the time. Let us consider an example. Any student with a very minimal set of beliefs (that one should endeavour to answer the questions; one should give the answers one believes to be true ...) would be predicted to do very well in his first-year logic exam. Similarly, a mathematics student who knows the axioms of some theory would be said to believe (and also know) all theorems of the theory—however complicated they may be. In fact, few students achieve 100% on their logic test and no one knows or believes all the theorems of arithmetic, say, let alone all the relevant meta-theorems. This is why logical and mathematical discoveries are surprising and informative. My bias in this discussion is therefore motivated by the following principle: we should not ascribe beliefs to agents that they could not, given their cognitive limitations, assent to. In following this principle, we also remove the temptation to think that the chess playing agent will always take the winning strategy.

This motivates the following question: what notion *do* we capture in treating agents as ideal reasoners? Certainly not belief (at least, as *we* use the term), for real agents are far from ideally rational when it comes to managing their own beliefs. But the assumption of perfect rationality nevertheless has a place, in showing what an agent’s rational commitments are in having certain beliefs and opinions. In judging the world to be a certain way, for instance, an agent commits itself to the consequences of that judgement. If an agent judges $\phi_1, \ldots, \phi_n$ to be the case and ψ is a consequence of these judgements but is rejected by the agent, then we could point out some error in the agent’s reasoning. In showing the agent that ψ is a consequence of judgements she has made, we would expect her to either change her mind about ψ or else reject one of the original judgements.

In talking about the consequences of an agent’s judgements, we may want to restrict the notion to *relevant* consequences, perhaps by taking relevant implication as our model. In this way, we can rule out the strange commitments involving material implications, such as one’s judgements about what to have for tea committing one to $p \rightarrow q \vee q \rightarrow r$, for any (completely unrelated) propositions p, q, r. In a similar way, the notion of commitment should avoid the *ex contradictione quad libet* principle, or principle of explosion, whereby contradictory judgements would commit an agent to every proposition whatsoever. An acceptable, non-explosive notion of consequence must therefore tolerate a degree of contradiction, as paraconsistent logics do. So, the notion of commitment, given what an agent judges, should be characterised along the lines of a paraconsistent, relevant consequence relation.

It is clear that this notion of commitment is too strong for an analysis of belief. An agent need not believe all of the things it commits itself to in making judgements; it could only do so if it were an ideal agent, with perfect rationality and unlimited cognitive capacity (memory, time to reason and so on). So, the commitments one forms in making judgements form an upper limit on what that agent believes. Moreover, the judgements an agent makes (the opinions it forms, the sentences it accepts or assents to) form a lower bound on what the agent believes. If an agent judges that ϕ then it believes that ϕ and it believes ϕ only if it is thereby committed to the truth of ϕ.

3 Avoiding Idealised Ascriptions

In the previous section, the consequences of an agent's beliefs were termed the commitments of those beliefs. The question that needs to be addressed now is: how can Dennett's predictive strategy result in a notion of belief which differs from (is weaker than) that of commitment? Dennett's suggestion is as follows.

> One starts with the idea of perfect rationality and revises downwards as circumstances dictate. That is, one starts with the assumption that people believe all the implications of their beliefs and believe no contradictory pairs of beliefs. ... one is interested only in ensuring that the system is rational enough to get to the particular implications that are relevant to its behavioural predicament of the moment. [Den87, p. 21]

Let us call this the *downwards revision* approach. Now, one might quite legitimately ask: just what is the measure of rationality appealed to here supposed to consist in? and just how does one revise downwards? I now take a look at two possible suggestions which attempt to explain downwards revision. Since a fully fleshed-out predictive strategy would incorporate a formal model of belief—a Baysian model, for example—an approach to downward revision should also be based on a more-or-less formal approach. Otherwise, we will not have a *method* at all; rather, we will be left with an *ad hoc* way of pruning beliefs.

A first suggestion is found in Hintikka's notion of *logical competence* [Hin75]. To be sure, one's logical competence does not exhaust one's rational ability but it is a component of it. If we cannot provide a method of downward revision to the way we ascribe logical competence to an agent, we cannot give a method of downwards revision for the way we ascribe rationality to that agent in general. Hintikka describes logical models which are inconsistent from a classical point of view, but "so subtly inconsistent that the inconsistency could not be expected to be known (perceived) by an everyday logician, however competent." [Hin75, p. 478] Suppose an agent considers the sentences satisfied by such a model to state genuine possibilities. That agent will thereby be taking some impossibilities to be possible and, in doing so, will not consider all valid sentences (or all consequences of its beliefs) to be true. We therefore have some handle on her logical competence, depending on the degree to which contradictions in the model manifest themselves.

The details of such models are provided by Rantala in [Ran75], where he uses the term *urn models*. I omit the details here (see [Jag06b, chapter 2] for a detailed discussion). The problem with such models is that agents remain believers in all instances of propositional tautologies. An agent's variable-free beliefs will be deductively closed and we will not be able to subtract from our initial assumption of perfect rationality in this domain. Moreover, we have no reason to suppose that an agent's competence will be a fixed parameter across the board. There are numerous sentences which the agent *could* derive, given her assumed degree of rationality and which she will therefore be ascribed belief in on the predictive strategy, which she will in fact not believe in the slightest. A mathematician who has spent months working towards proving a particular theorem is likely to have beliefs in that domain of far greater justificatory complexity than in other domains, or even in other mathematical fields. A logician might even be able to prove a complex theorem but have trouble with, what from the viewpoint of quantifier depth alone, appears to be less complex, such as deriving a corollary. This could not be explained using Hintikka's notion of logical competence.[3]

As an alternative approach, we might adapt Fagin and Halpern's account in [FH88] and begin by ascribing an idealised theory of belief to an agent, but then filter the results through an 'awareness' filter. Awareness is a purely syntactic notion and so it is possible to alter the properties of awareness without modifying the underlying idealised account of belief. We

[3] Similar examples are discussed in [Jag06b, ch. 1].

need not specify properties of the awareness set *a priori*, but "[o]nce we have a concrete interpretation in mind, we may want to add some restrictions" [FH88, p. 54]. However, it seems essential to the success of the awareness model that, in general, awareness sets have no closure properties whatsoever. As Fagin and Halpern comment,

> people do *not* necessarily identify formulas such as $\psi \wedge \phi$ and $\phi \wedge \psi$. Order of presentation does seem to matter. And a computer program that can determine whether $\phi \wedge \psi$ follows from some initial premises in time τ might not be able to determine whether $\psi \wedge \phi$ follows from those premises in time τ. [FH88, p. 53, their emphasis]

Given a concrete formulation of awareness we may ask, why could this notion not be used to define a notion of belief *directly*, using whatever principles were used to determine the properties of the awareness set? A potential notion of awareness given in [FH88, 54] is that the elements of the awareness set are precisely those sentences that the agent *could* determine as consequences of information they already possess in a specified space and/or time bound. This is, roughly, the notion of belief I will propose below, although I will do so directly, making no use of the evidently spurious notion of awareness.

A further suggestion as to how we might scale down our attributions of rationality from the ideal case is as follows. The beliefs we ascribe on the back of the predictive strategy are not ascribed piecemeal, but as part of a holistic network. Certain beliefs support certain others such that, in the case of a perfectly rational agent, believing the supports is sufficient for believing the supported beliefs. So we have a justification network: a network of beliefs with justifications marked within it. Such structures are common in current AI practise.[4] Some beliefs might be taken as supplied directly by experience and so have no support within the network of beliefs itself. These include the mundane, everyday beliefs which we are too busy to ever explicitly consider or judge, such as the belief that there is a chair in front of me, but no dancing elephants.

