Belief and Bounded Rationality

Mark Jago

1 Introduction

Since Quine's *Word and Object* [Qui60], there has been more-or-less general agreement on the correct treatment on the status of talk about 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"; they are "essentially dramatic" idioms [Qui60, p. 219].[1] There are, of course, disagreements within this general viewpoint, particularly as to the exact nature of the "dramatic" rôle played in ascribing intentional attitudes. 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.

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 much. I will then suggest another method, based on Dennett's, which avoids the problem. I will conclude by showing how this final method delivers a notion of epistemic possibility which, while satisfying many of the relevant intuitions that philosophers and logicians assure us they have, does not treat agents as perfectly rational reasoners.

[1] See also Sellars [Sel56].

2 The Intentional Method

Dennett's *intentional stance* 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 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 the method of ascribing beliefs along these lines the *intentional method*. There are two main problems with Dennett's account. To begin with, it is hard to explain false belief; and secondly, 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. I propose to ignore the former worry (which Dennett discusses at [Den87, pp. 18–19 and 83–88]; see [Sti81] for criticism of this view) and concentrate on the latter.

Dennett [Den81, chapter 16] distinguishes opinion, which is a classical on-off affair, from belief which may be a matter of degree, governed by Baysian rules. Opinion is a matter of assent (or, as Perry [Per80] has it, of accepting a sentence as true) 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 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 which they could not, given their cognitive limitations, assent to.

This motivates the following question: what notion do we capture when we treat agents as perfectly ideal Baysian 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. 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 of 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 of 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 which would otherwise, such as one's judgements about what to have for tea committing one to $p \to q \vee q \to r$, 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 Bounded Rationality

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 the intentional method 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 the intentional method incorporates a formal model of belief—a Baysian model, for instance—these approaches 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 competency to an agent, we cannot give a method of downwards revision for the way we ascribe rationality to that agent in general.

Hintikka's notion of logical competency comes in both a syntactic and a semantic form, but the results are the same either way. Here, I describe the semantic version,[2] which 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 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*. The domain is conceived as a huge urn from which individuals may be drawn (the urn metaphor is taken from elementary probability theory). Sequences of quantifiers embedded one within the scope of another are restrictions on draws from the urn. Now, a classical model is one in which the contents of the urn remains constant between draws—such models are known as *invariant* models. Rantala then considers *changing* models, whose urn has a mechanism attached which may alter the contents from one draw to the next. In this way, sentences which are classically invalid may nevertheless be satisfied by an urn model. The level of inconsistency in a urn model is viewed as the number of draws which occur before any change in the available individuals takes place. Suppose the largest number of nested quantifiers in a sentence ϕ is d (d is said to be the *depth* of ϕ). Then, if the domain/urn in a model M remains constant for at least the first d draws, M will agree with classical models as to the validity or logical falsehood of ϕ. Such models are called d-invariant.

Hintikka's idea is to use the parameter d as a measure of an agent's logical competency, for sentences with deeply embedded quantifiers are harder to understand than those without. The more competent the agent, therefore,

[2]The syntactic account is in terms of *surface information*—see [Hin73].

the larger the value of d. An agent whose competency is d will be able to recognise the validity of all valid sentences whise depth does not exceed d, but might get it wrong in the case of more complex sentences. We thus have a way of ascribing first perfect rationality to an agent, in line with the intentional method, and then revising downwards. To do so, we figure out the value of d needed to make predictions about the agent's behaviour and consider all d-invariant models. Since some of these models will be changing models—i.e. their domain will alter after n draws, for some $n > d$, the agent will believe in some classically invalid sentence, so could not be said to be perfectly rational.

However, agents remain believers in all instances of propositional tautologies on this account (for when there are no quantifiers involved, urn models agree with classical models). 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. Secondly, 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 intentional method, 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. Our 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]

Another 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 intentional method 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 to be primitive, or supplied 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 the chair will

[3]Similar examples are discussed in [Jag06b, ch 2, §1.4.2].

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

remain where it is whilst I attempt to sit on it.

Suppose we mark such beliefs as being evidentially (as opposed inferentially) justified and then calculate the justificatory complexity of the other beliefs based on the shortest path in the justification network from that belief to a set of beliefs which supports it. We could then revise downwards by throwing our those beliefs of higher justificatory complexity first. Our measure of rationality then would be the agent's ability to reason to beliefs of certain justificatory complexity form a set support set. However, this view falls to the second objection raised against Hintikka above. In typical cases, there is not a uniform degree of justificatory complexity throughout the beliefs at the periphery of an agent's justification network. We have an additional problem in the case of beliefs which may be justified is more than one way. For example, we cannot tell if 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 '$\vee$'. 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 we could not 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.

