Delusions and Madmen:
Against Rationality Constraints on Belief

Declan Smithies, Preston Lennon, and Richard Samuels

How does rationality constrain belief? We assume that there are norms of rationality that apply to anyone who has beliefs. If you have beliefs at all, then you ought to hold them rationally. Indeed, quite plausibly, this normative claim has the status of a conceptual truth.

Some philosophers go further than this. They combine this normative claim with the descriptive claim that if you have beliefs, then you must in fact hold them rationally. On this view, it is a conceptual truth that anyone who has beliefs conforms to some extent with the norms of rationality that apply to them. Needless to say, not all believers are perfectly rational. Nevertheless, the claim is that all believers are at least imperfectly rational.

Our goal in this paper is to evaluate – and ultimately to reject – the thesis that our concept of belief imposes any such constraint on the rationality of our beliefs. To a first approximation, the target of our critique can be stated as follows:

> *The Rationality Constraint*: It is a conceptual truth that anyone who has beliefs is rational enough to meet some minimal standard.

In contrast, we argue that our concept of belief imposes no limits on how much irrationality is compatible with having beliefs. Believers can be as bizarrely irrational as you like.

The Rationality Constraint has been a central theme in philosophy of mind over the last fifty years and it remains popular today. Prominent adherents include Davidson (1970), Dennett (1971), Lewis (1983), McDowell (1985), Loar (1986), Cherniak (1986), Stich (1990), Child (1994), Wedgwood (2007), Pautz (2013), and Williams (2020). In this paper, we focus on David Lewis's defense of the Rationality Constraint, since it has been so influential, although our discussion has more general ramifications for many others too. We begin in §1 by explaining how Lewis uses his distinctive brand of analytic functionalism to argue for the Rationality Constraint.

Despite its popularity, the Rationality Constraint has also been subject to intense criticism. Some philosophers have argued that it is undermined by empirical evidence of human irrationality from the psychology of reasoning (Stich 1985; Stein 1996) or the psychopathology of delusion (Bortolotti 2010). However, we argue in §2 & §3 that these empirical challenges are inconclusive. While they succeed in undermining more demanding versions of the Rationality Constraint that require perfect rationality, they pose no significant challenge to Lewis' version, which demands only minimal standards of rationality.

Instead, we develop a conceivability argument against the Rationality Constraint, which we call the *Continuity Argument*. We argue in §4 that human irrationality is continuous with more extreme forms of irrationality that are conceivable though non-actual. Ironically, our main weapon against Lewis is his own invention, the *madman*. We argue that it's conceivable that there are communities of Lewisian madmen whose beliefs are not even minimally rational. By excluding such cases, the Rationality Constraint confronts a version of the problem of chauvinism that plagues all functionalist theories of mind.

Finally, in §5, we diagnose where Lewis's argument for the Rationality Constraint goes awry. If mad belief is conceivable, as we contend, then his analytic functionalism must be abandoned. To avoid the problem of chauvinism, we must reject Lewis's thesis that our concept of belief is implicitly defined by its role in a theory – namely, folk psychology – that we use in predicting and explaining behavior. Instead, we should recognize that our understanding of the concept of belief has a first-person dimension, as well as a third-person dimension, since we experience our beliefs as feelings of conviction that we can know through introspection alone.

1. Lewis's Argument for the Rationality Constraint

For Lewis, the Rationality Constraint is not an unargued premise, but is rather a conclusion drawn from his analytic functionalism. In broad outline, his argumentative strategy is to derive the Rationality Constraint from a general thesis about the meanings of theoretical terms together with more specific claims about folk psychology. This section reconstructs Lewis's argument with the aim of critiquing it in §5.

Lewis's argument proceeds from a semantic thesis about the meanings of *theoretical terms*:

> (L1) Theoretical terms are functional terms that are implicitly defined by the causal roles specified in some associated theory.

What does it mean, for example, to say that something is a *gluon*? According to Lewis, to be a gluon is to play the *gluon-role* – that is, the causal role associated with gluons in particle physics. Moreover, the point holds more generally, since Lewis intends his semantic thesis to apply to all theoretical terms.

This semantic thesis grounds an analytic/synthetic distinction: it divides the content of a theory into an empirical component and a definitional component. It is a *synthetic* truth that gluons exist because there is no conceptual guarantee that anything plays the gluon-role. Since the meaning of 'gluon' is implicitly defined by its theoretical role, however, it is an *analytic* truth that if gluons exist, then they play the gluon role. More generally:

> (L2) For any theoretical term 't' of a theory T, it is an analytic truth that if t exists, then t plays the causal role specified by T.

How does Lewis's semantic thesis bear on the concept of belief? One part of the answer is that Lewis (1999: 298) endorses the so-called *theory-theory* of mindreading. On this view, our capacity for predicting and explaining behavior is explained by our implicit knowledge of a psychological theory comprising causal generalizations about how people are disposed to behave on the basis of their beliefs and desires. Lewis calls this theory *folk psychology*:

> (L3) We all have implicit knowledge of a psychological theory – folk psychology – that we use in predicting and explaining behavior.

The other part of the answer is that Lewis applies his semantic thesis to folk psychology. In particular, he claims that all the psychological terms in this theory, such as 'belief' and 'desire', are theoretical terms that are implicitly defined by their role in the theory:

> (L4) All folk-psychological terms, including 'belief', are theoretical terms of folk psychology.

From these premises, we can infer the following intermediate conclusions:

> (C1) All folk-psychological terms, including 'belief', are implicitly defined by the causal roles specified in folk psychology.
>
> (C2) It is an analytic truth that if anyone has beliefs, then their beliefs play the causal role specified in folk psychology.

We are not yet in a position to derive the Rationality Constraint. To see how Lewis reaches this conclusion, we need one more premise regarding the content of folk psychology:

> (L5) Folk psychology specifies the causal role of belief in such a way that anyone with beliefs is at least minimally rational.

Folk psychology is a *causal theory*: it specifies how beliefs, desires and other mental states are caused by sensory inputs, how they cause behavioral outputs, and how they causally interact with each other. Unlike many other causal theories, however, folk psychology has *normative* implications: it specifies causal roles for belief and desire that are at least minimally rational. More specifically, it implies that our actions are rationally based on our beliefs and desires, our beliefs are rationally based on perception and reasoning, and our desires are rationally based on beliefs about value. As Lewis writes: "Folk psychology says that we make sense. It credits us with a modicum of rationality in our acting, believing, and desiring" (1999: 320).

With this additional premise in play, we may now draw the further conclusion:

> (C3) It is an analytic truth that anyone who has beliefs is at least minimally rational.

And this is just what the Rationality Constraint maintains. As we'll see, Lewis qualifies the Rationality Constraint in later work by applying it to communities, rather than individuals, but we'll revisit this complication in §4.3. We now turn to our evaluation of the case against the Rationality Constraint in §§2-4 before revisiting Lewis's argument in §5.

2. The Psychology of Reasoning

Some have challenged the Rationality Constraint by appealing to empirical evidence of human irrationality from the psychology of reasoning (Stich 1985; Stein 1996). Since the late 1960s, a large and growing body of experimental results indicates that we routinely reason in ways that violate even the most basic principles of logic and probability theory. On the face of it, these results pose a serious challenge to the Rationality Constraint. After all, rationality is traditionally thought to require (among other things) reasoning in ways that respect formal principles of logic and probability theory. This is what Edward Stein calls the Standard Picture: "to be rational is to reason in accordance with principles of reasoning that are based on rules of logic, probability theory, and so forth" (1996: 4).

While some reject the Standard Picture (e.g. Gigerenzer 2006), it is widely assumed by friends and foes of the Rationality Constraint alike. Lewis is no exception: indeed, much of his work in probability theory and decision theory is devoted to defending the Standard Picture. Nevertheless, we'll argue that the empirical results from the psychology of reasoning can be reconciled with the Rationality Constraint without abandoning the Standard Picture.

2.1. Some Experimental Results

The relevant literature from psychology is extensive and well-known. Nonetheless, as a reminder, we start by briefly describing two classic findings from this much larger literature.

First, Peter Wason (1966) found that normal human subjects routinely fail the *selection task*. Subjects are presented with four cards, each of which has a letter on one side and a number on the other. Here is one example:

A	B	1	2

The task is to evaluate the truth of a simple conditional statement, for example: "If a card has an 'A' on one side, then it has a '1' on the other." Subjects typically recognize that they need to turn over the 'A' card, but most fail to realize that they need to turn over the '2' card, and many wrongly think they need to turn over the '1' card instead. These results demonstrate that we have difficulty reasoning with conditionals: we find it easier to reason in accordance with modus ponens than modus tollens, and we show some tendency to reason fallaciously by affirming the consequent. This is evidence that our reasoning fails to respect even the simplest deductively valid forms of logical argument.

Second, Amos Tversky and Daniel Kahneman (1982) found that normal human subjects routinely commit *the conjunction fallacy*. Subjects were presented with the following vignette:

> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

When asked to rank various statements in order of probability, most subjects judged the conjunction (a) that Lisa is a bank teller and is active in the feminist movement, to be more probable than its first conjunct (b) that Lisa is a bank teller. And yet it is a basic theorem of the probability calculus that the probability of a conjunction cannot be greater than the probability of either of its individual conjuncts. This is evidence that our probabilistic reasoning fails to conform to even the simplest theorems of the probability calculus.

These experimental findings are extremely robust: they have been replicated many times since they were originally discovered. Moreover, this is just the tip of an exceedingly large iceberg. There are many other well-documented examples of human judgment and decision-making that systematically violate familiar canons of rationality, including base-rate neglect, belief perseverance, the gambler's fallacy, and more (see Baron 2008; Kahneman 2013; Pohl 2017).