Suppose we mark such beliefs as being experientially (as opposed inferentially) justified and then calculate the justificatory complexity of the other beliefs based on the shortest path in the justification network from the belief in question to a set of beliefs which supports it. We could then revise downwards by throwing out those beliefs of higher justificatory complexity first. Our measure of rationality would be the agent's ability to reason to

[4] They are used extensively in the areas of belief revision and belief update, for example.

beliefs of certain justificatory complexity from a given support set. However, in typical cases, there need not be a uniform degree of justificatory complexity throughout the beliefs at the periphery of an agent's justification network. There are numerous sentences which an agent *could* derive, given her assumed degree of rationality, which she will in fact not believe in the slightest. A mathematician who has spent months working towards proving a particular theorem is likely to have beliefs in that domain of far greater justificatory complexity than in other domains, or even in other mathematical fields.

Moreover, beliefs may be justified in more than one way. For example, we cannot tell whether the set $\{\phi, \psi, \phi \rightarrow \psi\}$ was obtained by *modus ponens* from $\{\phi, \phi \rightarrow \psi\}$ or from $\{\phi, \psi\}$ by disjunction introduction and the rewrite rule for '$\rightarrow$' in terms of '$\vee$'. The problem is that, in starting from the viewpoint of perfect rationality, we cannot always decide *which* justification network to associate with a given set of beliefs. We might know that we need to treat our agent as believing ϕ, ψ and $\phi \rightarrow \psi$ in order to explain its behaviour, for example, but could not infer how these beliefs arose. We could not then say what degree of rationality we were thereby attributing to the agent and so could not say just how far we need to revise our initial assumptions of perfect rationality. If we were to precede in the other direction, from a small set of beliefs and build a justification network from bottom up, we would not have this problem. Then we would not need the initial assumption of perfect rationality; we would only need to assume reasoning ability to a certain level. This is roughly the model that I will present in section 5 below.

4 Sentential Accounts

Above, I made a distinction between belief on the one hand and opinion or assent on the other. We might also include *judgements* in the latter category whose members, according to Dennett [Den81, Chapter 16] and de Sousa [dS71], are not to be treated on a par with belief. One way to make the distinction, following Malcolm [Mal72], would be to claim that, whilst it certainly seems appropriate to say that a chicken believes (or thinks) that going to the farmer is a way of getting fed, it certainly has not judged or formed the opinion that this is so; nor has it assented to that statement. Forming judgements and opinions and assenting to statements are conscious mental acts, whereas having beliefs might be viewed as a different class of mental phenomenon altogether, operating on a more fundamental, subpersonal level. This is why it makes sense to attribute beliefs to an agent that it has not explicitly considered.

However, this does not licence the claim that, whilst judgement, opinion and assent are to be cashed out in terms of statements—i.e. unambiguous sentences—beliefs are to be ascribed in terms of non-linguistic entities. For example, Dennett (following Stalnaker [Sta76]) claims that "a particular belief is a function taking possible worlds into truth values" [Den81, p. 305], thus identifying a belief with what many take to be an intention or a meaning.[5] Now of course it may be the case that the processes in an agent's brain which give rise to the behavioural phenomena *via* which we attribute beliefs are themselves non-linguistic. However, we must remember that beliefs are ascribed at a certain level of description of the agent so that, even if the relevant processes subvenient to belief are intrinsically non-linguistic, we need not conclude that our ways of ascribing belief should be propositional, rather than sentential. As discussed above, there may be no interesting question as to what beliefs really are, so such considerations should not be allowed to persuade us of the supposed *de re* nature of belief.

The sense in which belief *is* a *de re* propositional phenomenon is as follows. Suppose two agents each have a belief that they would express as "it's raining." Agent a has the belief in London on Monday, b has it in New York on Wednesday. So a believes that it is raining in London on Monday, whereas b believes it to be raining in New York on Wednesday. They have different beliefs, and what distinguishes them is not anything linguistic, but rather the *de re* fact that London isn't New York, and Monday isn't Wednesday. However, for all practical purposes—explaining and making predictions about behaviour—the sentence "it's raining", understood in its appropriate context, is perfectly adequate. *Why did the agent take an umbrella? Because it believed that it was raining.*

Moreover, the *de re* content of the sentence that an agent would use to express her belief might not be adequate as an explanation or prediction of her behaviour. Consider an agent perpetually annoyed by mobile phones ringing on public transport who, upon hearing a phone continuously ring whilst on the train to London, gets increasingly annoyed. Each time it rings, she tries to locate the source of the annoying ring. Finally, she realises that she has left her own phone in her luggage at the end of the carriage, so comes to have a belief that she would most naturally express as 'it's *my* phone ringing.' This belief explains her subsequent actions—embarrassment, motion towards her luggage, apologies to the other passengers etc. John Perry

[5] See Lewis's [Lew75], for example).

considers a similar example in [Per93] and concludes that no replacement of the indexical characterisation of the agent's belief as 'it's *my* phone ringing' could account for this behaviour. The (true) belief that the annoying phone belongs to the passenger in seat 12A, for example, does not explain the behaviour unless we also add the belief that the agent would express as 'I am the passenger in 12A', itself an indexical sentence.

Following Perry [Per79], it is useful to distinguish between what the agent believes and her state of belief in so believing. As our embarrassed agent retrieves her phone, the other passengers in the carriage may well believe our agent to be the owner of the annoying phone, but they do not share our agent's feelings of embarrassment and the like. They all share the same belief—*who* owns the annoying phone—but they entertain that belief in different ways, and so are in very different belief states. Perry's conclusion is that there is something essential about the way we characterise such belief states in an agent centred way, using *I, me, here, now.* No substitute for 'I' or 'me' would allow us to explain the agent's egocentric behaviour. It is most natural, then, to classify belief states at a cognitive level, in terms of I-thoughts; and the way we typically attribute I-thoughts is through direct quotation: she believed "that's my phone." We classify belief states, therefore, using sentences. The same considerations apply when classifying desire states. If all the runners in the race want to win, for example, then they are all in the same (local, not total) desire state. Yet there is no one contender such that all the contenders want that person to win, so they all have different desires.

However, even with this distinction in place, it is still not correct to say that what an agent believes in having a belief is a function from worlds to truth values. On this view, one believe would believe the same thing in believing that Fermat's Last theorem is true and that $1 + 1 = 2$. Similarly, in believing any logical falsehood to be true, one would believe the same as one would in believing that $1 + 1 = 3$, i.e. the constant function taking any world to *false*. This same constant function would also account for beliefs about Superman and Pegasus. Moreover, one would believe the same thing in believing either that Bob Dylan or that Robert Zimmerman is a great songwriter. These are all unintuitive results; they do not square with what an ordinary speaker means by *what one believes* when one has a belief. What Stalnaker's view of propositions does achieve in a particularly elegant way is a characterisation of the truth-conditional content of a belief. We should conclude that whatever it is that people believe when they have a belief should not be identified with the truth conditional content of their belief.