The view that we treat the agent as a perfectly rational ideal reasoner in ascribing it beliefs goes hand in hand with the view that, in so doing, we are classifying the agent's belief state in terms of the possble worlds that it considers possible. This is the view presented in Hintikka's seminal *Knowledge and Belief* [Hin62], where the guiding idea is that, in believing something, one rules out all contrary possibilities. Since the theory of each world is deductively closed, so is the agent's set of beliefs. The resulting logic is therefore at least as strong as the modal logic **K**, with knowledge and belief operators distributing over '$\wedge$' and '$\rightarrow$'. It follows that any agent modelled in this way is modelled as having perfect rationality: agents believe all consequences of their beliefs, believe all tautologies, and have perfectly consistent beliefs. This has been dubbed the problem of *logical omniscience*. The problem, of course, is that real agents are not logically omniscient.

There have been numerous attempts to modify the possible worlds models to avoid logical omniscience, for example by basing the notion of a world on relevant rather than classical logic.[5] However, such attempts are badly motivated and it can be shown that agents remain logically omniscient in relevant (rather than classical) logic (see [Jag06b]). Another view, presented

[5]See, for instance, [Cre73, Lev84, Lak86, FHV90]. These accounts are reviewed in [FHMV95]and [Whi03].

in *Belief, Awareness and Limited Reasoning* [FH88] is that an agent's beliefs do indeed form a perfectly rational theory, but are then filtered through an 'awareness' filter. Awareness, on the other hand, is a purely syntactic notion. It is therefore possible to alter the properties of awareness without modifying the underlying framework of (idealised) belief. In fact, we 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]

Now, an agent's belief set $\mathbf{B}$ may only be deductively closed to the extent that the corresponding awareness set $\mathbf{A}$ is.[6] If our language is $\mathcal{L}$, define a closure operator Cl of type $2^{\mathcal{L}} \longrightarrow 2^{\mathcal{L}}$ such that $\phi \in \mathbf{B}$ implies $\phi \in Cl(\mathbf{B})$ and $Cl(Cl(\mathbf{B}) = Cl(\mathbf{B})$. Then it is easy to show that $\mathbf{B} = Cl(\mathbf{B})$ only if $\mathbf{A} = Cl(\mathbf{A})$ (the proof is simple, by induction on the structure any $\phi \in \mathbf{B}$ and the definition of belief in terms of awareness). I call this the *awareness closure principle*. So, 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 $\mathbf{A}$ are precisely those formulae which the agent *could* determine in a specified space and/or time bound. This is, roughly, the notion I will propose below, although I will make no use of the evidently spurious notion of awareness. Rather than describing an agent as a perfectly rational reasoner in terms of possible worlds, and then somehow using syntactic criteria to pare down the results, I want to argue that we can use syntactic elements to characterise an agent's beliefs directly.

[6] In the modal logic, define i's *belief set* $\mathbf{B}$ at world w as $\mathbf{B}_i^w \stackrel{df}{=} \{\phi \mid \mathcal{M}, w \Vdash \mathsf{Bel}_i\phi\}$. But we may ignore the agent and world parameters here, and consider the actual beliefs of a single agent.

4 Propositions vs sentences

Above, I briefly discussed some relations between belief and opinion, assent or acceptance of a sentence. We might also include *judgements* in the latter category which, according to Dennett [Den81, Chapter 16], Perry [Per80] 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 the chicken believes (or thinks) that going to the farmer is a way of getting fed, it certainly hasn't judged or formed the opinion that this is so; nor has it assented to or accepted that statement. Forming judgements and opinions, and assenting and accepting statements are conscious mental acts, whereas having beliefs might be viewed as a different class of mental phenomenon altogether, operating on a more fundamental, sub-personal level. This is why it makes sense to attribute beliefs to an agent which it has not explicitly considered.

However, this does not licence the claim that, whilst judgement, opinion, assent and acceptance is to be cased 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.[7]

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 also be semantic, rather than syntactic. 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 semantic or propositional nature of belief.

The sense in which belief *is* a semantic, 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

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

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" is perfectly adequate. *Why did the agent take an umbrella? Because it believed that it was raining.*

However, the *de re* content of sentence which 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 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 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, it is useful to distinguish between what the agent believes and their 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.

