Chapter 5
The Rationality Assumption

Richard Dub

Abstract Dennett has long maintained that one of the keystones of Intentional Systems Theory is an assumption of rationality. To deploy the Intentional Stance is to presume from the outset that the target of interpretation is rational. This paper examines the history of rationality constraints on mental state ascription. I argue that the reasons that Dennett and his philosophical brethren present for positing rationality constraints are not convincing. If humans are found to be rational, this will not be because a presumption of rationality must be built into the deployment of the Intentional Stance. It will be an empirical finding. Rationality will be an outcome of mental state ascription rather than a condition on ascription.

5.1 Forefathers

Daniel Dennett studied under Quine at Harvard and under Ryle at Oxford. It is only moderately procrustean to say that Intentional Systems Theory is what you get by stirring together Quinean and Rylean metaphysics of mind. Quine provided tough-minded naturalism and an emphasis on the holistic, indeterminate, and irreducible nature of intentional language; Ryle provided a sensitivity to ordinary language that resisted eliminating mental talk as a dispensable dramatic idiom. Dennett's signature ingenuity was the alchemical spark needed to catalyze the reaction between the two.

Nowhere are Dennett's twin influences as keenly felt as in his first book. *Content and Consciousness* offers a germinal version of the Intentional Systems Theory that Dennett still maintains to this day. The debt to his philosophical forefathers in the book is explicit, and this makes it an especially fruitful place to turn to when attempting to fit the Intentional Stance within a historical tradition. In this chapter, I'll be exploring the history of one of the more controversial features of Intentional Systems Theory: its adherence to *a rationality assumption on belief ascription*. According to Dennett (both then and now), to apply the Intentional Stance – that is, to interpret an individual as having a mind – involves an assumption that the individual is rational.

R. Dub (✉)
University of Geneva, Geneva, Switzerland
e-mail: richard.dub@gmail.com

C. Muñoz-Suárez, F. De Brigard (eds.), *Content and Consciousness Revisited*,
Studies in Brain and Mind 7, DOI 10.1007/978-3-319-17374-0_5

Many philosophers bristle at the suggestion. Without some fancy footwork, the claim that believers are necessarily rational simply looks empirically false. One occasionally gets the sense that some philosophers think the rationality requirement is not just wrong: they think it is absurd. Because of this, the force of their arguments comes down to their ability to convey goggle-eyed incredulity through text.[1] In addition to the "obvious" irrationalities we experience in ourselves and in others, there is scarcely an end to findings in psychology and behavioral economics purport to demonstrate various ways in which we are all exceedingly irrational. For instance, the work of Kahneman and Tversky is often presented as evidence for a natural human tendency to make various errors in probabilistic or conditional reasoning (Kahneman et al. 1982; Thagard and Nisbett 1983; Stich 1985; Cherniak 1986).

It is perfectly legitimate to argue against the rationality assumption by offering apparent counterexamples, but the method does not get to the heart of the matter. In what follows, I challenge the rationality assumption by challenging Dennett's need for such an assumption in the first place. Why does Dennett argue for a rationality assumption at all? What functions is it meant to serve? If these functions are legitimate, can they be served by other means? There are two similar but distinct arguments for the need for a rationality requirement in Intentional Systems Theory, each bequeathed to Dennett by his philosophical forebears. One of these arguments comes from Dennett's Quinean heritage; the other comes from his Rylean side. I'll develop these lines of argument, and show that neither is successful.

Thus, the main goal of this paper is to diagnose and reject Dennett's stated need for a rationality assumption. However, this leaves us with a new problem. What is the upshot if the arguments for the rationality assumption are unsuccessful? Should we drop the assumption? What would happen were it dropped? I'll argue (as a secondary thesis) that, in the end, not very much would change. A version of Intentional Systems Theory without a rationality assumption won't necessarily end up rendering the verdict that our neighbors are irrational. In fact, it might still well have us ascribe largely rational beliefs.[2] On this version of Intentional Systems Theory, the rationality of our neighbors (if they are indeed rational) will be an empirical finding rather than something to be settled before empirical investigation has begun.

5.2 The Quinean Lineage

Dennett is not the only figure who has argued for rationality constraints; Donald Davidson (1982) and David Lewis (1974) also include rationality constraints in their theories of mind in the form of "principles of charity." It's not surprising that there should be theoretical affinities between these three. All are interpretivists,

[1] A parody argument: "It's simply irrational to conclude that people are rational! Therefore, Dennett's theory is self-refuting."

[2] The extent to which people are actually rational or irrational is something that I will remain agnostic about for the purposes of this piece.

holding that interpretation is an important feature in the assignment of mental states. But more importantly, all are students of Quine. This gives us reason to analyze them together. (When studying an organism, if you don't know what a particular anatomical structure is for, it is sensible to look at homologous structures in the organism's ancestors and cousins. Likewise, it makes sense to look at the development of rationality requirements within this philosophical clade to see what similarities and differences we can tease out.) An ancestral form of the principle of charity can be found in Quine's *Word and Object* (1960b), so it makes sense as a starting point for our investigation.

Quine's principle of charity first appears during a discussion of radical translation. Quine famously argued that translation between languages would always be beset with indeterminacy. However, the existence of multiple contending translations does not mean that anything goes and that no translation is better than any other. Quine argued that in addition to respecting stimulus meanings, we ought to abide by certain maxims of translation which would have us prefer certain translation manuals to others. The principle of charity is one such maxim: it would have us rule out translations resulting in logical silliness. Take Quine's field linguist, charged with translating a language he has never heard before. He notices that speakers always assent to utterances of the form ⌜q ka bu q⌝. This counts as evidence against translating 'ka' as 'and' and 'bu' as 'not'. Such a translation would have the speaker assenting to contradictions, and so imputes unacceptable silliness. The principle of charity is what motivates Quine's (1960a) famous declaration that "prelogicality is a myth of bad translators."