2.2. Interpreting the Experimental Results

What do these empirical results tell us about the extent of human rationality? This is a complex issue, but we can draw five lessons that should be largely uncontentious.

First, it is old news that human reasoning is not always rational. Everyone knows that we sometimes make errors in reasoning – for example, when we are drunk, tired, or emotional. This knowledge is implicit in our folk psychology, which we use in predicting and explaining people's behavior. Even if folk psychology operates on the default assumption that people are minimally rational, it can still acknowledge the possibility of rational mistakes.

Second, the psychology of reasoning is concerned to ascertain whether all these rational mistakes can be explained as errors in *performance* or whether some of them reflect instead an underlying deficiency in our reasoning *competence*. According to the *rational competence view*, there is no rational defect built into our reasoning competence, although it generates rational errors because of limitations on such resources as time, memory, and attention. According to the *competence error hypothesis*, in contrast, our errors in reasoning sometimes manifest rational defects that are built into our reasoning competence itself.

Third, while some philosophers have endorsed the rational competence view (e.g. Cohen 1981), we are hard pressed to think of any current research program in the psychology of reasoning that does not accept some version of the competence error hypothesis. Given that standards of rationality are derived from logic and probability theory, as we assume here, no current empirical approach maintains that our underlying reasoning competence is precisely aligned with such standards. And this is because we violate such norms in systematic respects that prove hard to capture unless we specify the underlying competence in ways that deviate from rational norms. There is no reasonable prospect for defending the rational competence view except by challenging the Standard Picture of rationality (see, e.g., Gigerenzer 2006).

Fourth, while most contemporary research programs countenance competence errors, this is not to suggest that there are no disagreements regarding the *extent* to which our reasoning deviates from familiar canons of rationality. On the contrary, as we will explain, empirical theories of reasoning vary markedly in this regard. Some are more pessimistic than others.

Fifth – and most importantly – despite their differences, all extant views in the psychology of reasoning are distinctly *mixed* in their assessments of our underlying reasoning competence. Rhetorical flourishes aside, they all presuppose that while we make systematic errors in reasoning, we frequently reason in ways that are normatively *appropriate*. Indeed, they all presume that our competence, despite generating errors, is at least approximately rational in the sense that it reliably conforms to norms of rationality across a fairly wide range of cases.

To illustrate the extent of this consensus, let's contrast two views of human rationality that diverge in their assessment of the extent of human rationality. On the one hand, we find a rather pessimistic view of human rationality associated with psychologists in the *heuristics and biases* tradition, such as Daniel Kahneman, Paul Slovic, and Amos Tversky (1982). This view maintains that we routinely violate basic norms of rationality because our reasoning is not guided by principles of logic and probability, but rather by simple heuristics and biases that often lead us astray. For example, we commit the conjunction fallacy because we employ a *representativeness heuristic*, which assigns higher probabilities to outcomes that are more representative of prototypical exemplars. On this view, our reasoning competence is defective, since it ignores basic facts about the structure of the probability calculus. Let's call this the Pessimistic View (Samuels and Stich 2004).

On the other hand, we find a more optimistic view of human rationality associated with evolutionary psychologists, such as Leda Cosmides, John Tooby, and Gerd Gigerenzer. On this view, our reasoning competence is realized in mental modules that are reliable in the adaptive environments in which they evolved. This view is motivated in part by empirical evidence that performance on reasoning tasks improves when they are formulated in appropriate content-sensitive terms. For example, subjects perform much better in Wason's selection task when it is formulated as a *cheater detection* task. Cosmides and Tooby (1992) argue that this is because our ability to reason with conditionals is subserved by a cognitive module that evolved for social purposes, including detection of cheaters. Similarly, subjects perform better in probabilistic reasoning tasks when they are formulated as questions about frequencies. This leads Cosmides and Tooby (1996) to argue that our mechanisms for probabilistic reasoning are adaptively designed to operate on probabilistic information about natural frequencies. More generally, they claim that our reasoning modules are reliable in their adaptive environments. Let's call this the Optimistic View.

Although these two views diverge in their degree of pessimism about human rationality, both views offer a decidedly mixed assessment. On the one hand, the Pessimistic View is not entirely pessimistic. After all, our heuristics do not always lead us astray. Even in their earlier work, which is most easily interpreted as suggesting a bleak view of human rationality, Kahneman and Tversky note that our reasoning "sometimes yield reasonable judgments and sometimes lead to severe and systematic errors" (1973: 48). Moreover, this mixed view became increasingly explicit as their research program matured (e.g. Gilovich, Griffin & Kahneman 2002: 9). On this view, our reasoning is guided by heuristics that conform somewhat reliably, although far from perfectly, to principles of logic and probability. As such, these heuristics cannot be dismissed as entirely "stupid" or insensitive to the norms of rationality. On the contrary, they approximate towards ideal rationality insofar as their outputs are close enough to optimal across a wide enough range of cases (Samuels, Stich & Bishop 2002).

Similarly, the Optimistic View is less than entirely optimistic. While there are contexts in which people conform well enough to the canons of rationality articulated by the Standard Picture, there are also many contexts in which they don't, as evolutionary psychologists acknowledge. Although performance improves in some settings, the fact remains that we routinely violate these canons of rationality on many reasoning problems that arise in contemporary society. There is no denying the evidence that human reasoning is not always rational.

Similar points apply to the *dual processing theory* of reasoning, which is sometimes viewed as a "middle way" between the pessimism of the heuristics and biases tradition and the comparative optimism of evolutionary psychology (e.g. Samuels and Stich 2004). On this theory, reasoning is subserved by two distinct kinds of cognitive system that are characterized by a cluster of related properties. System 1 reasoning is presumed to be fast, holistic, automatic, unconscious, and requiring little cognitive capacity, while system 2 reasoning is slow, rule-based, controlled, conscious, and requires more cognitive capacity. (Evans 1989; Sloman 1996; Stanovich 1999).

Once again, a mixed assessment of human rationality emerges. On the one hand, the dual processing theory supports the pessimistic conclusion that our reasoning competence is not designed to solve all the reasoning problems that we face in modern society, since System 1 is innately fixed and emerged early in human evolution. On the other hand, it also provides some grounds for optimism about human rationality, since System 2 evolved more recently and there is some evidence that performance in reasoning tasks can be improved in ways that reflect the influence of culture and education on System 2 reasoning (Nisbett 1993). This optimism should be significantly tempered, however, since we can only rely on System 2 when our reasoning is slow, controlled, and conscious. Moreover, System 2 has its own characteristic limitations. For example, while it deals better with abstract problems that involve small quantities of information, performance deteriorates significantly in reasoning tasks that require weighing large amounts of information (Dijksterhuis 2004).

The consensus that emerges from this brief overview is that almost everyone working on the psychology of reasoning accepts a mixed view of human rationality: we are rational enough for government purposes, but very far from perfect. Indeed, it's hard to see how any adequate psychology of human reasoning could fail to generate such a mixed assessment. After all, there is good evidence that we systematically violate norms of rationality and there is also good evidence that under many conditions we accord with them. This mixed pattern in the empirical data is something that any adequate theory of human reasoning needs to explain.

2.3. Implications for the Rationality Constraint

What are the implications of this empirical research for the Rationality Constraint? We draw three main conclusions. First, and most obviously, we should reject any version of the Rationality Constraint that demands *perfect rationality*, since the empirical evidence confirms that our reasoning competence is far from perfectly rational. Arguably, however, this conclusion can be established on independent grounds, since there are principled reasons to suppose that perfect rationality is simply not computationally feasible for finite creatures like us (Cherniak 1986; Harman 1986; Nichols & Samuels 2017). This demanding version of the Rationality Constraint threatens eliminativism: if perfect rationality is required for belief, then human beings never believe anything at all.

Second, if the empirical argument undermines only this extremely demanding version of the Rationality Constraint, then it targets a strawman – a position that's easily rejected but accepted by hardly anyone. After all, proponents of the Rationality Constraint tend to reject the demand for perfect rationality in favor of some more minimal requirement. Lewis, for example, is very explicit about this:

> It wouldn't do to conclude that, as a matter of analytic necessity, anyone who can be said to have beliefs and desires at all must be an ideally rational *homo economicus*! Our rationality is very imperfect. (1999: 321)

On Lewis's version of the Rationality Constraint, we need not be perfectly rational to have beliefs and desires so long as we are imperfectly rational enough to meet some minimal threshold. This is not to reject the Standard Picture, according to which the standards of perfect rationality are articulated with reference to formal principles of logic and probability. Indeed, Lewis explicitly accepts the Standard Picture:

> I think that systematic theories of ideal rationality – decision theory, for instance, and the theory of learning from experience by conditionalizing a subjective probability distribution – are severely idealized parts of folk psychology. (1999: 321)

Instead, he claims that having beliefs requires only some minimal degree of approximation towards the ideal of perfect rationality, where this minimal threshold is set by folk psychology.

Third, and finally, the empirical research on the psychology of reasoning poses no obvious threat to Lewis's minimal version of the Rationality Constraint. It can be reconciled with the empirical evidence so long as it requires only a minimal enough degree of rationality. Indeed, the empirical evidence tends to confirm Lewis's assessment of the degree of human rationality that is assumed in folk psychology. As we have seen, the empirical evidence provides no support for the radically pessimistic conclusion that we have no rational competence at all. On the contrary, it comports rather well with Lewis's claim that our folk psychology credits us with at least a "modicum of rationality" (1999: 320).