Dennett's worry here is that language "*forces* us on occasion to commit ourselves to desires altogether more stringent in their conditions of satis-

faction 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." 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." Now, 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 some* 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 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 literally 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 conclusion that I wanted to arrive at: belief (and desire) states are to be characterised in terms of sentences.

5 The Fan of Bounded Rationality

Now I turn to what I take to be the correct way to characterise an agent's belief state. Dennett took the correct method to be one of first assuming the agent to be perfectly rational and then revising downwards. I have argued that this approach cannot be made to work. Rather, the correct method in classifying a belief state is as follows. We first add statements which

the agent's actions and expressed opinions indicate it as assenting to. We then add the mundane, everyday truths for which the agent's experience is adequate direct evidence. These are the truths which an agent may come to believe in an experiential and non-inferential way. The belief that my car is not in the parking bay falls into this category, whereas the the belief that my car has been stolen, inferred from this belief plus my recollection of having parked it in that very bay earlier (plus the general belief that cars tend not to move by themselves), does not. We "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] with the proviso that these truths are directly experienced and not inferred. We would also add relevant desires, in line with the intentional method. The difference between this step and Dennett's suggestion is that we do not attribute any inferential ability at all to the agent, but only the ability to gather and correctly conceptualise the evidence of its senses. Finally, we add statements which can be seen to be supported by sets of statements already added within the limits of the agent's bounded rationality. In this way, we can make sense of an agent being rational enough, given our interpretational purposes.

By way of example, suppose the first step of our characterisation of agent a's belief set includes the statements that *black clouds indicate that it is likely to rain later* and *umbrellas prevent one from getting wet in the rain* (perhaps because the agent's previous behaviour that the agent is of such an opinion) and the second interpretational step adds the statement that *there are black clouds overhead* and that a desires to avoid getting wet. Now, we can explain why the agent took an umbrella with her in the morning and we only have to ascribe a fairly limited amount of rationality to a in doing so. We certainly would not commit ourselves to a's believing all propositional tautologies or having the ability to solve the Riemann hypothesis on the basis of our explanation of why she took the umbrella.

Rationality has come into the picture in the following *additive* way. Where the agent has sufficient resources available for reasoning, we assume that it will use these resources in a more or less rational way. This is not to say that it will be a perfectly efficient reasoner and always choose the shortest path of reasoning to a given view. But we may assume that the agent will reason as well as we would, were we in the agent's situation. We don't have to concern ourselves with whether the agent reasons using *modus ponens*, for example. 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 domain and in a specific way.

The first two stages of the method just described give us sentences which

express what we might call the agent's *minimal beliefs*. These will include sentences expressing the agent's judgements and opinions, or the sentences that the agent assents to, which we get from observing the agent. That all such sentences express beliefs of the agent is echoed in Perry's thought that acceptance 'is an important component of belief. It is the contribution the subject's mind makes to belief. One has a belief *by* accepting a sentence" [Per80, p. 45]. However, following Dennett's line on the *status* of beliefs, we might want to say that an an agent believes much more than it has ever consciously considered, and therefore more than it has ever accepted, assented to or judged to be the case. This is why the second stage of the method is necessary. Suppose we call these first two stages of belief ascription the *minimal intentional method*. The exact details are not particularly important. What is important is that, in using such a method, we are not making grand demands of the agent's capacity for rational deliberation.

Now, of course it may be that Dennett's overall approach is radically wrong (although I suspect that it could not be far form the mark). One might hold, along with Fodor, that questions of psychological interpretation should be settled by appeal to primitive semantic properties as would be found, for example, in a language of thought. The problem here is that even an inner mental language would face the problems of interpretation which first moved Quine to write 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]. Although I believe this to show that an account along the lines of Putnam's (with the modifications suggested above) must be the right one, what I have to say below should be perfectly compatible with Fodor's general picture, or any account which provides a set of sentences which are taken to be the agent's minimal beliefs.

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 the 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 serial relation T, called the transition relation, which forms a tree on these

points.[8] Each point is labelled by a number of non-modal sentences of our language.

So much is standard fare for modal logics, give or take some terminology. The particularity of the models we are interested in comes in the states which may be related by T. Whenever Tsu holds, u must be labelled just like s except that, in addition, u is labelled by some additional formula. So, for some formula ϕ, we have $V(u) = V(s) \cup \{\phi\}$ whenever Tsu. Here, I say that u *extends* s by ϕ. Moreover, T is in a sense greedy, in that whenever a state can be extended by a formula ϕ, there is always a state u extending s by ϕ such that Tsu.