We might agree that belief is a relation between an agent an a proposition but that propositions should not be understood as functions from worlds to truth values. An alternative is to consider propositions to be structured entities containing semantic values, i.e. particulars, properties, relations and descriptive conditions. This is known as the *Russellian* view, popularised by Kaplan [Kap89] (amongst others) and adopted by direct reference theorists. King [Kin96] considers *structured propositions*, a development on the Russellian notion which includes the entire syntactic structure of a sentence, represented in tree form, with semantic values appended to leaves. In Kaplan's framework, an utterance (or a sentence in a context) first expresses a (Russellian) proposition, which is then evaluated for a truth value at a time and a world.

Such propositions play a rôle intermediate between truth-conditional content and belief state classification. They do not appear to offer any advantage over Stalnaker's view in terms of specifying truth-conditional content and do not represent a complete solution to the problem of belief state classification. Structured propositions allow one to distinguish the belief that Fermat's Last Theorem is true from the belief that $1 + 1 = 2$ but not between beliefs which differ only in the *salva veritate* substitution of one semantic value for another, without altering syntactic form. For example, the beliefs that Dylan is F and that Zimmerman is F relate an agent to precisely the same Russellian or structured proposition.

I do not think it necessary to assume the existence of propositions to explain (or act as the bearers of) the truth of sentences, even in direct discourse. A sentence can only express a proposition if it can be suitably disambiguated and situated in a relevant context. The sequence of shapes "I am not here today", worked into the sand by a crab crawling to and fro, does not express a proposition. So we may assume that we are dealing with *statements* when we talk about the truth of a sentence in a context, i.e. a fully disambiguated sentence. Given that the truth of a statement can be determined at a world, why not use whatever mechanisms are used to establish which proposition the statement expresses to establish its truth value directly?

The objection here has to do with the distinction between rigid and non-rigid terms. Descriptions, for example, pick out whatever satisfies their descriptive content at the world that they are evaluated at, whereas proper names pick out the same individual across all worlds (in which that individual exists). Kaplan's thought is that there needs to be two stages to the truth-determining process: firstly, that of a sentence (in a context) expressing a proposition and secondly of evaluating that proposition at a world. However, these facts of reference and modality do not entail that we must treat propositions at *entities*. Rather, we should evaluate suitable world-insensitive versions of statements and always bear both the current world and the actual world in mind when evaluating statements. For example, by replacing 'Tony Blair' with 'the actual referent of 'Tony Blair'', we can evaluate 'Tony Blair is prime minister of England in 2006' at any world and obtain the right results (namely that is is a contingent truth).[6]

Returning to belief contexts, the question of whether a particular belief is true or not is precisely the same as the question of whether the sentence that we use to classify the agent's belief state is true in the appropriate context. The picture we have is as follows. We classify an agent's belief state using sentences. Moreover, what the agent believes in being in such belief states is characterised by the disambiguation of, and the addition of ACTUAL operators to, these sentences. We can explain both the agent's state of mind and what is believed in believing something using suitable sentences.

Dennett's worry here is that language "*forces* us on occasion to commit ourselves to desires altogether more stringent in their conditions of satisfaction than anything we would otherwise have any reason to endeavor to satisfy." Language is too specific for the specification of desire, for "you often cannot say what you want without saying something more specific than you antecedently mean" [Den87, p. 20]. These worries apply equally to the classification of belief states, "where our linguistic environment is forever forcing us to give—or concede—precise verbal expression to convictions that lack the hard edges verbalization endows them with" [Den87, p. 21]. We may object here that language frequently does not look as precise as Dennett would have us believe. Vagueness, in particular, is an intrinsic feature of natural language. Our predicates tend not to neatly partition the domain, but instead direct us to a sample to which the present case may be more or less similar. We make extensive use of vague quantifiers such as *for most* and, even when we use a determinate quantifier *for all*, the domain of quantification is nearly always contextually specified, but need not do so in a precise way (this latter consideration also applies to definite descriptions).

We can point to numerous examples in which an expression of desire

[6] Note that bearing both the current and the actual world in mind does not require a two-dimensional modal semantics, i.e. we need not evaluate statements at a *current world-actual world* pair. The correct semantics can be captured in *hybrid logic*. One evaluates a statement 'ϕ' by first of all assigning the actual world to a world-variable x using the '$\downarrow x$' binder and then jumping to a suitable world using 'E'. Then both the actual and the relevant possible world are captured in the *syntax* of the formulae being evaluated, rather than in its satisfaction clause.

suggests satisfaction conditions broader than our antecedent desire. This does not show that the desire does not have an intrinsic language-like component, but only that the agent chose the wrong way of expressing her desire. Moreover, in expressing a desire linguistically, one can appeal to all the usual pragmatic features usually associated with discourse. A desire to eat a low-fat meal, which excludes eating dust as a satisfaction condition, is perfectly well expressed as ‘I’d like something low in fat’ in a restaurant setting. Anyone thinking that serving the utterer a plate of dust would satisfy the request is not playing within the conventions of the game. We often say things that, taken literally, are either more general or more specific than we mean, but this does not imply that meanings cannot be expressed linguistically. It merely highlights how conventional practise allows us to express ourselves concisely and efficiently. The same holds for desire and belief. This is the first positive conclusion that I want to draw: a belief (or desire) state and what is thereby believed (or desired) should be characterised in terms of sentences.

Just what is the relationship between the agent and the sentences we use to classify her belief state? Perry proposes the notion of *acceptance*:

> One has a belief *by* accepting a sentence. ... [Belief] states have typical effects which we use to classify them. In particular we classify them by the sentences a competent speaker of the language in question would be apt to think or utter in certain circumstances when in that state. To accept a sentence S is to be in a belief state that would lead such a speaker to utter or think S [Per80, p. 45].

Talking in terms of acceptance of a sentence highlights the difference between being cognitively related to ‘Dylan is F’ and to ‘Zimmerman is F’. Corazza asks us to imagine the subject as having an indefinite number of sentence tokens and tokens of the psychological verbs ‘believe’, ‘desire’ etc. placed in front of her. Then “[e]ach time our agent entertains an attitude she is asked to do two things: (i) pick out a psychological verb and (ii) choose from among the sentences the one she would use to express her attitude. The sentence she picks out or points to is the sentence she accepts” [Cor04, 260].

However, here we are tempted to think of acceptance as a conscious mental activity, along the lines of making a judgement or an act of assent. If accepting a sentence is a necessary requirement for entertaining a belief one would, on this view of acceptance, have very few beliefs indeed. I have many beliefs to which there corresponds no explicit act of acceptance, for example the beliefs that there is a chair in front of me, that it will not move as I sit down on it, that it will bear my weight and so on. If the notion of acceptance is to be of any use, therefore, accepting a sentence must not be treated as a conscious mental act. Rather, acceptance should be cashed out in counterfactual terms, such that "there are infinitely many sentences one is disposed to [accept]" [Cor04, p. 263], allowing the analysis to afford an agent infinitely many tacit beliefs.[7]

Corazza develops this counterfactual condition in terms of the mental representations of the attributer and attributee. One attributes the acceptance of a sentence to another when:

> the attributee's token mental representation (or cognitive particular) is similar to the one that would cause the attributer, in the attributee's context, to utter S. ... two agents' token representations are of the same type insofar as they [accept] the same sentences [Cor04, p. 262].