Of course, this response to the empirical challenge prompts an important question. If having beliefs requires that we are rational enough to meet some minimal threshold, then how much rationality is enough? Call this *the threshold question*. An adequate answer to this question need not specify a precise threshold. Presumably, the threshold for minimal rationality is both vague and context sensitive. Even so, we need some answer – even one that is vague and context-sensitive – to give content to the thesis that there are rationality constraints on belief. Otherwise, the Rationality Constraint risks collapsing into vacuity.

Lewis's answer to the threshold question is that *folk psychology* determines how much rationality we need to have beliefs and desires:

> Folk psychology can be taken as a theory of imperfect, near-enough rationality, yet such rationality as it does affirm can still be constitutive. (1999: 321)

The suggestion is that we presuppose a certain degree of rationality in our folk-psychological practice of predicting and explaining people's behavior in terms of their beliefs and desires. We don't need to assume perfect rationality for these purposes, but only some minimal degree of rationality. In order to count as having beliefs and desires, you must be rational enough that your behavior can be reliably predicted and explained in accordance with the principles implicit in folk psychology. In this way, the threshold for minimal rationality is fixed by the predictive and explanatory demands on the success of our folk psychology.

Lewis's answer to the threshold question does not entirely avoid the empirical challenge. After all, the empirical evidence from the psychology of reasoning suggests that we are considerably less rational than folk psychology assumes. While everyone knows that we sometimes make mistakes in reasoning, it is genuinely surprising to learn that we routinely violate even the most basic principles of logic and probability theory. Indeed, modern psychology is replete with surprising empirical discoveries of this kind in which we find that our reasoning and behavior deviates in surprising ways from what our folk psychology would lead us to expect.

This poses a problem for Lewis because he claims that the principles of folk psychology are "platitudes" that are common knowledge among the folk. If the meanings of folk-psychological terms, such as 'belief' and 'desire', are implicitly defined by their role in a theory, then it must be a theory that everyone knows, rather than a theory that is discovered in the laboratory. This is because we understand our folk-psychological terms well enough to know what they mean. If we build scientific discoveries into the theoretical principles that define these terms, then we jeopardize our claim to know what they mean. As Lewis says, "Esoteric findings that go beyond common sense must be kept out, on pain of changing the subject" (1983: 112).

The problem is that the empirical facts about human irrationality are surprising: these are "esoteric findings that go beyond common sense". Indeed, they conflict with the expectations set by folk psychology. The fact that these scientific discoveries come as a surprise to us suggests that we are much *less* rational than our folk psychology assumes. But if we violate the minimal standards of rationality that are built into folk psychology, then Lewis's answer to the threshold question threatens eliminativism. On pain of denying that we have beliefs and desires, we may seem forced to reject Lewis's version of the Rationality Constraint.

On reflection, however, this problem is not fatal. Lewis can assuage the problem by exploiting a general feature of his semantics for theoretical terms. According to Lewis's semantics, a theory need not be entirely accurate for its theoretical terms to secure reference. Theoretical terms can refer so long as there are entities that conform *well enough* to the principles of the associated theory. In the case of folk psychology, we can have beliefs and desires so long as we approximate towards satisfying the psychological principles that are implicit in our folk psychology. This answers the empirical challenge by allowing for some discrepancy between folk psychology and our best scientific psychology of human reasoning. As Lewis writes, "An imperfect but near-enough occupant of a folk-psychological role could thereby be an imperfect but near-enough deserver of a folk-psychological name" (1999: 321).

3. The Psychopathology of Delusion

We turn now from the psychology of reasoning to the psychopathology of delusion. In particular, we focus on so-called *monothematic delusions*, which have a single topic and are usually highly circumscribed. Examples include Capgras delusion, in which patients report that someone close to them – typically a spouse or a relative – has been abducted and replaced by an imposter, and Cotard delusion, in which a patient reports that they are dead.

Some philosophers, including Lisa Bortolotti (2010), argue that these monothematic delusions undermine the Rationality Constraint. We've already seen evidence that our ordinary methods of belief-revision are rationally problematic according to standard canons of rationality. But the evidence regarding monothematic delusion seems even worse: these delusional beliefs appear to be isolated from ordinary processes of belief-revision altogether. As such, they present apparent counterexamples to the Rationality Constraint in which delusional patients hold beliefs that are so irrational that they violate even minimal standards of rationality. Call this *the argument from delusion*.

The materials for such an argument are implicit in the standard definition of delusion in the *Diagnostic and Statistical Manual of Mental Disorders* (DSM) published by the American Psychiatric Association:

> Delusions are fixed beliefs that are not amenable to change in light of conflicting evidence. (DSM-5 2013: 87)

First, the DSM endorses a *doxastic conception of delusion*, according to which delusions are instances of belief. And second, the DSM distinguishes delusions from other beliefs in part by their irrationality: they are resistant to rational processes of belief-revision in such a way that they are firmly sustained even in the face of clear and compelling counterevidence. If delusional beliefs flagrantly violate the norms of epistemic rationality in this way, this threatens to undermine the thesis that there are any rationality constraints on belief at all.

We can represent this argument against the Rationality Constraint as an inconsistent triad:

(1) *The Doxastic Conception of Delusion*: Delusions are beliefs.
(2) *The Irrationality of Delusional Beliefs*: If delusions are beliefs, then they violate even minimal standards of rationality.
(3) *The Rationality Constraint*: All beliefs are at least minimally rational.

The argument from delusion maintains that we should we reject the Rationality Constraint by using (1) and (2) to undermine (3). Obviously, this argument is cogent only if it is reasonable to accept (1) and (2). Otherwise, we can resolve the inconsistency by rejecting one of these claims. In this section, we consider and reject two prominent attempts to block the argument from delusion in this way. Nevertheless, we maintain, the argument from delusion fails to undermine minimal versions of the Rationality Constraint defended by Lewis and others.

3.1. The Doxastic Conception of Delusion

One strategy for resolving the inconsistent triad is to reject the doxastic conception of delusion. For example, Gregory Currie and Ian Ravenscroft (2002) argue against the doxastic conception of delusion on the grounds that delusions violate the Rationality Constraint. In effect, they argue from (2) and (3) against (1). On their view, delusions are not beliefs, but psychological states of some other kind – namely, imaginative states that are misidentified as beliefs in acts of metacognitive monitoring.

Why should we follow the DSM in supposing that delusions are beliefs? The doxastic conception of delusion is often defended on broadly functionalist grounds by appealing to similarities between the causal roles of delusion and belief. The argument is that delusions should be classified as beliefs since they function in ways that resemble paradigm cases of belief (Stone & Young 1997; Bayne & Pacherie 2005; Bortolotti 2010).

Why should we suppose, for example, that Capgras patients genuinely believe the delusional hypothesis that their relative has been abducted and replaced with an imposter? First, and most obviously, because this is what they say: they assert with apparent sincerity and conviction that the delusional hypothesis is true. Second, they will sometimes defend their assertions by citing evidence that they take to support the delusional hypothesis, including facts about their own experience. Third, patients sometimes act and react as if they believe the delusional hypothesis: for example, feeling upset or acting aggressively towards the alleged imposter. Finally, delusions sometimes figure in inferences: for example, Young et al. (1992)

note that a patient with Cotard delusion made the following inference while visiting South Africa: (i) it's hot, and (ii) I'm dead, so (iii) I must be in hell. The general point is that delusions play enough of the functional roles that are associated with belief in our folk psychology that they should be classified as beliefs too.

3.2. The Irrationality of Delusional Beliefs

A second strategy for resolving the inconsistent triad is to argue that delusions satisfy the rationality constraints on belief: that is, to argue from (1) and (3) against (2). For example, Brendan Maher (1974) defends the doxastic conception of delusion by arguing that delusional beliefs are rational responses to anomalous experiences. Although he is mainly concerned with polythematic delusions in schizophrenia, we can extend his account to monothematic delusions too. Consider Ellis and Young's (1990) proposal that Capgras delusion involves a kind of *inverse prosopagnosia* in which patients experience diminished affective responses to intact perception of familiar faces. Can we regard the Capgras delusion as an epistemically rational response to an anomalous experience in which you see someone who looks just like your relative but who nevertheless seems unfamiliar?

Let's distinguish two hypotheses about the etiology of the Capgras delusion. First, consider the *explanatory hypothesis*, which says that delusional beliefs are based on abductive inference. On this view, the patient with Capgras delusion believes their relative has been replaced by an imposter because they regard this as the best explanation of why this person who looks just like their relative nevertheless seems unfamiliar. The problem is that this abductive reasoning is not epistemically rational, since the explanatory hypothesis is extremely improbable given the available evidence. There are much better explanations of why this person looks unfamiliar that are overwhelmingly more probable given the available evidence – for instance, my spouse seems unfamiliar because I've suffered a brain injury that diminishes my affective responses.

Second, consider the *endorsement hypothesis*, which says that delusional beliefs are based on endorsing the representational contents of perceptual experience. On this view, the patient with Capgras delusion endorses the content of an anomalous experience, which represents that this person who looks just like their relative is really someone else. This hypothesis is problematic for several reasons. First, it relies on the controversial assumption that perceptual experience represents the identities of specific individuals, rather than merely their visible properties (see Byrne and Siegel 2017). Second, it fails to explain aspects of delusional belief that go beyond the content of perceptual experience, such as the belief that one's relative has been abducted. And third, it fails to explain why delusional beliefs are routinely maintained in ways that are unresponsive to defeating evidence, as when patients with the Cotard delusion maintain that they are dead even while acknowledging that their heart is still beating.