Those following in Quine's footsteps took the principle of charity to be inculpated in projects wider than just linguistic translation. For Davidson and Lewis, the principle of charity is a constraint that preserves rationality during radical interpretation. Radical interpretation is unlike radical translation in that it is not purely linguistic; it also ascribes mental states to an agent. The principle of charity here is much the same. Quine introduced a principle of charity on radical translation to rule out translation manuals that would impute logical silliness; the reason for introducing a rationality constraint on radical interpretation is to pare down on an otherwise unbridled indeterminacy that would plague mental state ascription.

What is the source of such unbridled indeterminacy? In presenting his argument for the rationality assumption in *Content and Consciousness*, Dennett includes a particular argument of Quine's. It is worth quoting Dennett at length:

> Quine and Chisholm also present arguments about believing and intending, of which the central point is that efforts to provide behavioural analyses of these two phenomena are doomed by a vicious circle of implications. Take, for example, the belief that it is raining. What behavior would clinch it that A believes it is raining? No matter what is suggested, it will turn out that this is a clincher demonstrating that A believes it is raining *only* if we assume that A has some particular purpose or intentions. […] A's finding a tree or roof to stand under is no more evidence, for it depends on A's intending to stay dry. If ascription of belief always depends on an assumed ascription of intention, the converse holds as well. A's intention to stay dry is not behaviorally demonstrated by his cowering under the tree except on the assumption that he believes it is raining, that he believes that he would get wet if he

> did not stay under cover, and so forth. A survey of the other Intentional and mongrel Intentional idioms shows that the use of any one of them has implications about beliefs and intentions, so the circle that prevents a behavioural paraphrase of belief and intention sentences infects the whole realm of the Intentional (Dennett 1969, 31–2).

Dennett goes on to discuss how this argument establishes the holistic nature of mentalistic vocabulary, and therefore its irreducibility to a purely extensional language. But he also takes this section to establish that "intentional explanations presuppose the appropriateness of sequences they purport to explain." That is, this section is also taken to establish the rationality of the actor.

How does it do so? If we see A standing under a tree, we could interpret him as having a desire to stay dry, a belief that he'll stay dry if he stands beneath the branches, and an intention to do so. Or, we could interpret A as *wanting* to get wet, and believing that he'll get wet by going into the rain, but *irrationally* deciding to stay under the tree. We could, in other words, impute silliness to him. If interpretation is to make a lick of sense, silliness must be ruled out. The apparent need for maxims of interpretation is borne from the holistic nature of mental state ascription. Holism of the mental implies that many mental states get attributed at once, as a package deal. Absolutely unfettered interpretation would allow you to attribute whatever mental state you want, provided you compensate elsewhere.

This particular argument for a rationality requirement doesn't receive as much play in *Content and Consciousness* as does the one that I will call the Rylean argument, but it does play an increasingly prominent role in Dennett's writings as time goes on. For instance, he later writes,

> The assumption that something is an intentional system is the assumption that it is rational; that is, one gets nowhere with the assumption that entity x has beliefs p,q,r,… unless one also supposes that x believes what follows from p,q,r,…; otherwise, there is no way of ruling out the prediction that x will, in the face of its beliefs p,q,r,… do something utterly stupid, and, if we cannot *rule out* that prediction, we will have acquired no predictive power at all (Dennett 1978, 17, my italics).

According to Dennett, we need to "rule out" certain predictions. This is precisely why Quine, Davidson, and Lewis also hold fast to a principle of charity.[3] There are important differences between the three sons of Quine, of course. For one, they each have different opinions on the material one uses as input for the interpretive process. Davidson admitted publicly observable behavior, paying particular importance to the sentences that one asserts. Lewis allowed all physical facts, whether public or not, to be used as input for radical interpretation. Dennett can plausibly be read as allowing behavioral dispositions as well as the interpreted individual's (objective) goals or reasons as input.[4] Moreover, they all have different conceptions of what sort of norms of rationality are guaranteed. Still, they all agree that there is a need to

[3] Dennett also accepts Quine's argument in his (1989).

[4] Goals or reasons are characterized intentionally, which prevents Dennett from offering an account that fully naturalizes intentional descriptions to non-intentional descriptions. Note that taking reasons as input will not in itself guarantee rationality. Without a rationality constraint, it is still possible to interpret a person as irrationally ignoring what they have reason to do, or intending to do what they know is counterproductive to the attainment of their goals.

constrain interpretations with a rationality requirement in order to get any predictive or interpretive power whatsoever.[5]

In addition to Quine himself, a major source of historical support for this sort of rationality assumption came from formal decision theoretic models of economic behavior. Standard decision theoretic or game theoretic models, such as those of Von Neumann and Morgenstern (1944), are descriptive of human behavior only if humans act rationally and in accord with the dictates of the theory. Davidson, for one, was heavily influenced by Ramsey's "Truth and Probability" (1931). Ramsey gives a procedure for representing an agent's utilities and degree of beliefs in any proposition when simply given that agent's preferences; he then gives a representation theorem proving that if the agent's preferences satisfy certain requirements, the agent's degrees of belief will be coherent. Davidson took radical interpretation to involve something like Ramsey's procedure, and saw close affinities between Ramsey's procedure and Quinean radical translation. He writes, "Quine's solution resembles Ramsey's, in principle if not in detail."' (Davidson 1990, 319).[6] Dennett was less directly influenced by formal modeling (or at least, there is less textual evidence for its influence). He does at one point write that taking up the Intentional Stance involves interpreting an agent to have beliefs and desires "roughly as Bayes would have them" (Dennett 1978, 307), but formal decision theory has seemed not to have been a major influence. Still, it is worth noting that indeterminacy-reducing rationality constraints found wider appeal than simply among philosophers allied with Quine. Rationality requirements are what result from demanding that mental state ascription involve a procedure akin to Ramsey's. Choosing a formal theory that guarantees the ascription of rational beliefs is much the same as adopting a rationality constraint.