'Similar' here means 'of the same type'. Corazza is explicitly committing himself to a representational theory of mind: "A mental state ... is a mental representation plus the attitude relation (belief, desire, etc.) the agent bears to the proposition. ... A belief state, for example, is a mental representation embedded within the BELIEF operator" [Cor04, p. 257]. Whatever the truth of the claim that the mind is representational in all (or most) of its aspects, the appeal to mental representations in explaining intentional attitudes is overplayed. Corazza allows for tacit beliefs, such that 'yesterday, I believed that there were no pink elephants dancing on campus' may be a true self-ascription. But what would a mental representation of *no pink elephants dancing on campus* consist in? Certainly not in a mental representation of this or that aspect of the campus, for a mental representation of pink elephants dancing on campus is perfectly compatible with all of these. Similar worries apply in the case of the conditional belief that $p \rightarrow q$. The reply is to claim that there are mental representations of p and q, combined by syntactic rules into logically complex 'representations'. We then arrive at a fully fledged language of thought hypothesis (see, for example, Fodor [Fod87, Fod90]). On this view, questions of psychological interpretation are

[7] Corazza uses the term *n-acceptance*, which does not entail *accepting as true*. This allows for an analysis of desire, supposition etc. where the agent does not take the n-accepted sentence to be true. However, in the case of belief, n-acceptance *does* entail accepting a sentence as true; hence I have stuck to the usual term 'acceptance' in the passage.

settled by appeal to the primitive semantic properties of the language of thought.

This account flies in the face of Wittgensteinian considerations concerning privacy, for it seeks to locate meanings within a private mental sphere. To paraphrase Wittgenstein, suppose that there existed in the mind some private entity claimed to be a primitive semantic property. In considering the minds of others, which on this picture are as a black box to me, "it would be quite possible for everyone to have something different in his box"—i.e. each person having a *different* kind of mental entity claimed as a meaning—"[o]ne might even imagine such a thing constantly changing" [Wit02, §293]. But if we suppose that talk of meaning has a use in a public language or, as Wittgenstein would have it, in a *language game*, then the purported private semantic entity:

> has no place in the language game at all; not even as a *something*: for the box might even be empty.—No, one can 'divide through' by the thing in the box; it cancels out, whatever it is [Wit02, §293].

Thus, even if one accepts the existence of symbols in the mind (captured functionally in terms of brain processes) that are governed by syntactic rules, this inner mental 'language' would still need to be interpreted. The problems of interpretation are discussed by Quine, who comments that "[t]he metaphor of a black box, so often useful, can be misleading here. The problem is not one of hidden facts, such as might be uncovered by learning more about the brain physiology of thought processes" [Qui70, p. 180]. Putnam is in agreement, for ""[m]ental representations" require interpretation just as much as any other signs do" [Put83, p. 154]. Thus, it seems that positing a symbolic system in the mind does not, in itself, answer our questions concerning the semantics of belief.

Moreover, it is clearly not sufficient simply to have a mental representation of a snake lying in the garden, even if the cause of the representation is of the sort that usually results in belief (such as perception), in order to believe that there really is a snake there. If I know that my eyes are bad and that I often mistake coils of rope for snakes on foggy days, for example, I am unlikely to believe that the world is as it is presented to me. Note that this might be the case even in cases of veridical perception. The claim that "mental representations are all we need to explain attitude ascriptions" [Cor04, p. 263] is surely too fast here; we also need an account of just what the BELIEF relation is, and this is precisely what Dennett claimed requires an assessment of a particular "*predictive strategy*" [Den87, p. 15].

Even with an account of the BELIEF relation in place, it is not the case that similarity of belief, even in identical situations, requires similarity of mental representation. We might imagine interaction with an alien race who behave similarly to us in similar circumstances, such that we would be tempted to explain and predict their behaviour in terms of beliefs and desires. However, it might nevertheless turn out that their mental representations (if they have any at all) are wildly different from ours. Should we then say that our ascriptions of belief to the aliens were strictly speaking false? Only if we take truthmakers for such ascriptions to be hidden facts, settled by the realising physical properties of the agent's consciousness. This is, of course, not the way we ascribe such attitudes at all. As Quine says, "[t]he problem is not one of hidden facts" [Qui70, p. 180]. Rather, Dennett's recommendation that we first decide to treat the agent as an intentional system [Den87, p. 17] seems essential when explaining *our* notions of belief and desire.

Following this line of thought, such attitudes are best cashed out in terms of their relationships to other attitudes in a holistic way, such that ascriptions of attitudes are ultimately used as a tool to explain and predict behaviour. This is the approach advocated by Quine and Dennett, as discussed in section 2 above. To recap, Dennett distinguishes normative from projective versions of the predictive account. According to the former, we ascribe the attitudes that an agent should have, given her situation; according to the latter, one ascribes the attitudes that one would have oneself in those circumstances. Incidentally, Corazza agrees with the latter position, holding that "[w]hen we attribute an attitude to someone, we often imagine ourselves in her situation" [Cor04, p. 259]. Stich is also in agreement that "[i]n saying what someone else believes, we describe his belief by relating it to one we ourselves might have. And we indicate this potential belief of our own by uttering the sentence we would use to express it" [Sti83, p. 79]. But perhaps the *locus classicus* of this view is found again in Quine's views on indirect quotation:

> we project ourselves into what, from his remarks and other indications, we imagine the speaker's state of mind to have been, and then say what, in our language, is natural and relevant for us in the state thus feigned. [Propositional attitudes] can be thought of as involving something like quotation of one's own imagined verbal response to an imagined situation [Qui60, p. 219].

Ascription is thus "an essentially dramatic act" founded in part on our "dramatic virtuosity" [Qui60, p. 219].

It might be thought that there is an intrinsic difference between quotation and reporting a belief, namely that quotation seeks to report something about the attributee's relation to the quoted *words*, whereas belief reports say something about *the world* according to the attributee. This view is the result of giving too much consideration to the use-mention distinction. We are not tempted to say that cases of mixed quotation are merely about words. For example, the sentence

Quine held that ascription is "an essentially dramatic act"

says that Quine held that ascription is an essentially dramatic act. It also conveys that the choice of phrasing is Quine's, not mine. The sentence is just as much *about* dramatic acts as it is about Quine and ascription, despite the former worlds being enclosed within quotation marks. So it is with other intentional ascriptions.

To sum up the discussion so far, several sentential accounts of belief appeal to the notion of acceptance, which must be cashed out in counterfactual terms to allow for tacit beliefs. However, appeal to mental representations does not give us the right kind of counterfactual conditions and can commit one to the language of thought view of belief. Rather, the correct analysis of acceptance should either be that an agent accepts what it should assent to, given its situation, or that one treats an agent as accepting what one would oneself assent to in similar circumstances. Note that on either view, it is hard to follow Perry's claim that acceptance is the contribution that the mind makes to belief, for both accounts allow for the ascription of beliefs to which the attributee's mind has made no contribution whatsoever. It is perhaps best to do without the notion of acceptance altogether and talk of belief in terms of the agent's potential to assent to a sentence (a conscious mental act). The question then is whether the relation between belief and (potential) assent is a normative or a projective one.

We have already seen that the former account treats agents as ideal reasoners. Incidentally, in the case of an ideal agent making ascriptions, the normative and the projective accounts should agree on what they ascribe. This highlights a potential problem with the projective account, namely that it will be more or less accurate (as a prediction of behaviour, say) to the extent that the attributer and attributee share reasoning ability, resources with which to reason and the like. An advanced chess player playing a novice might find her opponent's last move incomprehensible but still hold that the opponent believed it to be a good move.