On either hypothesis, monothematic delusion involves serious defects in epistemic rationality. Anomalous experience may contribute towards explaining the Capgras delusion, but it cannot be the whole story. After all, patients with damage to the ventromedial prefrontal cortex experience diminished affective responses to intact perception of familiar faces without experiencing the Capgras delusion (Tranel et al. 1995). This motivates *two-factor theories* of monothematic delusion, which supplement the appeal to anomalous perception with some rational defect in cognition, such as a bias towards observational adequacy or a deficit in belief evaluation (e.g. Stone and Young 1997; Davies et al. 2001). While the exact nature of this cognitive factor is disputed, it remains plausible that monothematic delusions typically involve serious violations of the norms of epistemic rationality. Any defense of the doxastic conception of delusion needs to acknowledge this much.

3.3. The Rationality Constraint

A third strategy for resolving the inconsistent triad is advocated by proponents of the argument from delusion. On this view, we should argue from (1) and (2) against (3). For example, Lisa Bortolotti (2010) defends the doxastic conception of delusion by abandoning rationality constraints on belief. Thus, she writes:

> Instead of establishing that delusions are not beliefs on the basis of the rationality constraint on belief ascription, we should be open to rejecting the constraint because delusions and other irrational beliefs get ascribed and play the same role as rational beliefs in underpinning explanation and prediction of behaviour in intentional terms. (2010: 8)

What should we make of this? Once again, it's crucial to distinguish versions of the Rationality Constraint that demand perfect rationality from those that merely demand minimal rationality. Clearly, the argument succeeds in undermining the most demanding versions of the Rationality Constraint, but these demanding versions are rarely endorsed, since they are implausible on independent grounds. As we've seen, proponents of the Rationality Constraint, such as Lewis, typically endorse a rather more minimal constraint on rationality. And here the argument from (1) and (2) against (3) seems rather less compelling, since it is not at all clear that delusions violate this minimal rationality constraint.

One way to bring out the problem is to notice how Bortolotti's argument for the doxastic conception of delusion implicitly assumes some minimal version of the Rationality Constraint. She argues that delusions are beliefs because they satisfy various constitutive constraints on the functional role of belief, including the following:

> Beliefs have relations with the subject's other beliefs and other intentional states.
> Beliefs are sensitive to the evidence available to the subject.
> Beliefs are manifested in the subject's behavior. (2010: 12)

At the same time, she contrasts these minimal functional constraints on belief with the more demanding rationality constraints listed below:

> Beliefs are procedurally rational if they are *well*-integrated in a system with other beliefs or intentional states

> Beliefs are epistemically rational if they are *well*-supported by and responsive to the available evidence.
>
> A subject is agentially rational if she can endorse beliefs by giving *good* reasons in support of their content, and by acting in a way that is *consistent with* and *explicable by* their content. (2010: 14)

We can draw this contrast because rationality comes in degrees: one's beliefs can be rationally sensitive to evidence, integrated with other beliefs, and manifested in behavior without exhibiting these rational virtues to a very high degree. Nevertheless, it would be a mistake to suppose that Bortolotti's functional constraints on belief have nothing to do with rationality. Although she doesn't explicitly use normative language in stating these constraints, we need to make normative assumptions in order to give them any substance. As stated, they remain crucially underspecified. We need to ask how exactly your beliefs must be sensitive to evidence, integrated with other beliefs, and manifested in behavior. Presumably, not just any causal relations will do no matter how crazy they are. That would impose no significant functional constraint on belief at all. Instead, these causal relations must be at least minimally rational. If so, then Bortolotti cannot abandon all rationality constraints on belief. On the contrary, her argument for the doxastic conception of delusion assumes that there are at least some minimal rationality constraints on belief.

This reveals a tension between Bortolotti's argument for the doxastic conception of delusion and her argument against the Rationality Constraint. The functional similarities between belief and delusion undercut the reasons for supposing that delusions provide any special threat to the Rationality Constraint. After all, as Bortolotti argues, the irrationality of delusion is continuous with the irrationality of ordinary, non-pathological cases of belief, including superstition, racial prejudice, and self-deception. She writes:

> The irrationality of delusions is not different in kind from the irrationality of everyday beliefs. Delusions are on a continuum with irrational beliefs, and you are likely to find them towards the 'very irrational' end of the line. (2010: 260)

If the irrationality of delusions is continuous with the irrationality of ordinary belief, then we don't need empirical evidence from the psychopathology of delusion in order to provide counterexamples to the rationality constraint. Ordinary examples of irrationality will suffice.

What this means, however, is that Bortolotti's argument against the Rationality Constraint succeeds only on an implausibly demanding conception of its strength. If the standards of rationality required for belief are very demanding, then they are violated even in ordinary cases of irrational belief, which means we don't need delusions to generate counterexamples. But if the required standards of rationality are minimal enough to be satisfied in ordinary cases, then arguably they are satisfied in pathological cases as well. After all, the irrationality of delusion is continuous with the irrationality of ordinary belief.

In conclusion, delusions pose no more of a threat to the Rationality Constraint than ordinary cases of irrational belief. If we operate with sufficiently minimal standards of rationality, then we can resolve the inconsistent triad by rejecting (2) while endorsing (1) and (3). Moreover, we can do this while still recognizing that delusional beliefs are epistemically irrational. Although delusional beliefs are perhaps more egregiously irrational than ordinary beliefs, the difference is one of degree rather than kind. The empirical challenge from the psychopathology of delusion fails, since the functional similarities between belief and delusion suggest that delusions are not so egregiously irrational that they fail to meet some very minimal threshold. Once again, the empirical evidence is not only consistent with the thesis that delusions are minimally rational; it coheres with it rather well.

4. The Continuity Argument

So far, we've argued that neither the psychology of reasoning nor the psychopathology of delusion poses any serious problem for Lewis's minimal version of the Rationality Constraint. Indeed, in both cases, we suggested that the empirical evidence coheres rather well with Lewis's own assessment of the extent of human rationality. For he claims that we need not be perfectly rational so long as we are rational enough to meet some minimal threshold.

To put pressure on the Rationality Constraint, we need to move from actual cases of human irrationality to more extreme cases of irrationality that are non-actual but nevertheless conceivable. In this section, we argue that there are conceivable cases of *mad belief* in which communities of alien madmen have beliefs whose causal roles are not even minimally rational. Of course, the danger in moving from empirical arguments to conceivability arguments is that it becomes too easy to deny that these extreme forms of irrationality are genuinely conceivable. However, we'll support our claims about conceivability by appealing to the continuum between actual and non-actual cases of irrationality. Our strategy is to move, via successive coherent transitions, from empirical evidence about actual cases of irrationality in human beings to intuitions about conceivable cases of irrationality in alien madmen. We call this the *Continuity Argument*, since we put significant weight on the continuity between actual and counterfactual cases of irrationality.

Cases of irrationality fall on a continuum: some departures from perfect rationality are more egregious than others. Given this continuum, it seems *ad hoc* and unprincipled to insist on some minimal threshold for rationality that is weak enough to include all actual cases of irrationality but strong enough to exclude our counterfactual cases. We contend that there is no such threshold built into our ordinary concepts of belief and desire. There may be contingent empirical limits on how far we actually depart from the ideal of perfect rationality, but there are no principled conceptual limits on how far any possible believer may depart from the ideal. The assumption of rationality is not so central to our concept of mind that we cannot coherently conceive of Lewisian madmen whose beliefs are not even minimally rational.

Our argument proceeds in three steps. The first step is to explain why the doxastic conception of delusion remains plausible in actual cases of Capgras delusion, despite considerable variation in the extent to which rational functions of belief are preserved. In all actual cases, patients make assertions that express their feeling of conviction that the delusional hypothesis is true. The second step is to argue that this rationale for the doxastic conception of delusion extends to counterfactual cases of mad belief in which all the rational functions of belief are excised. Finally, the third step is to extend the madness from individual beliefs to whole belief-systems and from individual believers to whole communities.

We claim that these madman scenarios are coherently conceivable. If conceivability is a guide to possibility, then this is defeasible evidence that the scenarios are genuinely possible. But we take no stand here on the relationship between conceivability and possibility. Our claim is simply that there is no conceptual incoherence in the hypothesis that a community of Lewisian madmen might have beliefs that are not even minimally rational. This is enough to cast serious doubt on Lewis's claim that it is an analytic truth that our beliefs must be rational enough to meet some minimal threshold.

4.1. Capgras Delusion

The first step in our argument begins with actual cases of Capgras delusion. As we saw in §3, patients with Capgras delusion hold beliefs that are deeply irrational. At the same time, we must acknowledge that some rational functions of belief are preserved in Capgras delusion. Delusional beliefs continue to play some minimally rational role in responding to evidence, guiding behavior, and integrating with other beliefs in inference. Hence, these cases do not constitute clear counterexamples to the Rationality Constraint so long as the threshold for minimal rationality is set low enough.

Crucially, however, Capgras patients vary in the extent to which the rational functions of belief are preserved in their actions and reactions. Some are disposed to act and react in predictable ways given the content of their delusional belief: for example, feeling upset or exhibiting hostility towards the alleged imposter. And yet many patients appear wholly unmoved by their delusional belief: they display no concern about the fate of the abducted spouse and seem content to live alongside the alleged imposter as if nothing untoward had happened.

Such cases present a challenge for the doxastic conception of delusion. Given the varying degrees of rationality exhibited in actual cases of Capgras delusion, why should we have any confidence that they all satisfy minimal rationality constraints on belief? If believing a hypothesis is simply a matter of having a functional state that occupies enough of the causal role associated with belief in our folk psychology, then some cases of Capgras delusion would seem to be borderline cases at best. Perhaps, as Eric Schwitzgebel (2012) suggests, these are "in-between" cases in which it's indeterminate whether the delusion plays enough of the right causal role to count as a genuine belief.