The sorts of considerations just mentioned make it *seem* like we need rationality constraints to get interpretation off the ground. But the arguments are not decisive. To my mind, the arguments fail to satisfyingly answer the following two questions:

1. Must some constraint on interpretation be a *rationality* constraint?
2. Is a constraint on interpretation really required *at all*?

[5] As an aside: it is worth nothing that although the philosophers above are interpretivists – they hold that the *content* of our mental states is determined through a process of interpretation – the apparent need for a rationality constraint hits non-interpretivists as well. Suppose that a computer or a brain contains an inscription written in Mentalese. Is this particular Mentalese sentence in a "belief box"? Or is it in an "imagination box" and the agent irrationally acts as if her imaginations are beliefs? Non-interpretivists find themselves facing the same problems that interpretivists do: they seem to require a rationality constraint to appropriately ascribe *attitudes* to an agent. We need to be careful and distinguish theories of semantic content from theories of mental attitudes with those contents. (This fact is sometimes glossed over by non-interpretivists. For instance, Fodor doesn't recognize this in his response to Stich's Mrs. T thought experiment, in which a woman assents to the claim that McKinley was assassinated while also being unable to say anything else related to assassination. Does she believe that McKinley was assassinated? Fodor should, I think, say she does not. She has the concept ASSASSINATED (fixed by asymmetric dependence), but it languishes in her head without playing a role in any of her beliefs. But this is not Fodor's response (see Fodor 1987, 62).)

[6] See Rawling (2003) for more on Quine and Ramsey's influence on radical interpretation.

I plan to argue that we already have principles that constrain indeterminacy, and an additional rationality requirement is neither motivated nor desirable. However, in order to talk about the principles that "we already have," I first need to unravel a persnickety issue that all-too-often complicates conversations about rationality constraints and interpretivism.

5.3 Types of Ascription

The rationality assumption is a constraint on theory construction. What sort of theory – and whose theory – requires constraint?

There are (at least) two sorts to consider. Firstly, individual human agents ascribe mental states to other agents. This is often called 'mindreading' or 'mentalizing'. One popular account of mindreading holds that we interpret other people around us by fitting our observations of them to a tacit folk psychological theory. The fitting of such a theory might involve an assumption of rationality. Secondly, philosophers and psychologists ascribe mental states to others by building, and subsequently applying, mature theories of the mind. This sort of theory-construction, too, might demand rationality constraints. Let's call these types of ascription *individual ascription* and *scientific ascription*, respectively. They are distinguished by who it is that does the ascription: the first is employed by individuals in real-world situations, and the second is employed by scientists and philosophers in the development of theories. Either investigation can have a descriptive or normative focus. One might be interested in how individuals actually do go about mindreading, or one can make suggestions about how people ought to mindread. Similarly, one can describe how psychologists actually do build theories that attribute mental states to observed actors, or one can offer suggestions about how their theories could be improved. Investigations into individual ascription are traditionally descriptive; investigations into scientific ascription are traditionally normative.

In *Content and Consciousness*, Dennett is clear that his concern is mental ascription of the second type. The goal is to build a mature theory of intentionality and mental states, and it is permissible to deviate from the terms of "ordinary" mental ascription. For instance, he writes, "the centralist makes his initial characterization Intentional, describing the events to be related in law-like ways using either ordinary, or semi-ordinary, or *even entirely artificial* Intentional expressions" (Dennett 1969, 41–2, italics mine).

The ground shifted somewhat when Dennett developed the Intentional Stance. The Intentional Stance became a piece of *individual* ascription: interpretation was now spoken as something that we *all* naturally do.[7] It is, of course, a legitimate

[7] E.g. "According to Intentional Systems Theory, [questions about the conditions under which a thing can be truly said to have a mind] can best be answered by analyzing the logical presuppositions and methods of our attribution practices, when we adopt the intentional stance toward something" (Dennett 2009, 339).

hypothesis that mindreading works through an application of a tacit theory of mind. However, building a psychological theory and mindreading are two separate enterprises, subject to different demands. Speed of processing is a worry in mindreading, for instance; the psychologist in her lab is under less time pressure.

The two enterprises became conflated in the literature. In "Mid-Term Examination: Compare and Contrast" (1989), Dennett takes a tour of his various philosophy of mind contemporaries, writing that "two chief rival" principles of interpretation have emerged: Normative Principles and Projective Principles. Normative Principles constrain interpretation by ascribing propositional attitudes that a creature *ought* to have; Projective Principles attribute the propositional attitudes that one supposes one *would* have in that very scenario. Dennett counts himself, Lewis, and Davidson among defenders of Normative Principles, and affirms that it all arose from Quine. Something strange has gone on here, however, for Projective Principles, with their egocentric focus ("interpret others as believing what *you* would believe in their shoes"), can only be understood as constraining individual ascription. To cast them as a competitor to the Normative Principles espoused by Quine, Lewis, and the Dennett of'69, suggests that these authors present their Normative Principles as also governing individual ascription, but this was not the case. Dennett, after all, suggests that a mature Intentional Systems Theory might invoke entirely artificial intentional expressions, formerly unknown to folk psychology. He can't be giving a theory about how we actually individually mentalize.