In the following section, the example of such a *bounded* (but not out-and-out irrational) reasoner will be used to motivate an account of belief which locates an agent's beliefs, as we should ascribe them, between ideal rationality on the one hand and a description of the agent's experience on the other. The resulting agent is described as neither irrational nor as a perfect reasoner.

5 Bounded Rationality

In the previous section, it was concluded that belief states are best classified using sentences. There is also no barrier to classifying *what* the agent believes using sentences. The remaining question is how belief relates an agent to a sentence and whether the relation is best captured by ascribing the sentences that the agent should believe, given its circumstances, or by ascribing the sentences that the attributer would believe herself in those circumstances. Unsurprisingly, the focus of the discussion will center on which type of account provides the best explanation of agents with bounded rationality, i.e. agents which are neither irrational nor ideal reasoners.

Both types of account, normative and projective, make use of the notion of the agent's *circumstances*, which we may take to mean *relevant experience*. An agent's experience relates directly to belief through observation or rather, an agent's beliefs are ultimately based on the sentences that classify her observations, i.e. *observation sentences*, in Quine's sense [QU70]. This is precisely Quine's methodology: "let us ask no longer what counts as an observation, but turn rather to language and ask what counts as an observation sentence" [QU70]. Just what makes a sentence an observation sentence? Quine's answer is that:

> any second witness would be bound to agree with me on all points then and there, granted merely an understanding of my language. ... In short, an observation sentence is something that we can depend on other witnesses to agree to at the time of the event or situation described. [QU70].

Observation sentences provide the link between observation and assent, for a requirement of an observation sentence is "that all reasonably competent speakers of the language be disposed, if asked, to assent to the sentence under the same simulations of their sensory surfaces" [QU70].

It is clear that observation sentences cannot be of high logical complexity. That is, even if p is an observation sentence corresponding to an event e, '$p \wedge (\phi \rightarrow p)$' need not be counted as an observation sentence for e, even though it is logically equivalent to 'p'. Someone might observe e and assent to 'p' but not to '$p \wedge (\phi \rightarrow p)$', say if '$\phi$' is particularly irrelevant or complicated to understand. The only way to guarantee that any two distinct competent speakers will both assent to a sentence S on the basis of witnessing some event or situation, is for S to be logically simple or a report. Observation sentences are justified by experience alone, not by inference.[8]

It is beneficial to be generous with the notion of justification here, such that "the man in front of me is in pain" can count as an observation sentence. Knowledge of other minds, therefore, should not be counted as inferred knowledge. It is important too that reports are included as observation statements, provided that the way in which the embedded content is reported meets the criteria of an observation statement. Thus 'the man standing in that room said that S' will, but 'John said that S' will not count as an observation sentence, for not all observers could recognise the man standing in the room to be John. Observation sentences give us a base case for our ascriptions of belief to an agent: we ascribe, at the least, all the observation sentences that the agent's experience to date has justified.[9] As Quine says, "the ultimate evidence that our whole system of beliefs has to answer up consists strictly of our own direct observations—including our observations of our notes and other people reports" [QU70]. Let us call these the agent's *minimal beliefs*. The question that then arises is, what other beliefs should we ascribe to an agent, based on its minimal beliefs?

In the example of the two chess players, one of whom is far more competent than the other, the situation might be as follows. The first explains the rules of chess to the second and consequently should ascribe belief in these rules and knowledge of the positions of the pieces on the board to the other. The less competent agent then makes a legal but inadvisable move, which the more competent player certainly would not have made, in the same situation. In light of this, what additional beliefs should be ascribed to the less competent player? One line of reasoning runs as follows: people play games such as chess to win (amongst other reasons) and experience dictates that, in order to win, one should make the best move available. If the players did not share these beliefs, they could hardly be said to be *playing* chess at all—rather, they would more accurately be described as merely moving the pieces around. We should then infer that the less competent agent made the inadvisable move because she (falsely) believed it to be the best move

[8] There is of course a complication here, in that theory (arrived at through inference) can sometimes dictate what counts as an observation sentence.

[9] Compare Dennett's predictive strategy, according to which all of "the truths relevant to the system's interests (or desires) that the system's experience to date has made available" count as beliefs [Den87, p. 18].

available to her. However, this is neither the belief that the agent *should* have, given its experience, nor the belief that the more competent agent would have in that situation.

Rather, it is a belief that an agent might have who cannot follow through all the consequences of her chess-playing beliefs. Most people can only imagine how the game would progress given this or that move to a very limited extent and thus, a bad move may nevertheless appear to be beneficial when one of its consequences is unforeseen. It is not that the agent lacks the *ability* to reason in the right way—she knows all the rules of chess and how to apply them—but rather that she lacks the cognitive resources to make the right inferences. Compare this to a player who makes a bad move because he thinks that her opponent's queen may only move one square at a time. We can certainly blame this agent for the mistake, whereas the previous agent's mistake might be perfectly blameless. If one misses an opportunity to take the game, but it takes an advanced chess-playing computer several hours to discover the required sequence of moves, we would not hold the missed opportunity against the novice player.

So it is with many kinds of inference and in particular the inferential relations between beliefs gained directly as observation sentences and beliefs inferentially supported by observation sentences. An agent may be blamed for having incorrect beliefs when those beliefs are *mis*inferred from observation sentences (for example, by denying the antecedent), but failing to believe all the commitments of one's beliefs is, for the main part, blameless. Agents typically have limited resources with which to figure out what they should and should not believe in a certain circumstances. The chess playing agents might have a time limit, but they also have a finite amount of memory which limits their ability to 'look ahead' to a certain number of moves. Also, agents allocate these resources relative to their personal interests and needs such that considering a chess move might make less resources available for concurrent reasoning in other areas.

It should be clear that the kind of account required must be able to distinguish between what an agent *should* believe, given a set of observation sentences and what it *could* believe, given that same set and limited resources with which to reason. This is the distinction between what an ideal agent would believe in the same situation and what another agent with similarly bounded resources could come to believe. Belief should be captured by considering those sentences that an agent could explicitly assent to, using her allocated resources, determined according to Dennett's maxim that "one is interested only in ensuring that the system is rational enough to get to the particular implications that are relevant to its behavioural predicament of the moment" [Den87, p. 21]. Thus, given that a particular set of sentences treated as beliefs would explain the agent's behaviour and that a certain amount of rationality must be attributed to the agent in order to credit her with these beliefs, given what she has observed, we should credit her with that amount of rationality in the other beliefs we attribute.

In this way, tacit beliefs, such as the belief that there are no dancing elephants in this room, can be explained. An agent may not explicitly assent to 'there are no dancing elephants in this room' at a particular moment of time but believe it nevertheless. If questioned on the matter, she would respond that *there are no such elephants in the room.* There are a number of conventional as well as cognitive bounds on the resources available to an agent, for example, the time in which a response is required by normal discourse conventions, or the time in which a move must be made in a game of chess. This is why how an agent replies or responds within that allotted time bound is a good indicator as to what the agent believes.