In our opinion, however, Schwitzgebel's view underestimates the plausibility of the doxastic conception of delusion. Perhaps the most compelling point in its favor is that delusional patients *assert,* with apparent sincerity and conviction, that the content of the delusion is true.

In their seminal defense of the doxastic conception, Tony Stone and Andy Young rest significant weight on this point:

> Individuals experiencing the Capgras or Cotard delusion appear to be *sincere* when they make assertions about imposters, or about themselves being dead. Again, Quine (1990) has pointed out that asking is the easiest way to find out what someone believes so long as you can assume your informer's sincerity. In the absence of contrary evidence, and given that such patients provide consistent answers to questions about their delusions, we therefore maintain that these sincere assertions provide a prima facie reason for thinking that the patients are expressing a belief, albeit with bizarre content. (1997: 354).

As Stone and Young note, we treat apparently sincere and steadfast assertion as evidence that settles the question of what someone believes. But this is not easy for Schwitzgebel's view to explain either in cases of delusion or in non-pathological cases, such prejudice and superstition. After all, assertion is merely one kind of verbal behavior that must be weighed against the rest of an individual's verbal and non-verbal behavior. But then why should we give any privileged role to the speech act of assertion in our ordinary practices of belief-ascription?

We suggest that this is because belief has a phenomenal dimension and not just a causal one. One aspect of believing a hypothesis is being disposed to *feel conviction* that the hypothesis is true. We express our feelings of conviction in assertion, at least when we are sincere, and we expect others to do the same. So, when someone asserts that *p* with apparent sincerity, we treat this as evidence that they feel convinced that *p*. And, crucially, we treat the disposition to feel conviction that *p* as a sufficient condition for believing that *p*. That is why we treat their apparently sincere assertions as evidence that delusional patients believe what they say. Their linguistic behavior is compelling evidence that they are disposed to feel conviction that that the delusional hypothesis is true, notwithstanding the conflict with other aspects of their linguistic and non-linguistic behavior. As we argue in §4.2, nothing more is required for believing that *p* than being disposed to feel convinced that *p* when you consciously consider whether *p* is true.

Schwitzgebel argues on independent grounds that the disposition to feel conviction that *p* is not sufficient for believing that *p*. He gives the example of Juliet, a philosophy professor who feels convinced that all races are equal in intelligence, and consistently affirms as much in discussion, but who is systematically racist in her spontaneous reactions, unguarded behavior, and judgments about particular cases. According to Schwitzgebel, this is an "in-between" case: it's indeterminate whether Juliet believes that all races are intellectually equal.

We are not persuaded by this example. If we say it's false, or indeterminate, that Juliet believes what she asserts, then we deprive ourselves of the resources to explain her linguistic behavior. Why does she consistently assert that all races are intellectually equally? The obvious explanation – and, we maintain, the correct one – is that she gives sincere expression to what she believes. This is not to deny that Juliet also holds conflicting beliefs. Indeed, Schwitzgebel asks us to imagine that Juliet is prone to make judgments about particular cases that conflict with her general belief in racial equality. And this goes some way to explain the tendency to deny that Juliet really (or wholeheartedly) believes what she asserts to be true. After all, there is some inconsistency – or other rational tension – within her belief system.

We need not deny, of course, that patients with Capgras delusion have similar rational tensions within their belief system. We are concerned only to maintain that they believe the delusional hypothesis that they affirm in speech. Our claim is that they are using the speech act of assertion in the normal way to give sincere expression to their belief that the content of their assertion is true. Although patients with Capgras delusion vary considerably in their degree of rationality, they all share a common phenomenal core – namely, the disposition to feel conviction that the delusional hypothesis is true. This phenomenal core is what unifies all actual cases of Capgras delusion.

4.2. From Capgras Delusion to Mad Belief

The next step in our argument moves from actual to counterfactual cases of delusion. In actual cases, some of the rational functions of belief remain intact, although individual cases vary in how much rationality is preserved. In some cases, the only remaining vestige of the rational profile of belief is the disposition to affirm its content in speech. We can now imagine a more extreme version of the Capgras delusion – *super-Capgras delusion* – in which all the rational functions of belief are excised, including its rational role in speech. Crucially, however, we hold fixed the patient's disposition to feel conviction that the delusional hypothesis is true.

As we envisage this example, the patient with super-Capgras delusion is disposed to feel utterly convinced that the delusional hypothesis is true when he consciously entertains it. At the same time, however, his feelings of conviction have no intelligible basis in evidence and play no minimally rational role in reasoning, action, or speech. These cognitive feelings are "mad" in the technical sense that Lewis (1983: 122) defines: they don't play the causal roles that feelings of conviction normally play in human psychology. On one version of the case, they are *epiphenomenal*: they play no causal role in the patient's psychology at all. On another version, they play a highly *eccentric* causal role. Either way, and however we fill in the details of the case, it remains plausible that these delusions are beliefs insofar as they dispose their subjects to feel convinced that their contents are true. Hence, there are conceivable cases of *mad belief*.

When we imagine this case, it is important to consider how things seem from the patient's own perspective. From the third-person perspective of a radical interpreter, it remains obscure what the patient believes, since their linguistic and non-linguistic behavior is completely haywire. And yet we cannot lose sight of the fact that, from the patient's own perspective, it seems evident that the delusional hypothesis is true. The patient may reach this conclusion and use it in reasoning in ways that seem utterly bizarre to us. But none of this shakes his own firm and stable feeling of conviction that the delusional hypothesis is true.

Someone might protest that this scenario is not conceivable at all. In response, however, we can surely imagine scenarios in which the phenomenal character of an experience is dissociated from its normal causal role. For example, we can imagine *zombies* in which the causal role of pain is played without any feeling of pain. And, conversely, we can imagine *madmen* in whom the feeling of pain plays an abnormal causal role. As Lewis writes:

> There might be a strange man who sometimes feels pain, just as we do, but whose pain differs greatly from ours in its causes and effects. (1983: 122)

If we can imagine feelings of pain in the absence of their normal causal role, as Lewis concedes, then we can do the same for feelings of conviction.

Like other experiences, the feeling of conviction is an *intentional state*: in other words, it has an intentional content. You cannot experience the feeling of conviction without thereby feeling conviction that some content is true. Moreover, it's plausible that the phenomenal character of the experience reflects its content: for instance, what it's like to feel convinced that *p* is different from what it's like to feel convinced that *not-p* or to feel convinced that *p* only if *q*. When external circumstances are held fixed, the phenomenal character of the experience varies depending on which content you feel convinced is true. This leads some philosophers to argue that the phenomenal character of cognitive experience plays an essential role in determining its content, perhaps in combination with some causal contribution from the external environment (Strawson 1994; Siewert 1998: Ch. 8; Horgan and Tienson 2002; Pitt 2004; Smithies 2019: Ch. 4).

Obviously, the task of defending any specific theory of content is one that goes beyond the scope of the current paper. Our main goal here is to pose a challenge for Lewis's analytic functionalism, which he intends as a theory of contents as well as attitudes, rather than to defend an alternative theory. If our challenge succeeds, this poses a problem not only for Lewis's analytic functionalism, but also for many other functional role theories of content that are inspired by it. Nevertheless, the appeal to constitutive functional roles is not the only resource available for giving a theory of content. In particular, as we just noted, one attractive option combines an appeal to the phenomenal character of experience with causal relations to the external world. There may be other options too. However, we must leave the task of developing a theory of content for another occasion.

Another objection is that although we can imagine madmen whose feelings of conviction play eccentric causal roles, we do not thereby succeed in imagining cases of *mad belief*. But if we can imagine mad pain, then why not mad belief too? Some philosophers who reject the functional analysis of phenomenal consciousness nevertheless express sympathy for a functional analysis of cognition (e.g. Block 1978). But what justifies this asymmetry? And why think Lewisian madmen are any less capable of holding beliefs than feeling pain?

One possible answer is that pain is a feeling that has phenomenal character, whereas belief is a dispositional state that can persist through time without impacting the stream of phenomenal consciousness. In response, however, our dispositional beliefs are disposed to manifest themselves in phenomenal consciousness as occurrent feelings of conviction. And these cognitive feelings do have phenomenal character: there's something it's like to feel convinced that something is true. If occurrent feelings of conviction are instances of belief – what philosophers sometimes call 'occurrent beliefs' – then this is already enough to establish the conceivability of mad belief. Alternatively, we might construe belief as a standing dispositional state, which is distinct from its occurrent manifestations in phenomenal consciousness. Here, again, we can allow for cases of mad belief by divorcing the disposition to feel conviction from its normal causal role.

Eric Schwitzgebel (2012) claims that mad belief is inconceivable on the grounds that belief is a *multi-track disposition*. Although beliefs are disposed to manifest themselves in phenomenal consciousness as feelings of conviction, they are also disposed to manifest themselves in other ways, including action and reasoning. And, crucially, Schwitzgebel maintains that all of these dispositions – or enough of them – are essential to belief. This means that a Lewisian madman who is disposed to feel conviction that *p* doesn't count as believing that *p* because his states don't play enough of the other functional roles associated with our folk concept of belief. At best, it's indeterminate whether or not the madman believes that *p*.

As we saw in §4.1, however, Schwitzgebel's account generates implausible verdicts in some actual cases of Capgras delusion. It's plausible that all Capgras patients believe the delusional hypothesis, since they assert that it's true with apparent sincerity and conviction. And yet Schwitzgebel's account implies that this is not determinately true, since many Capgras patients are not disposed to act and react as if their assertions are true. What these cases seem to reveal is that the disposition to feel conviction, as expressed by the sincere assertion that *p*, carries more weight in our ordinary practices of belief-ascription than other aspects of the normal functional role of belief. That is why we can generate cases of mad belief by holding fixed the phenomenal disposition to case feelings of conviction, while varying other aspects of its functional role.