Dennett puzzles over the fact that Quine's *Word and Object* contains the seeds of both Normative Principles and Projective Principles. He resolves the potential conflict between the principles by arguing that for Quine, it did not matter much which principle yielded the actual propositional attitudes: since mental talk is a dramatic idiom that we employ simply for practical purposes, we can afford whatever indeterminacy is yielded by having two separate methods of ascription (344). I endorse a different solution: Quine presented the Projective Principle as part of a theory about how individuals actually understand the statements of others, and the principle of charity as a part of a theory about how linguists ideally ought to understand the statements of others. There is no conflict between the two principles because they are enlisted for two different projects. It is entirely consistent to be a simulationist with respect to individual ascription without being a simulationist with respect to scientific ascription: that is, while also being an interpretivist about the metaphysics of belief.[8]

Sometimes skeptics of rationality constraints admit that there is a need for something *like* a rationality constraint in order to act as an heuristic that can be used in real-time cognizing. This is not an admission that should be made if one is trying to determine whether we ought to invoke a rationality constraint when interpreting

[8] Goldman (2006) charges Dennett and Davidson with occasionally taking their theory of mindreading to be identical with their theory of the metaphysics of mental states, and their commitments to the metaphysics of mental states leads them to reject simulationism (a theory of *individual* mental state ascription) right off the bat.

others through a psychological theory. For instance, Cherniak offers a *minimal rationality constraint* because human beings are in "the *finitary predicament* of having fixed limits on their cognitive capacities and the time available to them" (Cherniak 1986, 8).[9] Bortolotti endorses an *intelligibility requirement*: "intentional behavior must be intelligible or amenable to rationalization" (Bortolotti 2009, 100), but she suggests that we should consider the interpreter's assumptions about intelligibility to be "flexible and revisable heuristics, not constraints. They are supposed to guide the interpreter and help her to ascribe intentional states with determinate content to a variety of subjects in a variety of situations" (107). The rest of their work makes it clear that they are really concerned with scientific ascription and the metaphysics of belief, so it is odd for them to discuss time-sensitivity and other concerns that clearly belong to the domain of individual ascription.

Now that we've established that the main project in *Content and Consciousness* is one of scientific and not individual ascription, an argument against the need for a rationality assumption can present itself.

5.4 Undoing the Quinean Lineage

We left our discussion of the Quinean lineage on a cliffhanger. Does Dennett have a good answer to the following two questions?

1. Must some constraint on interpretation be a *rationality* constraint?
2. Is a constraint on interpretation really required *at all*?

These are best dealt with in turn. Firstly, note that if the sole goal is to reduce indeterminacy of mental state ascription, it is far from obvious that a rationality constraint is the only constraint or assumption that would accomplish the task. It is one viable option, but there are others. One way to see this is to consider the argument from Ramsey's representation theorem. Ramsey showed that, given a preference ordering with certain features, humans can be formally represented as having rational and coherent degrees of belief, but this means nothing in itself, for they can also be formally represented as *irrational*. Zynda (2000) has shown that for any preference ordering that allows one to be representable as having degrees of belief that obey the laws of probability, that same preference ordering allows one to be representable as having of degrees of belief that *don't* conform to the laws of probability. In order to establish that humans are rational, it is not enough to simply establish that humans are representable as having consistent and rational beliefs; there are other representations that say otherwise.

This is just to say that the data are indeterminate without interpretation. But what's important is that the representations that lead to *ir*rationality are well-behaved, which means that the representation that guarantees rationality is only

[9] See also (Dennett 1987, 98).

one of many. What gives that particular interpretation a place of pride? It can't *simply* be its ability to reduce indeterminacy, because all sorts of representations have that feature.

What's even worse is that it appears that models that *don't* preserve rationality can actually be more predictive and empirically adequate. Since the original suggestion that unelaborated Ramseyan decision theory could be used as an empirical model of actual human decision-making (Edwards 1954), the claim has been steadily attacked; psychologists and behavioral economists have developed competing accounts of decision-making and competing research programs. Why should we think that the best formal theory of mental state ascription should be contained within the set of formal theories that guarantee rational beliefs? There are other formal models: some posit mental states other than belief and desire (such as intention or emotion); some do not assume that our preference ordering is transitive; some allow for unsharp probability functions. Perhaps a model that does not guarantee rationality will do a better explanatory job.

How does Dennett respond to apparent breaches of rationality in everyday life? After all, taking up the Intentional Stance involves interpreting an agent as having coherent and rational degrees of belief, but people obviously don't act exactly like perfect Bayesian agents all the time. Dennett accepts this, but he maintains that this doesn't imply the surprising fact that no one is a believer. He has two responses. Stich (1981) calls these "the hard line" and "the soft line" on rationality constraints.

On the hard line, the Intentional Stance is useful because people closely *approximate* rational agents. The property of *being a believer* is somewhat like the property *being a rabbit-shaped image* (Dennett 1991). Some images only vaguely resemble rabbits; others might be smudgy or pixellated. As the fidelity of the image goes down and noise is introduced, it becomes less of a perfect rabbit image, but it still has the same basic pattern that a perfect image would. People are, metaphorically, "smudgy images" of fully rational Bayesian agents. To ask whether a schizophrenic *really believes* that someone else has inserted thoughts into her head is akin to asking whether a shape in a smudgy picture *really is* rabbit-shaped. It's like a rabbit image in some respects but not in others – its status is indeterminate and there is no fact of the matter.[10] On the soft line, the form of rationality that is assumed by the rationality constraint demands less than perfect Bayesian consistency and coherence. For instance, it becomes rational to "satisfice" (to use Herb Simon's term). In Dennett's (1987) response to Stich, he adopts both strategies. So, upon seeing someone apparently act irrationally, we can either understand them by seeing them as approximating a rational being (and deviating slightly); or we can understand them as actually being rational according to some different standard.

The third strategy that Dennett does not adopt, of course, is just to give up on the assumption of rationality. Consider the hard line strategy: taking up the Intentional Stance just is representing or modeling an individual as having coherent degrees of

[10] Whether this account demands ontic vagueness is an open question; accounts of indeterminacy that are purely linguistic don't seem to capture what Dennett has in mind.

belief, and that people resemble perfect Bayesian creatures to some extent. Now, however, recall that there are other cognitive models waiting in the wings. Consider a new stance – a "schmIntentional Stance" – according to which individuals are represented or modeled by some different formal structure. Perhaps this representation assumes that we are predictably irrational whenever we reason about certain topics: perhaps it models us as systematically overestimating (or underestimating) the likelihood of events that would be bad (or good) for us. Perhaps it models us as having intransitive preferences (is this ruled out by the Intentional Stance?). Perhaps it posits various mental states that the Intentional Stance does not and to which it is difficult to apply folk notions of rationality. These models might very well do a better job of predicting human behavior.