It should be noted that *amount of rationality* is being used in a rather precise way here to indicate how the resources available to the agent determine which of its commitments it may and may not become aware of. There are several problems associated with this notion. If ψ is a consequence of $\phi_1, \ldots, \phi_n$ (say, in classical propositional logic), then

$$\frac{\phi_1 \ \cdots \ \phi_n}{\psi}$$

is a correct rule of inference. Then, as far as our ascription is concerned, any agent for whom $\phi_1, \ldots, \phi_n$ are beliefs obtained directly from observation sentences should be able to become aware of its commitment that ψ in one inference step. Then, for any of the agent's commitments, we would have to say that the agent believes it. Of course, this is not the result we desire and so we must not use *any* acceptable inference rule in making our ascription.

Related to this point is the differing psychological complexity of inference rules. Many find *modus tollens* more difficult to apply than *modus ponens*, such that an agent with sufficient resources to apply *modus ponens* n times may not be able to apply *modus tollens* as many times before its resources run out.[10] These examples show that we must fix a set of inference rules to use in our ascription such that other, more psychologically complex, rules can be explained. For example, in taking our set of rules to be the standard introduction and elimination rules for the Boolean connectives, the implication $(p \to q) \to (\neg q \to \neg p)$ corresponding to *modus tollens* can be derived

[10] My thanks to Stephen Mumford for making this point during a talk I gave at Nottingham.

in seven steps.[11] Thus, by counting inference steps in a system including *modus ponens* but not *modus tollens*, the latter rule is modelled as being more complex than the former. Of course the reverse also holds: this is why we *model*, rather than *explain*, the relative complexity of different rules of inference. That people frequently find reasoning with *modus tollens* more complex than with *modus ponens* shows that the model including the latter but not the former rule is to be preferred. In the case of artificial agents, we might have to revise these assumptions, for example, if we know that the agent reasons in a specific way using a fixed set of rules.

What of agents whose beliefs are to be explained through *mis*inference? An agent who is told that he will fail if he does not put the work in may well feel surprised or let down on being told that he has failed, after putting in plenty of work. We might explain his disappointment in terms of fallacious reasoning, from 'if x does not work hard, then x will fail' and 'x worked hard' to 'x will not fail' but our explanation might instead attribute the additional inductive belief that, in most cases, hard work results in a pass. If we have no additional knowledge of just how the agent came about his expectation to pass, we would be tempted to attribute this latter belief rather than explain the expectation as a mistaken inference through denying the antecedent of a prior belief. This highlights our need to rationalise the behaviour of others, such that we should prefer a rational explanation unless the available evidence strongly indicates otherwise. An example of this latter sort is found in Book 1 of Locke's *Essay Concerning Human Understanding*, where he explicitly says that universal assent to an Idea would imply that the idea was innate, but that since no Idea is universally assented to, none can be innate.[12]

The reason we feel the need to attribute intentional attitudes to agents in the first place is tied up with our need to rationalise and predict their behaviour. There may certainly be cases in which a reporter is forced to consider an agent's inferences to be invalid or irrational but, wherever possible, the assumption of rationality is the course the reporter should take. Our very practise of ascription is premised on the foundation of rationality such

[11] By assuming $p \rightarrow q$, then $\neg q$, then p). Deriving $(p \rightarrow q) \wedge \neg q \rightarrow \neg p$ also requires seven steps: assume the antecedent and eliminate '$\wedge$', then assume p.

[12] Locke takes the argument from universal assent to be that there could be innate Ideas "if it were true in matter of fact, that there were certain truths wherein all mankind agreed" [Loc97, §1.2.3] and goes on to argue that "this argument of universal consent, which is made use of, to prove innate principles, seems to me a demonstration that there are none such; because there are none to which mankind give an universal assent" [Loc97, §1.2.4].

that, if widespread irrationality were the norm, intentional explanation and prediction would cease to be of benefit. This is not to deny that irrational opinion is a worldwide phenomenon. But, as soon as we consider just how many mundane, everyday true beliefs even the most irrational agent has—that chairs tend not move as one sits down, that chairs tend to bear one's weight and the like—we can see that even agents with the most irrational opinions are likely to have mostly rational beliefs (that is, beliefs which are faultless, given their resource bounds).[13]

6 The Fan of Bounded Rationality

Above, I argued that candidate formal models of downwards revision could not be made to work. So as not to shirk my responsibilities, some formal account of additive rationality ascription is now required. I will give only a brief and fairly non-technical outline here. A model M is a relational structure which may be described by a modal logic containing the '$\Diamond$' operator. The domain of M is simply a set of points S (which, following standard practise will be called *states*), some of which will be related by a nontransitive serial relation T, called the transition relation, which forms a tree on these points.[14] Each point is labelled by a number of non-modal sentences of our language by the *labelling function* V, such that $V(s)$ is a set of non-modal sentences for each $s \in S$ (note that $V(s)$ need not be classically consistent or deductively closed).

The particularity of the models we are interested in comes in the way in which we fix T. Whenever Tsu, we say that there is a *transition* from s to u. These transitions model potential *atomic* inferences: the act of an agent inferring just one new formula from those it already knows. Thus whenever Tsu holds, u must be labelled just like s except that, in addition, u is labelled by one additional formula (i.e. for some ϕ, $V(u) = V(s) \cup \{\phi\}$ whenever Tsu). Here, I say that *u extends s* by ϕ. A state s may be extended by a formula ϕ when ϕ is the conclusion of a rule of inference that we expect the agent to use, whose premises match the formulas which label s (or rather, since such rules tend to be meta-rules containing sentence-variables, we should talk about ϕ being the conclusion under some substitution instance of a rule whose premises, under that same substitution, are all labels of s). In a model M, whenever a state s may be so extended, there is a state u suitably extending s

[13] Dennett makes a similar point at [Den87, p. 19, footnote 1].

[14] The restriction to models in tree form is inessential, as it is a theorem of normal modal logics that every model is bisimilar to a tree model. See, for example, [BdRV02].

such that Tsu. Intuitively, models correspond to possible chains of reasoning that the agent in question could perform, starting from the sentences which label the root. Suppose an agent believes each of the sentences that label the root of a model. Then a state at depth n in that model is labelled by the sentences that the agent could realise to be consequences of its beliefs in n steps of reasoning.

We apply a model as follows. First, we assign a minimal set of beliefs to the agent, following the method I have described. Call this set B_0. We label the root of our model M with all and only the elements of B_0. Now we have to fix what rules our agent reasons with, which will automatically fix T. Just which rules we select will depend on our setting and our purpose. If we are to model an AI system, for example, it makes sense to select the rules of inference that the system actually uses.[15] In the cases of human belief, we assume that the agent reasons using whatever rules we expect or are typical of human reasoning, including inductive reasoning and inference to best explanation.[16] Once we have fixed a set of rules, our model itself is fixed.

Let us look at the model we have built. In models that include certain deductive rules—natural deduction-style introduction rules, say—there will be no finite bound on the length of branches through the model. In the purely deductive case, the least transfinite fixpoint of each branch gives us the deductive closure of the sentences which label the root of the model (the minimal set of beliefs B_0). Such points are the closest states to the root lying on a branch but not reachable from the root in a finite number of transitions. They represent the commitments of any agent whose beliefs include B_0. Section 2 concluded that an agent's beliefs should be located between its minimal beliefs and its commitments. In terms of our model, the beliefs we should ascribe to the agent must lie somewhere between its root and its leaves. Just how far from the root they lie is a matter of deciding the degree of rationality we want to treat the agent as having: "one is interested only in ensuring that the system is rational enough to get to the particular implications that are relevant to its behavioural predicament of the moment"

[15] Many systems in AI are explicitly programmed in a rule-based fashion. Rule-based programming allows for a great degree of abstraction in specifying behaviour and consequently several rule-based agent architectures have been developed, e.g. SOAR [LANR87] and Sim-Agent [SL99]. Rule-based programming extensions are also increasingly being offered as add-ons to existing, lower-level, agent toolkits, e.g., JADE [BPR01] and FIPA-OS [PBH00].