Unlike Schwitzgebel, Lewis does not dispute the conceivability of mad belief. On the contrary, he explicitly acknowledges that madmen can have beliefs and desires as well as feelings of pain. For example, in the postscript to "Radical Interpretation", he writes:

> In "Mad Pain and Martian Pain", I argued that a "madman" might be in pain not because his state occupied the causal role of pain in him but rather because that state occupies that role, for the most part, in members of the kind to which he belongs. The same possibility should be recognized for attitudes as well. Karl might believe himself a fool, and might desire fame, even though the best interpretation of Karl considered in isolation might not assign those attitudes to him. (1983: 119)

In §4.3, we criticize Lewis's attempt to reconcile this concession with his analytic functionalism.

4.3. From Mad Beliefs to Mad Believers

The third and final step in our argument extends the madness from individual beliefs to whole systems of belief and from individual believers to whole communities of believers.

One surprising feature of Capgras delusion is the way in which failures of rationality are localized to a specific topic. As Stone and Young note:

> It is characteristic of people suffering from the Capgras delusion that their bizarre beliefs can be highly circumscribed – they have an island of irrationality within a sea of rationality. (1997: 357)

The same is true in our example of super-Capgras delusion. But if we can imagine mad beliefs that are circumscribed in this way, then we can imagine a madman whose entire system of beliefs is infected with madness. In this madman, there is no trace of the rational function of belief anywhere in his psychology. Nevertheless, it remains plausible that the madman has beliefs insofar as he experiences occurrent feelings of conviction when he consciously entertains their contents and considers whether they are true. It's just that these feelings of conviction play no minimally rational role in his action, speech, or reasoning.

Going further, we can imagine a whole community composed exclusively of madmen. As we use the term, following Lewis, a madman's *community* (or 'population') is constituted by members of his kind – for example, his species. We can imagine that any member of the madman's species is a madman too. In the limit case, we can imagine a solitary madman in a community of one. And when we imagine the solitary madman, we needn't imagine a human being. We can imagine a member of some alien species, such as a Martian. When we imagine the solitary mad Martian, we imagine that he feels just as we do when we experience feelings of pain or feelings of conviction, although his feelings differ from ours in both physical realization and causal role. We thereby imagine that the solitary mad Martian feels pain, just we do, and shares our beliefs, though these mental states do not play their normal causal roles.

Lewis rejects this third and final step in our argument. As we've seen, he acknowledges that we can imagine cases of mad pain and mad belief. At the same time, however, he maintains that we can imagine such cases only as exceptions to a more general rule. On this view, the madman has mental states only because he has tokens of the same physical type that normally play the requisite causal role either in him or in other members of his species. Thus, Lewis abandons the individualistic version of the Rationality Constraint that he defended in "Radical Interpretation" and replaces it in the "Postscript to Radical Interpretation" with a communal version inspired by his later work in "Mad Pain and Martian Pain". On this communal version, not all believers are required to be minimally rational so long as this is the normal case.

This means that Lewis's analytic functionalism succumbs to the *problem of chauvinism* that it was originally designed to avoid. A chauvinistic theory of mind is one that "withholds mental properties from systems that in fact have them" (Block 978: 265). Consider the mind-brain identity theory, which identifies mental states with neural states of the brain. This theory is chauvinistic because it withholds mental properties from creatures that do not have neural states at all. As Lewis notes, for example, “There might be a Martian who sometimes feels pain, just as we do, but whose pain differs greatly from ours in its physical realization” (1983: 123).

One motivation for analytic functionalism is to avoid the problem of chauvinism by allowing that mental states can be *multiply realized*. Lewis can allow that Martians share our mental states by sharing causal roles that have different physical realizers. Similarly, he can allow that madmen share our mental states by sharing physical realizers that play different causal roles. He can even allow for a mad Martian who shares our mental states by sharing physical realizers with other Martians who share causal roles with us. And yet he cannot accommodate mental states in an alien species of Martians composed exclusively of madmen. In particular, as he explicitly acknowledges, he cannot accommodate mental states in the limit case of a solitary mad Martian. After all, there is no overlap in either physical realization or causal role that would explain how they share mental states with us. And he cannot appeal to the phenomenal overlap between us, since this presupposes what his theory is supposed to explain. This prevents Lewis from including the isolated mad Martian as an unusual member our own community.

This chauvinistic consequence of Lewis’s analytic functionalism undermines his theory. When we imagine mental states in madmen and Martians, we needn’t assume that they overlap with us in either physical realization or causal role. We need only imagine that they feel just as we do when we experience feelings of pain or feelings of conviction. And, worse still, Lewis’s theory has the bizarre consequence that whether someone feels pain, or feels conviction, depends on highly extrinsic facts about their membership in a species. A mad Martian feels pain, for example, if and only if they share physical realizers with other members of their own species in whom these physical states play the normal causal role of human feelings of pain.

Lewis acknowledges that these implications are deeply counterintuitive, but he sees no better option for a functionalist theory of mind. This is because such theories are forced to choose between two intuitions:

> Any broadly functionalist theory of mind is under intuitive pressure from two directions. On the one hand, it seems wrong to make it invariable or necessary that the mental states occupy their definitive causal roles. Couldn’t there be occasional exceptions . . .? On the other hand, the mental states of Karl seem intrinsic to him. Why should whether he now feels pain – or believes himself to be a fool, or desires fame – depend on what causes what in the case of someone else? I do not see any acceptable way to respect both intuitions in their full strength. (1983: 120)

As Lewis explains, functionalist theories of mind cannot respect both intuitions: they cannot allow for madmen to share our mental states without implying that mental states depend extrinsically on membership in a species. But why regard this as a forced choice? There is no problem respecting both intuitions so long as we deny that mental states are individuated by their functional roles. We regard this as a good reason to reject Lewis's analytic functionalism.

We conclude that the conceivability of mad belief in populations of Lewisian madmen constitutes a powerful challenge to Lewis's communal version of the Rationality Constraint. Moreover, this is just one manifestation of the more general problem of chauvinism that plagues all functionalist theories of mind. This is not entirely surprising, since Lewis argues that the Rationality Constraint is a consequence of his analytic functionalism. In §5, we diagnose where this argument goes wrong.

5. Against Analytic Functionalism

Our main goal in the previous section was to argue against the Rationality Constraint, which states that it's analytic that anyone (or, in its communal version, any *normal* person) who has beliefs is at least minimally rational. We argued in contrast that it's conceivable that a community of Lewisian madmen has beliefs without being even minimally rational. As we saw in §1, however, Lewis argues that the Rationality Constraint is a consequence of his analytic functionalism. Our goal in this final section is to diagnose where Lewis's argument goes wrong.

5.1. On Lewis's Argument for the Rationality Constraint

Lewis derives the Rationality Constraint from a combination of five premises. Since the argument is valid, and we reject the conclusion, we are committed to rejecting at least one of these premises. All of Lewis's premises are controversial and have well-known detractors. Nevertheless, we want to concede as much as we can so as to pinpoint what we regard as the fundamental problem in Lewis's analytic functionalism.

Lewis's first premise is his semantic thesis about the meanings of theoretical terms:

> (L1) Theoretical terms are functional terms that are implicitly defined by the theoretical roles specified in some associated theory.

Lewis's semantics for theoretical terms is highly controversial. For example, it is one of the primary targets of Kripke's (1980) critique of descriptivism in *Naming and Necessity*. Lewis and others respond to these criticisms using the apparatus of two-dimensional semantics (Chalmers 1996; Jackson 1998; Lewis 1999). However, it is beyond the scope of this paper to adjudicate these semantic issues and so we take no stand on them here.

Lewis's second premise results from his use of the semantic thesis to rehabilitate a distinction between analytic and synthetic truths:

> (L2) For any theoretical term 't' of a theory T, it is an analytic truth that if t exists, then t plays the theoretical role specified by T.

Many philosophers influenced by Quine (1951) remain skeptical about this distinction. For current purposes, however, this aspect of Lewis's position is dispensable. We can recast the Rationality Constraint in Quinean terms as the thesis that the assumption of rationality lies at the core of our theory of mind, rather than the periphery. As Lewis writes: "I do not think I need to claim analyticity: it is enough that the constraining principles should be very firmly built into our common system of belief" (1983: 112).

Lewis's third premise records his commitment to our knowledge of folk psychology:

> (L3) We all have implicit knowledge of a psychological theory – folk psychology – that we use in predicting and explaining behavior.

This premise is sometimes challenged by those who reject the *theory-theory* of mindreading in favor of the *simulation theory* (Heal 1986; Gordon 1986; Goldman 2006), On this view, our capacity for mindreading is explained in terms of imaginative simulation, rather than implicit knowledge of a psychological theory. These are sometimes regarded as competing theories of how our capacity for mindreading is implemented at the sub-personal, computational level. And yet Lewis himself makes no empirical commitments regarding the format in which our knowledge of folk psychology is represented: for example, he remains agnostic about whether it is represented in mentalese sentences in a language of thought. He assumes only that we have implicit knowledge of psychological information that we use in predicting and explaining behavior, however exactly this information is represented in the brain. For all Lewis says, this implicit knowledge might be realized by the capacity to simulate other minds (Davies and Stone 2001). We are prepared to concede this premise at least for the sake of argument.

Lewis's fifth premise records his commitment to the claim that the assumption of rationality is somehow built into the content of folk psychology:

> (L5) Folk psychology specifies the causal role of belief in such a way that anyone (or any *normal* person) with beliefs is at least minimally rational.