Dennett often speaks as if, when an individual can't profitably be understood on the Intentional Stance, we need to plunge down to the design stance or physical stance. But why? Why not look for models of human psychology that are similar to (but distinct from) the one that you get by applying the Intentional Stance? We should not conclude that humans must be interpreted according to some psychological model just because they *can* be successfully interpreted according to that psychological model. There might be a more predictive model out there. We can update the Intentional Stance. That's what we do whenever cognitive psychology discovers new mental states.[11] The rationality constraint pushes us toward one of many possible interpretations of behavior. But in many cases, this means it pushes us away from interpretations that would be comprehensible and yield predictions.

Considering the second question (is a constraint on interpretation really required at all?) lets us go even further in questioning the need for the rationality constraint in theory building. Intentional Systems Theory models the mind, and we already have various maxims that regulate our theory construction. We do not need an additional constraint to reduce indeterminacy. Consider the observational data we acquire when building theories of physics. We take measurements, we construct atom chambers and run experiments, we build instruments, etc. The actual theory we construct is underdetermined by this data. We posit atoms and subatomic particles, but an evil demon manipulating all our observations will fit the data equally well. What prevents us from inviting in rampant indeterminacy in our commitments are certain epistemic principles or scientific virtues that guide our theorizing: simplicity, conservatism, scope, fecundity, and so on. If rationality were a constraint on mental state ascription, it would be serving as another such scientific virtue. It would be another such principle that we would use to reduce indeterminacy.[12]

[11] Two responses that Dennett might make here are responses that I will deal with in my discussion of the Rylean lineage in the next section. (A preview: they are that rationality is guaranteed by natural selection, so as evolved agents we are forced to make that assumption; and that the "schmintentional stance" is a discussion-changer: its declarations would be so remote from our ordinary mentalistic vocabulary that we could not properly call its posited states 'beliefs' and 'desires'.) All I am trying to establish here is that the need to reduce indeterminacy in ascribing mental states to our friend who is huddling under a tree in the thunderstorm does not *in itself* necessitate a *rationality* constraint, which is an argument that Dennett and others seem to make at times.

[12] The virtues listed above are some of those listed by Quine and Ullian (1970).

There is something very odd about this principle in that it is relative to a particular special science: psychology (and perhaps economics). No other sciences seem to require an additional virtue. This should make us suspicious of its necessity. In fact, it's not clear that the other virtues cannot do the job we want rationality to do. The problem is that ascribing irrationality to a subject is wholly uninformative. To interpret a person huddling under a tree as irrationally *not* intending to do so doesn't predict much else about them. Why will they say they are huddling under the tree? Would they huddle if they didn't want to huddle? The theory doesn't say. Attributing rationality and the intention to stay dry under the tree, on the other hand, offers up wealth of other information about their potential behaviors in various situations. This is very close to Quine's explanation: we attribute rational beliefs because we get predictive power by doing so. But the work here is not being done by an assumption that theories that postulate rationality are better: it's done by the assumption that theories with more predictive power are better.

Let's consider a version of Quine's field linguist, who, seeing speakers assent to an instance of ⌜p ka q⌝ when they assent to p and dissent from q, prefers to translate 'ka' as 'or' rather than as 'and'. Why should he prefer this hypothesis? On Quine's account, it would be because translating it as 'and' violates a requirement of rationality. Can we get the same result without appealing to such a constraint?

If we posit that 'ka' means 'or' and that the speaker is rational, we end up making all sorts of other predictions. For one, we anticipate that he will accept *any* instance of ⌜p ka q⌝ for any p or q. The hypothesis systematizes a whole lot of possible data about the speaker's dispositions. On the other hand, if we posit that 'ka' means 'and' and that the speaker is irrational, and if we don't have a theory about how the speaker is irrational, then we can't predict much else. We don't know how the speaker will respond to pretty much any instance of ⌜p ka q⌝. Thus, whatever scientific virtues push one to prefer simple and predictive systematizations of the facts will suggest a theory in which the agent is rational. We have a theory that tells us what can be expected when an agent is rational; claiming that an agent is irrational jettisons all those predictions. Consider: if a psychotic patient has the delusion that he is Napoleon, we can predict at least *some* things about his behavior (such as the fact that he will say that he is Napoleon). If we simply say that the agent has the irrational belief that he is Napoleon, then we should be hesitant to draw very few conclusions at all. We lose information. It's the epistemic virtues of predictiveness and systematization that keep us from attributing irrational beliefs, not a distinct rationality requirement.

Note that a rationality requirement can't be straightforwardly derived from predictiveness and systematization, because if we have a theory of how irrational actors will act, the most predictive, systematized, and empirically adequate theory might be one that interprets actors as irrational. Suppose we do come up with a theory of the speaker's irrationality. Suppose we notice that the speaker's behavior is altogether rationally consonant with 'ka' meaning 'and', but that the speaker tends to make errors when forming complex statements involving some particular sentence. We might then hypothesize that it's difficult for the speaker to reason about that sentence – maybe it introduces a lot of cognitive load. This hypothesis once again

lets us systematize the speaker's dispositions to assent: we expect that the speaker will assent to ⌜*p* ka *q*⌝ iff he assents to *p* and to *q* unless either *p* or *q* is one of the sentences identified to introduce cognitive load, in which case he dissents from the whole thing. The epistemic virtues should cause us to prefer a theory in which the agent is irrational if and only if the various irrational inferences the agent is disposed to make are patterned instead of piecemeal, and can be systematized into a theory of the agent's cognitive system that yields the patterns of irrationality.[13]