Just what does it mean to say that a state *can* be extended by a formula?—or, rather, when could a state not be so extended? Here we come to the principle idea: transitions between states model the agent's *atomic* inferences—the act of inferring just one new formula from those it already believes. In principle, these inferences could be of any type—deductive, inductive and abductive. We might characterise all such types of reasoning as *rule-based* reasoning (this is what gives it its normative flavour), and purely deductive rules as those which always give rise to monotonic reasoning (reasoning in which derived conclusions never need to be withdrawn without also withdrawing their premises). A state s may be extended by a formula ϕ when ϕ is the conclusion of a rule of inference 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 such that Tsu. We don't have to represent these reasoning rules explicitly in the model; they are captured by regularities in the way T relates similar states to similar states.

So much for the details of our models. What use are they? We apply them as follows. First, we run what I have called the minimal intentional method and arrive at a set of sentences, all of which express beliefs of the agent (it is likely, of course, to have many more beliefs than this). 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 (we may assume we have an infinite supply of states and that V labels states randomly, so that we have all possible states available—

[8] 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].

note that we make no assumption that each $V(s)$ be classically consistent or deductively closed). 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.[9] 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.[10] 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—i.e. the set of minimal 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 that any agent would enter into on believing the sentences that label the root of the model to be true. Section 2 concluded that an agent's beliefs lie in 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 (note that this is a decision required by any account based around the intentional method).

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. We are in effect saying that, in order to make sense of the agent's behaviour for our purposes, we only need to consider the agent rational enough to reason to depth δ in the model. The

[9] 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].

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

sentences that we should say the agents believes, then, are those sentences labelling any state of depth δ. In the pure deductive case, the labels of states at any depth subsume those of less depth, but this will not be the case in general in a non-deductive setting. By picking only those labels at depth δ, we ignore both those sentences that the agent would not be sufficiently rational to assent to (these are the labels on states of greater depth than s) and those sentences that the agent could assent to, but then later realise to be mistaken and withdraw its assent from (the labels on states of depth less than s). 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. We may use a modal language $\mathcal{L}^\delta$ parameterised by δ containing a sentential operator 'Bel' such that $\mathsf{Bel}\,\phi \stackrel{df}{=} \Diamond^\delta\phi$.

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 are of length δ, held over the tree so that its sides run parallel to the outermost branches of the tree. The area within the fan represents belief states which the agent could reason itself into from the set of minimal beliefs we attribute it. The states we find along the bottom edge of the fan are thus the most advanced belief states which this agent could reach, given its bounded rationality. We thus say it believes 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 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 $V_1, \ldots, V_n$ 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 $\mathsf{Bel}_i\phi \stackrel{df}{=} \Diamond^{\delta_i}\mathsf{B}_i\phi$.

As it stands, this account is subject to the same criticism 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 which can reached from the root with no irrelevant inferences. Suppose the sequence of states $s_0 s_1 s_2$ occurs on a branch b and that s_1 extends s_0 by the sentence "I should avoid wearing heels on icy days", and that s_2 extends s_1 by "murder is wrong". 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 her high heels but then after checking the weather decided on a pair of flats, we can ignore branches such as b which model 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 as 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 the 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 may not be a perfectly precise matter. Suppose we follow the method I have suggested and attribute belief in ϕ and $\phi \rightarrow \psi$ to agent a. Must we also say that the agent believes ψ? It seems odd that the agent would not have this belief and yet we may have picked the least deep state in the tree at which ϕ and $\phi \rightarrow \psi$ both appear, in which case our method need not say that the agent believes ψ. With a choice of $\delta + 1$, on the other hand, we would say that a does 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.

The problem arises 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 is 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 just is 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 produced similar belief sets 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 endpoints are clearly different colours. If we follow Dennett in 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 around whatever reason we are predicting the agent's behaviour. Thus, "agent a believes that ϕ and $\phi \rightarrow \psi$, but not ψ" is by no means contradictory. Rather, the account I have presented here, on which this ascription is satisfiable, highlights how our practises in ascribing belief fits in with our ascription of predicates such as "is bald" and "is red" in general.

The kind of model developed here is versatile. In [Jag06a], I show 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 [Jag06b], 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.

[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.

[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 ideal rationalality. `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.

[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.

[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.

[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.

[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.

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