There is considerable plausibility to the idea that our folk-psychological practice of predicting and explaining behavior is informed by an assumption of rationality. For example, we expect someone to take an umbrella when they believe it will rain and want to stay dry because this is the rational thing to do. More generally, we assume that people will act more or less rationally on the basis of their beliefs and desires. Even so, questions remain about the status of this assumption. Why regard this as an analytic truth that is built into our psychological concepts, rather than a synthetic truth about human psychology? Or, in Quinean terms, why regard the rationality assumption as a core principle of folk psychology, rather than a peripheral one?

One influential answer is that the rationality assumption is indispensable to our capacity for predicting and explaining people's behavior in terms of their beliefs and desires. Here, for example, is Fodor's articulation of this idea as he finds it in Dennett's work:

> Unless we assume rationality we get no behavioral predictions out of belief/desire psychology since without rationality any behavior is compatible with any beliefs and desires. (1985: 80)

As Fodor notes, however, this idea is questionable. First, it's not clear *how much* rationality we need to assume in order to predict and explain behavior. It is common knowledge that we are neither perfectly rational nor wholly irrational but somewhere in between. Intuitively, however, it's not incoherent to suppose that someone with beliefs and desires could be much *more* rational or much *less* rational than we actually are. So long as we know how much rationality to expect, we can use this knowledge to predict and explain behavior. Second, we need not assume that people are rational at all in order to predict and explain their behavior in psychological terms. It's enough that we know they are *irrational* or *arational* in systematic ways that we can exploit in prediction and explanation. By assuming that people are prone to blush when embarrassed, for example, we can explain why someone goes red when he realizes that his trousers have fallen down. Third, and finally, we can understand the hypothesis that someone has certain beliefs and desires even if we cannot use this hypothesis to predict and explain their behavior. In much the same way, we can understand competing hypotheses about the origins of the universe even if we cannot decide between them on the basis of the limited empirical evidence that is available now. Scientific realists no longer subscribe to the verificationist dogma that the coherence of a hypothesis requires that it is empirically verifiable or falsifiable. So why make an exception for the coherence of a psychological hypothesis?

The extra assumption Lewis needs to derive the Rationality Constraint is that folk-psychological terms are theoretical terms that are defined by their role in folk psychology:

> (L4) All folk-psychological terms, including 'belief', are theoretical terms of folk psychology.

It is important to be clear that (L4) is not just the anodyne claim that folk-psychological terms, such as 'belief', figure in folk psychology. After all, not every term that figures within a theory is one of its theoretical terms. According to Lewis, a theory contains not only *theoretical terms* (or "T-terms") that are defined by their role in the theory, but also *old terms* (or "O-terms") that are antecedently understood. But then why should we suppose that 'belief' is a T-term, rather than an O-term? In our opinion this is where Lewis's argument goes awry.

5.2. The Myth of Our Rylean Ancestors

Why does Lewis think folk-psychological terms are theoretical terms that are defined by their role in folk psychology? He motivates this claim by appealing to Wilfrid Sellars' (1956) myth of our Rylean ancestors:

> Imagine our ancestors first speaking only of external things, stimuli, and responses . . . until some genius invented the theory of mental states, with its newly introduced T-terms, to explain the regularities among stimuli and responses. But that did not happen.

> Our common-sense psychology was never a newly invented term-introducing scientific theory – not even of prehistoric folk-science. The story that mental terms were introduced as theoretical terms is a myth. It is, in fact, Sellars' myth of our Rylean ancestors. And though it is a myth, it may be a good myth or a bad myth. It is a good myth if our names of mental states do in fact mean just what they would mean if the myth were true. I adopt the working hypothesis that it is a good myth. (1999: 258-9)

According to Sellars' myth, our understanding of what it is to have a mental state is exhausted by our knowledge of its theoretical role. This claim is plausible for many theoretical terms. We don't have any grasp of what it is to be a gluon, for instance, that goes beyond our knowledge that gluons are things that play the theoretical role specified in particle physics. But this is not nearly so plausible for folk-psychological terms. As Lewis admits, Sellars' myth is not literally true: folk-psychological terms were not introduced by the invention of a new theory. But then why should we regard this as a good myth – that is, as Lewis claims, one that captures the meaning of folk-psychological terms?

Lewis says very little to justify this assumption. In fact, he says only this:

> Many philosophers have found Rylean behaviorism at least plausible; more have found watered down, 'criteriological' behaviorism plausible. There is a strong odor of analyticity about the platitudes of common-sense psychology. The myth explains the odor of analyticity and the plausibility of behaviorism. (1999: 259)

To our minds, however, behaviorism has little to recommend it aside from discredited verificationist doctrines about meaning. Among other problems, it implies the chauvinistic result that Putnam's (1980) Super-Spartans on X-world cannot feel pain, since they are not disposed to exhibit pain behavior. In §4, we argued against the analyticity of folk psychology on similar grounds by appealing to the conceivability of mental states in Lewisian madmen. The conceivability of such cases is evidence that Sellars' myth is a bad myth: our folk-psychological terms don't mean what they would mean if the myth were true. In other words, not all folk-psychological terms are theoretical terms. Instead, we'll suggest, some folk-psychological terms are O-terms, rather than T-terms: their meaning is not exhausted by their theoretical role.

The problem with Sellars' myth is that we have something our Rylean ancestors don't have. Our understanding of the mind has a first-person dimension, as well as a third-person dimension, since we can know our own minds by means of *introspection*. Introspection gives us knowledge of the phenomenal character of experience. Moreover, this introspective knowledge doesn't depend on any antecedent knowledge of a psychological theory that we use in predicting and explaining other people's behavior. On the contrary, we have a first-personal way of knowing about our own minds that is independent from any third-personal way of knowing about other minds. Here, we follow the usual convention of treating the term 'introspection' as a placeholder for this distinctively first-personal way of knowing about our own minds (Smithies & Stoljar 2012). We remain neutral between competing theories of introspection – for example, we do not claim that introspection is a form of inner observation. When we say that some psychological terms are O-terms, we don't mean to imply that they are *observational* terms. The claim is merely that our understanding of what they mean goes beyond our knowledge of their role in folk psychology.

Our key point is that introspection gives us knowledge of the phenomenal character of experience that is not exhausted by our knowledge of any psychological theory. For example, you can know by introspection what it's like to feel pain, or what it's like to feel conviction, without knowing anything about the causal role of these feelings. Moreover, once you know what it's like to feel pain, you can understand what it's like for anyone else to feel pain – that is, to have a state that *feels like this*. In this way, introspection gives you distinctively first-personal ways of thinking about your own experience, which are sometimes called *phenomenal concepts* (Chalmers 1996).

Our possession of phenomenal concepts explains why we can imagine Lewisian madmen whose mental states feel just like ours while functioning very differently in their psychology. We can exploit our introspective knowledge of what it's like to have an experience that feels a certain way from the inside. We can know what it's like to have these feelings without relying on any antecedent knowledge of their causal role. And so we can imagine having these feelings in isolation from their normal causal role. This is how we can make sense of the hypothesis that a Lewisian madman might have the very same feelings – characterized in terms of what it's like to have them – while nevertheless playing a highly eccentric causal role (or none at all).

This is perhaps most evident for the folk-psychological terms that we use to describe our own experience. For example, there is an occurrent sense of the term 'pain' that we use to describe someone who is currently experiencing the feeling of pain. Plausibly, however, there is also a dispositional sense of the term 'pain' that we use to describe someone who is disposed to experience this feeling, although they are not feeling it right now – perhaps because they are asleep or experiencing some temporary break in an extended process of suffering (e.g. Lewis 1983: 130, n. 4).

As many philosophers have noted, we can draw a similar distinction between occurrent and dispositional senses of the term 'belief'. We can use the term in its occurrent sense to describe someone who is currently experiencing the feeling of conviction that something is true. But we can also use the term in its dispositional sense to describe someone who is disposed to experience this feeling of conviction, although they are not feeling it right now – perhaps because they are not currently thinking about the relevant topic. Arguably, we have phenomenal concepts of dispositional states as well as occurrent experiences. This is because we can think of them as dispositions to have experiences that we conceptualize under phenomenal concepts.

Our hypothesis, then, is that we have a phenomenal concept of belief as an occurrent feeling of conviction or a disposition to experience such feelings. Introspection gives us knowledge of what it's like to experience these feelings, which doesn't depend on our knowledge of their causal role. We can use this knowledge to imagine madmen in whom the disposition to feel conviction is divorced from its normal causal role. In this way, our hypothesis explains what Lewis's semantics cannot explain – namely, our ability to imagine madmen whose beliefs play eccentric causal roles. The conceivability of mad belief shows that there is more to our understanding of belief than Lewis acknowledges. Our understanding of belief is not exhausted by our knowledge of its normal causal role. We therefore deny that 'belief' is defined by its role in folk psychology in such a way that its meaning can be captured by an application of Lewis's semantics for theoretical terms.

6. Conclusion

In this paper, we examined the Rationality Constraint, according to which our concept of belief imposes limits on how much irrationality is compatible with having beliefs at all. We argued that empirical evidence of human irrationality from the psychology of reasoning and the psychopathology of delusion undermines the most demanding versions of the Rationality Constraint, which require perfect rationality as a condition for having beliefs. At the same time, however, we suggested that this evidence is consistent with more liberal versions of the Rationality Constraint, which only require meeting some minimal threshold for rationality. Nevertheless, we argued that these minimal versions of the Rationality Constraint are undermined by the conceivability of more extreme forms of irrationality that are continuous with actual cases of human irrationality. In particular, we argued that there are conceivable cases of "mad belief" in which populations of Lewisian madmen have beliefs that are not even minimally rational. This undermines the claim that our ordinary concept of belief is a theoretical concept implicitly defined by its role in folk psychology. We argued instead that introspection gives us some first-personal understanding of the concept of belief, which is not exhausted by our third-personal capacity for predicting and explaining other people's behavior.