Sometimes it is argued that we should prefer models that assume humans to be rational because they are simpler than other models. Sober (1978) argues for this. Heil writes that it is useful to regard charity "as parsimony applied in the mental realm" (1994, 120). These sorts of warnings do not put any additional strictures on psychological theory-construction. We already have parsimony in the mental realm: it goes by the name 'parsimony'. Moreover, we don't want to *equate* charity with parsimony in the mental realm, because we cannot guarantee from the outset that the most parsimonious (or otherwise virtuous) theory will be the one with the result that people are rational. Thagard and Nisbett (1983) respond to Sober by presenting psychological evidence that people apparently behave irrationally in various domains; explaining away these apparent irrationalities will probably be less parsimonious than just positing a streamlined model that predicts irrationality in these domains. They present a moderate version of a principle of charity: "Do not judge people to be irrational unless you have an empirically justified account of what they are doing when they violate normative standards." This is not a bad general methodological principle (in psychology's current state). "Do not judge entities to be *X* unless you have an empirically justified account of how they can be *X*" is a reasonable scientific proscription whether building a theory of the mind or of tornados or of ducks. We have a simple theory of rational agents, we have some reason to think that rationality would be evolutionarily adaptive, and agents do seem to often be rational, so the rationality hypothesis is a reasonable default hypothesis. This is a far cry from saying that it is a constraint that cannot be overturned. If we find what appears to be systematic irrationality in people, then we needn't torture ourselves trying to interpret them as *really* being rational. We should just admit that the rationality hypothesis is no longer supported and then give it the boot.

I hope to have successfully challenged arguments that a rationality assumption is needed to do indeterminacy-reducing work because the work cannot be done by more standard scientific norms. If we interpret agents as rational because we are led to do so by scientific norms of predictiveness, systematization, and empirical adequacy, then rationality need not be a *constraint* on interpretation, nor need it play

[13] This account has affinities with Cherniak (1986), who argues that we don't only holistically ascribe mental states and language meanings: we holistically attribute mental states and the meanings of our words along with a theory of the agent's cognitive system. This is in order to account for the ascription of irrational inferences that are the product of memory constraints and computational difficulty or intractability. Cherniak, however, takes his project to be one of individual psychological ascription rather than the ascription of our best scientific theory, and still thinks that a constraint of minimal rationality is needed on top of all this.

any sort of role on the *input* side of psychological theory-building. It could be an *outcome*, or *finding*, of (current) psychology that agents are (largely) rational.

Consider, similarly, that it is an outcome of physics that there exist particles that have negative charge. We do not need to mandate anything like a negative-charge constraint on physics. It might be that a psychology that postulates rationality – or a physics that postulates electrons – makes better predictions, but this would only be contingently true, and not because of any necessary restrictions on theory-construction.[14]

This deals with the motivations for a rationality assumption that stem from Quine. The need to reduce indeterminacy in order to get psychology off the ground does not require anything that is unknown to the other sciences.

5.5 The Rylean Lineage

When Dennett entered his graduate studies at Oxford, ordinary language philosophy's Last Days of Empire were in full effect. When describing his time there, he emphasizes the atmosphere of disdain toward science that he experienced.[15] Attempted naturalizations of the mind were considered vulgar. Dennett broke from the tribe and auto-didactically immersed himself in psychology, neuroscience, and computer engineering, but even in so doing, he was moved by certain arguments of the anti-naturalists around him. The two books on intentionality that had the largest influence on him were Anscombe's *Intention* (1957) and Taylor's *The Explanation of Behaviour* (1964) (Dennett 1996). *Content and Consciousness* is studded with references to the two.

Dennett saw, in their anti-reductionist arguments, a recapitulation of Quine's arguments for the holistic nature and hence irreducibility of intentional discourse. While these arguments drove Quine to disparage mind-talk, in places advocating its dispensability and in other places treating it as pragmatic crutch that deserved scant respect, mind-talk was dead serious for the Oxbridgians. Their ordinary language

[14] There is a sense in which physicists do have something like a negative-charge constraint. If some feature of a theory has been pretty much conclusively established, scientists are free to dismiss theories that claim otherwise. Established physicists receive letters from all sorts of cranks who claim to have "disproved relativity," and these crackpots are rightfully ignored. The constraint in this case isn't a restriction on theory-building, but an heuristic used to guide the theorist's attention away from likely falsehoods. This does not always seem to be what Dennett has in mind when he speaks of a rationality assumption (for instance, when he argues that prediction could not get off the ground at all if it were not for an assumption of rationality).

Please note that in drawing the comparison between mental states and electrons, I do not mean to suggest that both are what Dennett calls 'illata' and that mental states are not personal-level states. Mental states are abstracta. Nonetheless, my comparison is apt because abstracta and illata are both potential objects of empirical investigation. Determining whether an agent has any particular personal-level state is an empirical matter. As I've been arguing, there's no compelling reason to think that empirical investigation into these sorts of states needs to involve a special sort of rationality assumption. (Note also that the positing of non-mental abstracta, such as centers of gravity, does not involve a rationality assumption.)

[15] Dennett (1996, 2012)

analyses of mentalistic terms proved attractive to Dennett. "The philosophy of mind initiated by Ryle and Wittgenstein is in large measure an analysis of the concepts we use at the personal level" (Dennett 1969, 95) and their sensitivity to the features of these concepts was crucial in the development of Dennett's theories. Ryle's notions of separate "logical categories" and the category mistakes that result from illicit admixtures of terminology from two different categories, foreshadows Dennett's construction of an Intentional Stance distinct from the Physical and Design Stance.