[16] In the formal model described in [Jag06c] and [Jag06b], I only consider deductive rules; formulating formal rules for abductive reasoning is no small task!

[Den87, p. 21].

Suppose we find in our model a particular set of sentences which, treated as beliefs and together with the desires we ascribe, explain what we want to explain, e.g. the agent's behaviour. We look for the smallest such set of sentences, and find the state closest to the root of M which is labelled by all of these sentences. Call this state s; it has a certain depth δ in the model (not to be confused with quantifier depth), equal to the number of transitions required to reach s from the root. The parameter δ is in effect telling us how many steps of reasoning would be required for the agent to realise that the sentences that we are interested in are in fact consequences of its beliefs.[17] In order to make sense of the agent's behaviour for our purposes, we only need to consider the agent to be rational enough to reason to depth δ in the model.

The sentences that we should say the agents believes, then, are those sentences labelling any state of depth δ. In terms of our modal language, in which '$\Diamond\phi$' holds at a state u iff there is a transition to a state v at which ϕ holds, we say that our agent believes that ϕ iff $\Diamond^\delta\phi$ (that is, ϕ proceeded by δ '$\Diamond$'s) is satisfied at the root of the model. In fact, we can generalise this definition to any state in our model, since every state is the root of the tree formed by its descendants. If we parametrise our modal language by δ, we can define a sentential operator 'Bel' such that $\mathsf{Bel}\,\phi \overset{df}{=} \Diamond^\delta\phi$.

If the entire tree represents the reasoning possibilities of an ideal agent, with one possible line of reasoning per branch, we have limited our attribution of rationality by chopping off each of the branches at depth δ. We might imagine a wedge-shaped fan, whose sides are of length δ, held over the tree so that its sides run parallel to the outermost branches of the tree and so I call the model we have built a *fan model*. The area within the fan represents the belief states that the agent could reason to from the minimal set of beliefs we attribute it. The states we find along the bottom edge of the fan are thus the most advanced belief states that this agent could reach, given its bounded rationality. We should, therefore, attribute as beliefs whatever labels we find at states along the bottom edge of the fan.

As with other modal epistemic logics, it is easy to extend the account to incorporate multiple agents. Suppose we want to model agents $a_1, \ldots, a_n$. Let $\Delta = \delta_1 \cdots \delta_n$ be a sequence of length n, where each $\delta_{i \leq n}$ is the measure of rationality we want to assign to agent i. Models contain a family $V_1, \ldots, V_n$

[17] Our model is not in itself a justification network, but such a network can be extracted by inspecting which rule is fired and which sentence added in each transition. Then the δ parameter represents the justificatory complexity of the sentences we are interested in.

of labelling functions, one for each agent. The language $\mathcal{L}^{\Delta}$ is parameterised by Δ and contains belief operators $\mathsf{Bel}_1, \ldots, \mathsf{Bel}_n$ and a family of additional operators $\mathsf{B}_1, \ldots \mathsf{B}_n$ such that $\mathsf{B}_i\phi$ holds at a state s iff $\phi \in V_i(s)$. Then we define $\mathsf{Bel}_i\phi \overset{df}{=} \Diamond^{\delta_i}\mathsf{B}_i\phi$.

As it stands, this account is subject to one of the criticisms levelled against attempts to downwardly revise assumptions of perfect rationality in section 3 above, namely that an agent is assumed to be rational to degree δ across the board. But agents typically direct their rational enquiry in one direction or another. An agent who has followed through the consequences of her beliefs about quantum physics, for example, is not guaranteed to have been just as rational in her beliefs about ethics, or what constitutes sensible footwear.

However, the account presented here is unlike those criticised above in that this problem can be overcome by restricting our selection of states at depth δ and less to those that can be reached from the root without irrelevant inferences. Suppose the sequence of states $s_0 s_1 s_2$ occurs on a branch b such that s_1 extends s_0 by the sentence 'murder is wrong' and that s_2 extends s_1 by 'I should avoid wearing heels on icy days'. Under most classifications, the topic has shifted quite dramatically from one inference to the next. If we want to explain why the agent first put on high heels but then after checking the weather decided on a pair of flats, we can ignore branches such as b which include off-topic or irrelevant inferences.

Concretely, we might place all sentences in the language in an abstract relevance network, such that the longer the shortest distance between any two sentences, the less relevant they are to one another (the relevance relation is reflexive, such that every sentence is of the highest degree of relevance to itself). Then, we decide just how relevant we want our agent to be, say to degree r. We then return to our original chosen state s, whose labels allow us to explain the agent's behaviour, and look up all sentences ϕ of distance no more that r in the relevance network from one of the labels of s. A branch is then excluded from our considerations iff a state on that branch of depth no greater than δ extends a previous state by a sentence not selected from the relevance network.

It should be pointed out that, in practise, our choice of a degree of rationality δ is often not a perfectly precise matter. It seems odd that, on a particular choice of δ, an agent might believe ϕ and $\phi \rightarrow \psi$ but not ψ. With a choice of $\delta + 1$, on the other hand, we would say that a does indeed believe that ψ. This sounds somewhat unintuitive, but this is only to be expected in an account in which agent's beliefs are not deductively closed. This only becomes a problem when we try to classify belief states in terms of strict, numerical identity, i.e. when we say that a belief state including $\phi \rightarrow \psi$ and ϕ must also include ψ because the latter belief state must be identical to the former. This is really just a way of saying that the identity conditions on belief states includes the deductive closure condition. As I argued above, this is just not the case. Rather, we should say that the two belief states are sufficiently similar, in fact so similar that we feel it odd to say that an agent believing $\phi \rightarrow \psi$ and ϕ would not also believe that ψ. One-step inference always produces similar belief states but chains of inference may not.

The case is somewhat similar to Sorites-style problems involving vague predicates. Given a sequence of colour patches from dark red to light orange, we would find it rather artificial to impose a sharp boundary between the red and the orange samples, yet of course the end points are clearly different colours. If we agree with Dennett (as I have here) that belief "can be discerned only from the point of view of one who adopts a certain *predictive strategy*" [Den87, p. 15], then a particular predictive strategy may well impose a sharpened boundary, based on our reasons for predicting the agent's behaviour. Thus, "agent a believes that ϕ and $\phi \rightarrow \psi$, but not ψ" is by no means contradictory. Rather, such ascriptions fit in with our ascriptions of predicates such as "is bald" and "is red" in general.