Acknowledgements

We thank Uriah Kriegel and several anonymous referees for comments on a previous version.

References

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.). Washington, DC: Author.

Bayne, T. & E. Pacherie. (2005). In Defence of the Doxastic Conception of Delusions. *Mind & Language*, 20(2), 163–188.

Baron, J. (2008). *Thinking and Deciding* (4th ed.). Cambridge: Cambridge University Press.

Block, N. (1978). Troubles with Functionalism. *Minnesota Studies in the Philosophy of Science*, 9, 261–325.

Bortolotti, L. (2010). *Delusions and Other Irrational Beliefs.* Oxford: Oxford University Press.

Chalmers, D. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

Cherniak, C. (1986). *Minimal Rationality*. Cambridge, MA: MIT Press.

Child, W. (1994). *Causality, Interpretation, and the Mind*. Oxford: Oxford University Press.

Cohen, L. J. (1981). Can Human Irrationality be Experimentally Demonstrated? *Behavioral and Brain Sciences*, 4(3), 317–331.

Cosmides, L. & J. Tooby. (1992). Cognitive Adaptations for Social Exchange. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture* (pp. 163–228). Oxford: Oxford University Press.

Cosmides, L. & J. Tooby. (1996). Are Humans Good Intuitive Statisticians After All? Rethinking Some Conclusions from the Literature on Judgment Under Uncertainty. *Cognition*, 58, 1–73.

Currie, G. & I. Ravenscroft. (2002). *Recreative Minds: Imagination in Philosophy and Psychology*. Oxford: Oxford University Press.

Davidson, D. (1970). Mental Events. In L. Foster and J. Swanson (Eds.), *Experience and Theory* (pp. 79–102). Amherst, MA: University of Massachusetts Press.

Davies, M., M. Coltheart, R. Langdon, & N. Breen. (2001). Monothematic Delusions: Toward a Two-Factor Account. *Philosophy, Psychiatry, & Psychology*, 8(2/3), 133–158.

Davies, M. & T. Stone. (2001). Mental Simulation, Tacit Theory, and the Threat of Collapse. *Philosophical Topics* 29, 127–173.

Dennett, D. (1971). Intentional Systems. *The Journal of Philosoph*y, 68(4), 87–106.

Dijksterhuis, A. (2004). Think Different: The Merits of Unconscious Thought in Preference Development and Decision Making. *Journal of Personality and Social Psychology*, 87(5), 586–598.

Ellis, H. & A. Young. (1990). Accounting for Delusional Misidentifications. *The British Journal of Psychiatry*, 157(2), 239–248.

Evans, J. (1989). *Bias in Human Reasoning: Causes and Consequences*. Mahwah, NJ: Erlbaum.

Fodor, J. (1985). Fodor's Guide to Mental Representation: The Intelligent Auntie's Vade-Mecum. *Mind*, 94, 76–100.

Gigerenzer, G. (2006). Bounded and Rational. In R. Stainton (Ed.), *Contemporary Debates in Cognitive Science* (pp. 115–133). Oxford: Wiley-Blackwell.

Gilovich, T., D. Griffin, & D. Kahneman (Eds.). (2002). *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge: Cambridge University Press.

Goldman, A. (2006). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford: Oxford University Press.

Gordon, R. (1986). Folk Psychology as Simulation. *Mind & Language*, 1, 158–171.

Harman, G. (1986). *Change in View*. Cambridge, MA: MIT Press.

Heal, J. (1986). Replication and Functionalism. In J. Butterfield (Ed.), *Language, Mind, and Logic* (pp. 135–150). Cambridge: Cambridge University Press.

Horgan, T. & J. Tienson. (2002). The Intentionality of Phenomenology and the Phenomenology of Intentionality. In D. J. Chalmers (Ed.), *Philosophy of Mind: Classical and Contemporary Readings* (pp. 520–533). Oxford: Oxford University Press.

Jackson, F. (1998). *From Metaphysics to Ethics*. Oxford: Oxford University Press.

Kahneman, D. (2013). *Thinking, Fast and Slow*. New York: Farrar, Straus, and Giroux.

Kahneman, D. & A. Tversky. (1973). On the Psychology of Prediction. *Psychological Review*, 80(4), 237–251.

Kahneman, D., P. Slovic, & A. Tversky. (Eds.). (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge, UK: Cambridge University Press.

Kripke, S. (1980). *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Lewis, D. (1983). *Philosophical Papers, Vol. 1*. Oxford: Oxford University Press.

Lewis, D. (1999). *Papers in Metaphysics and Epistemology*. Cambridge: Cambridge University Press.

Loar, B. (1981). *Mind and Meaning*. Cambridge: Cambridge University Press.

Maher, B. (1974). Delusional Thinking and Perceptual Disorder. *Journal of Individual Psychology*, 30(1), 98–113.

McDowell, J. (1985). Functionalism and Anomalous Monism. In B. McLaughlin & E. LePore (Eds.), *Actions and Events: Perspectives on the Philosophy of Donald Davidson* (pp. 387–398). Oxford: Blackwell.

Nichols, S. & R. Samuels. (2017). Bayesian Psychology and Human Rationality. In T. Lane & T. Hung (Eds.), *Rationality: Constraints and Contexts* (pp. 17–35). London: Elsevier Academic Press.

Nichols, S. & S. Stich. (2003). *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*. Oxford: Oxford University Press.

Nisbett, R. E. (1993). *Rules for Reasoning*. Hillsdale, NJ: Erlbaum.

Pautz, A. (2013). Does Phenomenology Ground Mental Content? In U. Kriegel (Ed.), *Phenomenal Intentionality* (pp. 194–234). Oxford: Oxford University Press.

Pitt, D. (2004). The Phenomenology of Cognition or "What Is It Like to Think That P?" *Philosophy and Phenomenological Research*, *69*(1), 1–36.

Pohl, R. F. (Ed.). (2017). *Cognitive Illusions: Intriguing Phenomena in Thinking, Judgment, and Memory* (2nd ed.). Hove, UK: Psychology Press.

Putnam, H. (1975). Brains and Behavior. In H. Putnam (Ed.), *Philosophical Papers* (pp. 325–341). Cambridge: Cambridge University Press.

Quine, W. (1951). Two Dogmas of Empiricism. *Philosophical Review* 60: 20–43.

Samuels, R. & S. Stich. (2004). Rationality and Psychology. In P. Rawling and A. Mele (Eds.), *The Oxford Handbook of Rationality* (pp. 279–300). Oxford: Oxford University Press.

Samuels, R., S. Stich, & M. Bishop. (2002). Ending the Rationality Wars: How to Make Disputes about Human Rationality Disappear. In R. Elio (Ed.) *Common Sense, Reasoning, and Rationality* (pp. 236–268). Oxford: Oxford University Press.

Schwitzgebel, E. (2012). Mad Belief? *Neuroethics*, 5(1), 13–17.

Sellars, W. (1956). Empiricism and the Philosophy of Mind. *Minnesota Studies in Philosophy of Science*, 1, 253–329.

Siegel, S. & A. Byrne. (2017). Rich or Thin? In B. Nanay (Ed.), *Current Controversies in Philosophy of Perception* (pp. 59–80). New York: Routledge.

Siewert, C. (1998). *The Significance of Consciousness*. Princeton, NJ: Princeton University Press.

Sloman, S. (1996). The Empirical Case for Two Systems of Reasoning. *Psychological Bulletin*, 119, 3–22.

Smithies, D. & D. Stoljar. (2012). Introspection and Consciousness: An Overview. In D. Smithies & D. Stoljar, *Introspection and Consciousness* (pp. 3–26). Oxford: Oxford University Press.

Smithies, D. (2019). *The Epistemic Role of Consciousness*. Oxford: Oxford University Press.

Stanovich, K. (1999). *Who is Rational? Studies of Individual Differences in Reasoning.* Mahwah, MJ: Erlbaum.

Stein, E. (1996). *Without Good Reason: The Rationality Debate in Philosophy and Cognitive Science*. Oxford: Clarendon Press.

Stich, S. (1985). Could Man Be an Irrational Animal? *Synthese*, 64(1), 115–135.

Stich, S. (1990). *The Fragmentation of Reason: Preface to a Pragmatic Theory of Cognitive Evaluation*. Cambridge, MA: MIT Press.

Stone, T. & A. Young. (1997). Delusions and Brain Injury: The Philosophy and Psychology of Belief. *Mind & Language*, 12(3-4), 327–64.

Strawson, G. (1994). *Mental Reality*. Cambridge, MA: MIT Press.

Tranel, D., H. Damasio, & A. R. Damasio. (1995). Double Dissociation Between Overt and Covert Recognition. *Journal of Cognitive Neuroscience* 7: 425–32.

Tversky, A. & D. Kahneman. (1982). Judgments of and by Representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment Under Uncertainty: Heuristics and Biases* (pp. 84–98). Cambridge, UK: Cambridge University Press.

Wason, P. (1966). Reasoning. In B. M. Foss (Ed.), *New Horizons in Psychology* (pp. 135–151). Harmondsworth: Penguin.

Wedgwood, R. (2007). *The Nature of Normativity*. Oxford: Oxford University Press.

Williams, J. R. G. (2020). *The Metaphysics of Representation*. Oxford: Oxford University Press.

Young, A, I. Robertson, & D. Hellawell. (1992). Cotard Delusion after Brain Injury. Psychological Medicine 22: 799–804.