It's a conceptual analysis of mentalistic vocabulary that leads Dennett to his second version of the rationality requirement: it arises from the supposition that the meanings of mentalistic terms are fixed by their holistic connections with other mentalistic terms. Dennett points out the "conception causes pregnancy" is analytically true, because an event only counts as a conception if it causes pregnancy. Asking why a conception led to a pregnancy (rather than some other state) while using those terms is silly and unnecessary: the occurrence of the pregnancy is already entailed by there being a conception.[16] Dennett thinks mental vocabulary works in the same way. He writes,

> In Intentional explanation, on the other hand, the sequences of events are so characterized that the occurrence of a particular consequent action is explained by the occurrence of a particular antecedent, say a perception or a belief or intention, and there is no room for the question of why this consequent should follow this antecedent, and hence no room for any general law 'explaining' this sequence. For example, having said that my intention to leave was followed by my walking to the door, there is no room for the question: why should that result (as opposed to, say, opening my mouth or raising my arm) follow my intention to leave. The 'covering law' to the effect that all intentions to leave are followed by walking to the door is silly and unnecessary; the occurrence of my walking to the door has already been explained by citing my antecedent intention. In this way Intentional explanations assume the environmental appropriateness of the connections between antecedent and consequent (Dennett 1969, 37).

If you have a conception, then you certainly have a pregnancy, and this is guaranteed by the meanings of the terms. Similarly, if you have an intention to leave a room, then *ceteris paribus* and barring other mental states that would intervene, you'll move to leave the room; this is guaranteed by the meaning of the term "intention." If you acted irrationally instead of appropriately – if you opened your mouth or raised your arm – then you couldn't have had the intention in the first place. Whatever you had, it wasn't an intention to leave the room. To think otherwise would be to misuse the (ordinary language) word. For years, Dennett has presented various thought experiments to prompt the intuition that when rationality breaks down, we very much balk at ascribing beliefs to an agent: we don't know what to say. Let's draw another analogy with theories in physics. To be an electron, a subatomic particle must have certain features. It must have negative charge; it must have intrinsic angular momentum of 1/2, and so on. If some particle under observation does not display these properties, it isn't an electron. Similarly, for a

[16] This isn't actually true: 'in vitro conception' is in common use and not a contradiction in terms. (Admittedly, this is a cheap shot, as the technique was invented after the publication of Dennett's book. But this does go to show just how difficult it is to find analyticities.)

mental state to play a belief-role, it might need to stand in rational relations with other mental states.

The claim that beliefs are constitutively rational can be read in two ways, and they are not always distinguished. Firstly, one might mean that the *process of interpretation* involved in mental state ascription is constrained by a principle that guarantees the rationality of the interpreted agent. Alternately, one might mean that it is characteristic of the *functional role* of belief that it is rational: if a mental state doesn't play the role of a rationally formed and maintained belief that motivates behavior in a rational way, then it doesn't play the role of a belief. One way to think of this is that on the first thesis, rationality is a condition on the interpretive process. On the second, rationality is a mandated feature of the outputs of the process of interpretation. The first sort of rationality constraint is Quinean, and the second is Rylean.

5.6 Undoing the Rylean Lineage

Suppose we grant that the meaning of 'intention' in everyday folk language does, in fact, imply that individuals act appropriately on their intentions. Why must Intentional Systems Theory hang onto the meanings given to us by folk theory unaltered?

I am not driving toward eliminativism; I'm not suggesting that we replace belief-desire psychology with something radically different. My goal is less contentious. I am simply pointing out that once we separate the project of explaining individual ascription from the project of scientific ascription, we should recognize that it is perfectly admissible to make modifications to folk theory if it gains us predictive and explanatory power. Dennett himself does this: recall his claim that a successful Intentional Systems Theory might describe mental events using "ordinary, or semi-ordinary, or even entirely artificial Intentional expressions" (42). In a chapter of *Brainstorms*, he introduces *opinion* as a novel sort of propositional attitude, and touts it as "a *reform* of our ordinary concept of belief" (Dennett 1978, xxii). It's true that opinions were introduced in order to *preserve* rationality: when an agent says P, and it would render him irrational were he to believe P, we can say instead that he merely has the opinion that P. But the damage is done: folk psychology is up for amendment if in the service of constructing a better theory. Why not think that the features of folk explanation that presume appropriateness are similarly up for grabs? The simple fact that folk psychological terms assume rational relations does not in itself say anything about whether the terms of a mature theory ought to similarly assume rational relations. We might find it best, at some point, to adopt the schmIntentional Stance instead.

Thus, even if the terms of folk psychology analytically ensure the rationality of any agent they are attributed to, this would not, in itself, restrict future theory-building. We regiment folk terms all the time in all the sciences; why are these terms sacrosanct? One might think that it is just central to the meaning of 'intention' that

it implies rational relations to other mental states. If this were true, then amending intention to be intention-like really would be considered a version of eliminativism. To my ears, this sounds like a semantic dispute over what states merit the name 'intention'.[17] Arguing over whether an irrational intention is an intention does not sound much different to me than arguing whether a wrap is a sandwich.

5.7 Preserving Intentional Systems Theory

I believe that electrons exist; I also believe that they have negative charge. I do not think that we are in much danger of a future generation discovering that there are no electrons. Imagine, then, that I encountered someone who believed in a *negative charge requirement* on the construction of physical theories. He presents me with the following two arguments: firstly, it is an indeterminacy-reducing constraint on theory-building that physics posit a subatomic particle with negative charge; secondly, it is a central part of our current concept that electrons have negative charge, and the concept is too useful and predictive to ever want to give up.

This "negative charge requirement" is wholly unneeded. The existence of subatomic particles with negative charge was a discovery, not an a priori condition on scientific inquiry. If anything, having this sort of requirement stunts potential scientific investigation: on the remote chance that a there is a more virtuous theory waiting in the wings that would dispense with negatively charged particles, the requirement would have us dismiss it out of hand.

Rationality requirements are in much the same boat. I think we have good reason to suppose that our mental states are (mostly) appropriate and rational, but this is a well-established *discovery*, not a condition on all future psychologizing. One argument Dennett makes for rationality requirements which I haven't mentioned until now appeals to natural selection. Having mostly true and rationally-formed beliefs is conducive to fitness, so we should expect our attitudes to be rational. I think this is a good argument.[18] However, I have a hard time seeing how it could act as an argument for a rationality constraint on mental ascription. Our evolved nature is a source of evidence that should cause us to *expect* our attitudes to be rational, but this evidence could plainly be defeated by other sources of evidence. Perhaps we will find that it was fitness-conducive in our primitive niche for us to be overly credulous or skeptical, or to be subservient to authority – biases that are harmful in our current environment. The various cognitive biases that psychologists discover do provide *some* evidence that we are irrational; they can't always be written off (as performance errors or whatnot) in allegiance to an unshatterable rationality assumption.