The fan models developed here are versatile. In [Jag06a], I discuss the advantages of using such models to capture epistemic possibility. In these terms, an account of dynamic information can be developed which avoids the traditional problem of considering agents to be ideally rational reasoners with unbounded resources. In [Jag06c], on the other hand, I develop a temporal account of the *explicit beliefs* (what Dennett would term *opinions*) of AI agents, allowing one to build a model of an agent and check whether, for example, the agent could come to believe some sentence ϕ within a fixed time bound. As well as being versatile, the models developed here have many interesting logical properties, as discussed in [Jag06b]. For example, when modelling an agent with a fixed program (set of inference rules), the satisfaction relation '$\Vdash$' is decidable. Such properties make these models easy to work with. This adds support to my claim that the assumption of perfect rationality in modelling psychological notions is unnecessary, both conceptually and practically. I have presented a genuine account of belief states according to which agents are not modelled as perfectly rational reasoners. When combined with the logical results given in [Jag06b], we see that the formal models of this account are just as useful to logicians in modelling agents but, in the case of resource bounded agents, produce far more accurate results.

References

[BdRV02] Patrick Blackburn, Maarten de Rijke, and Yde Venema. *Modal Logic*. Cambridge University Press, New York, 2002.

[BPR01] F. Bellifemine, A. Poggi, and G. Rimassa. Developing multi-agent systems with a fipa-compliant agent framework. *Software Practice and Experience*, 21(2):103–128, 2001.

[Cor04] Eros Corazza. *Reflecting the Mind: Indexicality and Quasi-Indexicality*. Oxford University Press, 2004.

[Cre73] M.J. Cresswell. *Logics and Languages*. Methuen and Co., 1973.

[Dav85] D. Davidson. *Inquiries into Truth and Interpretation*. Clarendon Press, Oxford, 1985.

[Den81] D. Dennett. *Brainstorms*. MIT Press, Harvard, MASS., 1981.

[Den87] Daniel C. Dennett. *The Intentional Stance*. MIT Press, 1987.

[dS71] R. de Sousa. How to give a piece of your mind:or, the logic of belief and assent. *Review of Metaphysics*, 25:52–79, 1971.

[FH88] R. Fagin and J.Y. Halpern. Belief, awareness and limited reasoning. *Artificial Intelligence*, 34:39–76, 1988.

[FHMV95] R. Fagin, J.Y. Halpern, Y. Moses, and M.Y. Vardi. *Reasoning About Knowledge*. MIT press, 1995.

[FHV90] R. Fagin, J.Y. Halpern, and M.Y. Vardi. A nonstandard approach to the logical omniscience problem. In R. Parikh, editor, *Theoretical Aspects of Reasoning about Knowledge: Proc. Third Conference*, San Fransisco, California, 1990. Morgan Kaufmann.

[Fod87] Jerry Fodor. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. MIT Press, Cambridge, Mass., 1987.

[Fod90] Jerry Fodor. *A Theory of Content and Other Essays*. MIT Press, Cambridge, Mass., 1990.

[Hin62] J. Hintikka. *Knowledge and belief: an introduction to the logic of the two notions*. Cornell University Press, Ithaca, N.Y., 1962.

[Hin73] J. Hintikka. Surface semantics and its motivation. In H. Leblanc, editor, *Truth, Syntax and Modality*. North-Holland, Amsterdam, 1973.

[Hin75] J. Hintikka. Impossible possible worlds vindicated. *Journal of Philisophical Logic*, 4:475–484, 1975.

[Jag06a] Mark Jago. Imagine the possibilities: Information without overload. `http://www.nottingham.ac.uk/philosophy/staff/mark-jago/`, April 2006.

[Jag06b] Mark Jago. *Logics for Resource-Bounded Agents.* PhD thesis, University of Nottingham, 2006. Forthcoming.

[Jag06c] Mark Jago. Rule-based and resource-bounded: A new look at epistemic logic. `http://www.nottingham.ac.uk/philosophy/staff/mark-jago/`, April 2006.

[Kap89] D. Kaplan. Demonstratives. In J. Almog, J. Perry, and H. Wettstein, editors, *Themes from Kaplan*, chapter 17, pages 481–563. Oxford University Press, New York, 1989.

[Kin96] Jeffrey King. Structured propositions and sentence structure. *Journal of Philosophical Logic*, 25:495–521, 1996.

[Lak86] G. Lakemeyer. Steps towards a first-order logic of explicit and implict belief. In J. Y. Halpern, editor, *Theoretical Aspects of Reasoning About Knowledge: Proceedings of the 1986 Conference*, pages 325–340, San Francisco, Calif., 1986. Morgan Kaufmann.

[LANR87] J. E. Laird, A. A. Newell, and P. S. Rosenbloom. Soar: An architecture for general intelligence. *Artificial Intelligence*, 33:1–64, 1987.

[Lev84] H. J. Levesque. A logic of implicit and explicit belief. In *National Conference on Artificial Intelligence*, pages 1998–202, 1984.

[Lew75] David Lewis. Language and languages. In K. Gunderson, editor, *Language, Mind and Knowledge*, pages 3–35. University of Minnesota Press, 1975.

[Loc97] John Locke. *An Essay Concerning Human Understanding.* Penguin, 1997.

[Mal72] Norman Malcolm. Thoughtless brutes. In *APA Proceedings and Addresses*, 1972.

[PBH00] S. Poslad, P. Buckle, and R. G. Hadingham. The fipa-os agent platform: Open source for open standards. In *Proceedings of the Fifth International Conference and Exhibition on the Practical Appli- cation of Intelligent Agents and Multi-Agents (PAAM2000)*, pages 355–368, Manchester, April 2000.

[Per79] John Perry. The problem of the essential indexical. *Noûs*, 13:3–21, 1979.

[Per80] John Perry. Belief and acceptance. *Midwest Studies in Philosophy*, 5:553–54, 1980.

[Per93] John Perry. *The Problem of the Essential Indexical.* Oxford University Press, Oxford, 1993.

[Put83] H. Putnam. *Realism and Reason, Philosophical Papers III*, chapter Computational Psychology and Interpretation Theory. Cambridge University Press, Cambridge, 1983.

[QU70] W.V.O. Quine and J.S. Ullian. *The Web of Belief.* Random House, New York, 1970.

[Qui60] W. V. O. Quine. *Word and Object.* MIT Press, Cambridge, Mass., 1960.

[Qui70] W. V. O. Quine. On the reasons for indeterminacy of translation. *Journal of Philosophy*, 67:178–83, 1970.

[Ran75] V. Rantala. Urn models. *Journal of Philosophical Logic*, 4:455–474, 1975.

[Sel56] W. Sellars. Empiricism and the philosophy of mind. In H. Feigl and M. Scriven, editors, *The Foundations of Science and th Concepts of Psychology and Psychoanalysis.* University of Minnesota press, 1956.

[SL99] A. Sloman and B. Logan. Building cognitively rich agents using the sim agent toolkit. *Communications of the ACM*, 42(3):71–77, March 1999.

[Sta76] R. Stalnaker. Propositions. In A. MacKay and D. Merrill, editors, *Issues in the Philosophy of Language*. New Haven, Yale, 1976.

[Sti81] S. Stich. Dennett on intentional systems. *Philosophical Topics*, 12:38–62, 1981.

[Sti83] S. Stich. *From Folk Psychology to Cognitive Science*. MIT press, Cambridge, Mass, 1983.

[Whi03] M. Whitsey. Logical omniscience: a survey. Technical Report NOTTCS-WP-2003-2, School of Computer Science and IT, University of Nottingham, 2003.

[Wit02] Ludwig Wittgenstein. *Philosophical Investigations*. Balckwell, 2002.