Does it drastically damage Intentional Systems Theory if we scrap the rationality requirement and simply replace it with a claim that we have a lot of *good evidence* that our mental states are rationally arranged? I can't see that it does. Dennett still

[17] See Stich (1996) for more on these tricky semantic issues.

[18] Pace Stich (1985).

has the ability to claim that mentalistic vocabulary is holistic and irreducible. He can still hold that we are goal-oriented, sensitive to reasons, and have his version of free will worth wanting. He can also still hold that *individual* ascription involves the use of a rationality assumption: it's just one of many heuristics that we might use in order to enable real-time mentalizing.

On the other hand, it might be thought that I haven't made much of a change. Is progress really made by saying that, instead of there being a *rationality assumption* on ascription, it is epistemically *safe to assume* that minds are mostly rational? Yes, I think so. It's a small point, but an important one: by removing rationality as a condition on all theories of mind, we remove a barrier that could influence or stand in the way of creative theory construction. Philosophy of mind has of late been replete with proposals for new attitude types much like Dennett's own opinions, from Gendler's aliefs (2008) to Egan's bimagination (2009) to Schwitzgebel's in-between beliefs (2001) to Frankish's superbeliefs (2004). These are exciting and creative times. I worry that the rationality constraint is too aprioristic, and it will dissuade us from imaginative reform of our cognitive theories.

References

Anscombe, E. (1957). *Intention*. Cambridge, MA: Harvard University Press.
Bortolotti, L. (2009). *Delusions and other irrational beliefs*. Oxford: Oxford University Press.
Cherniak, C. (1986). *Minimal rationality*. Cambridge, MA: MIT Press.
Davidson, D. (1982). Psychology as philosophy. In *Essays on actions and events* (pp. 229–238). Oxford: Oxford University Press.
Davidson, D. (1990). The structure and content of truth. *The Journal of Philosophy, 87*(6), 279–328.
Dennett, D. (1969). *Content and consciousness*. London: Routledge and Kegan Paul.
Dennett, D. (1978). *Brainstorms: Philosophical essays on mind and psychology*. Cambridge: Bradford Books.
Dennett, D. (1987). Making sense of ourselves. In *The Intentional Stance*. Cambridge, MA: MIT Press.
Dennett, D. (1989). Mid-term examination: Compare and contrast. In *The Intentional Stance*. Cambridge, MA: MIT Press.
Dennett, D. (1991). Real patterns. *Journal of Philosophy, 88*(1), 27–51.
Dennett, D. (1996). An overview of my work. In K. Ouyang & S. Fuller (Eds.), *Contemporary British and American philosophy*. New York: Nova Scientific Publishers.
Dennett, D. (2009). Intentional systems theory. In B. P. McLaughlin (Ed.), *The Oxford handbook of philosophy of mind*, chapter 19 (pp. 339–350). Oxford: Oxford University Press.
Dennett, D. (2012). Daniel Dennett: Autobiography (part 1). *Philosophy now*.
Edwards, W. (1954). The theory of decision making. *Psychological Bulletin, 51*(4), 380.
Egan, A. (2009). Imagination, delusion, and self-deception. In T. Bayne & J. Fernandez (Eds.), *Delusion and self-deception: Affective and motivational influences on belief formation* (pp. 263–280). New York: Psychology Press.
Fodor, J. A. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. Cambridge, MA: MIT Press.
Frankish, K. (2004). *Mind and supermind*. Cambridge: Cambridge University Press.
Gendler, T. S. (2008). Alief and belief. *Journal of Philosophy, 105*(10), 634–663.

Goldman, A. (2006). *Simulating minds*. Oxford: Oxford University Press.
Heil, J. (1994). Going to pieces. In G. Graham, & G. L. Stephens (Eds.), *Philosophical psychopathology* (pp. 111–134). Cambridge, MA: MIT Press.
Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, MA: Cambridge University Press.
Lewis, D. (1974). Radical interpretation. *Synthese*, 27(July–August), 331–344.
Quine, W. V. (1960a). Carnap and logical truth. *Synthese, 12*(4), 350–374.
Quine, W. V. (1960b). *Word and object*. Cambridge, MA: MIT Press.
Quine, W. V., & Ullian, J. S. (1970). *The web of belief*. New York: Random House.
Ramsey, F. P. (1931). Truth and probability. In R. B. Braithwaite (Ed.), *The foundations of mathematics and other logical essays* (pp. 156–198). London: Routledge and Kegan Paul.
Rawling, P. (2003). Radical interpretation. In K. Ludwig (Ed.), *Donald Davidson, contemporary philosophers in focus* (pp. 85–112). New York: Cambridge University Press.
Schwitzgebel, E. (2001). In-between believing. *Philosophical Quarterly, 51*, 76–82.
Sober, E. (1978). Psychologism. *Journal for the Theory of Social Behavior, 8*, 165–191.
Stich, S. (1981). Dennett on intentional systems. *Philosophical Topics, 12*(1), 39–69.
Stich, S. P. (1985). Could man be an irrational animal? *Synthese, 64*(1), 115–135.
Stich, S. P. (1996). Deconstructing the mind. In *Deconstructing the Mind* (pp. 3–90). Oxford: Oxford University Press.
Taylor, C. (1964). *The explanation of behaviour*. London: Routledge and Kegan Paul.
Thagard, P., & Nisbett, R. E. (1983). Rationality and charity. *Philosophy of Science, 50*(2), 250–267.
von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton: Princeton University Press.
Zynda, L. (2000). Representation theorems and realism about degrees of belief. *Philosophy of Science, 67*, 45–69.