Inferential Transitions

[Penultimate version — forthcoming in *Australasian Journal of Philosophy*]

Jake Quilty-Dunn[a] and Eric Mandelbaum[b]

[a]University of Oxford; [b]The Graduate Center and Baruch College, CUNY

ABSTRACT

This paper provides a naturalistic account of inference. We posit that the core of inference is constituted by *bare inferential transitions* (BITs), transitions between discursive mental representations guided by rules built into the architecture of cognitive systems. In further developing the concept of BITs, we provide an account of what Boghossian [2014] calls 'taking', that is, the appreciation of the rule that guides an inferential transition. We argue that BITs are sufficient for implicit taking, and then, to analyse explicit taking, we posit *rich inferential transitions* (RITs), which are transitions that the subject is disposed to endorse.

KEYWORDS

inference, thought, rule-following, association

> 'And what is *thinking*? Well, don't you ever think? Can't you observe yourself and see what is going on? It should be quite simple.'
>
> —Ludwig Wittgenstein, *Philosophical Investigations*, §327

1. Methodology and the Analysis of 'Inference'

Thinking is easy to do, but difficult to understand. Often enough our minds are consumed with cascading thoughts, at least some of which are fully conscious. Yet we're hard pressed to say exactly what it is we're *doing* when we're thinking. As far as philosophy of mind and cognitive science are concerned, the question, 'What is thinking?' decomposes into at least two questions: What kinds of mental states are thoughts, and what kinds of mental transitions do those states figure in? This paper is concerned with the second question.

At least one kind of transition between thoughts—perhaps the fundamental kind—is *inference*. There are at least two main ways of approaching inquiry into the nature of inference. According to one approach, we should ask how inferential transitions successfully transmit epistemic warrant between thoughts. This epistemological approach fixes the referent of 'inference' by means of a description such as, *whatever type(s) of mental transition(s) are apt to transmit epistemic warrant from premises to conclusions*. This approach fosters interest in active, reflective, and conscious aspects of human reasoning, particularly given internalist epistemological assumptions on which consciousness, reflection, and/or personal activity are necessary for the transmission of warrant (e.g. Boghossian [2014]; Valaris [2014]).[1]

A second approach is to ask what the psychological parameters of inferential transitions are irrespective of how (or whether) they successfully or unsuccessfully transmit warrant. Those adopting this naturalistic approach seek to specify the purely descriptive features of inference as a type of logical, reason-responsive transition between mental states. Part of this project is demonstrating how inference differs experimentally from other types of transitions, such as associative transitions or noninferential computations. Cognitive scientists have been interested in such a notion of inference to account for the reason-responsiveness of propositional attitudes, which cannot be understood in purely associationist terms [De Houwer 2009; Mandelbaum 2016]. This psychological approach focuses attention on the involuntary, unconscious, and perhaps normatively degenerate aspects of inferential transitions as they're empirically shown to occur in cognition (e.g. Kornblith [2012]; for a similar but more *a priori* approach, see Richard [forthcoming]).

Nonetheless, there are few philosophers today that use empirical results to characterize inference as a natural kind. This is no accident: some philosophers in the first camp think that

[1] Siegel [2017] is an exception, in that she offers an epistemological account that allows for unconscious inference.

inference is inherently non-naturalizable [Boghossian 2014], in which case the project of situating inference in a scientific taxonomy of mental transitions looks hopeless. The data, however, may call for inferential transitions to be enumerated alongside other sorts of empirically tractable mental operations.

Our project in this paper is to sketch a naturalistic theory of inference. We aim to isolate a distinctive type of transition between thoughts by appeal to (i) intuitive paradigm cases, (ii) other cases uncovered or suggested by recent empirical data previously unmentioned in the literature, and (iii) contrasts with noninferential transitions, especially associative ones. We'll use Boghossian [2014] as a foil for our naturalistic approach. Like Boghossian, we take the basic phenomenon of inference to be the kind of mental transition typically involved in paradigm cases such as a person's movement from thinking (1) that it's raining, and (2) that if it's raining then the streets are wet, to thinking (3) that the streets are wet. Unlike Boghossian, however, we think unconscious inferences are every bit as central to an account of inference as the conscious cases.

There's a difference between merely going from the thought that *p* to the thought that *q*—via conditioned association, say, or a kind of arbitrary mental 'jogging' [Broome 2013: 226]—and undergoing an *inferential* transition from *p* to *q*, even if the latter doesn't involve conscious deliberate thought. We'll argue that the data call for a characterization of inference as a basic kind of transition between thoughts that's neither conscious, deliberate reasoning nor mere mental jogging. Such an account may not turn out to satisfy certain internalist assumptions about transmission of warrant. Indeed, trying to satisfy the internalist desiderata leads Boghossian [2014] to conclude that there's a hard problem for inference analogous to the hard problem for consciousness. Perhaps this is right, or perhaps those desiderata should be rejected. Our descriptive project is silent on warrant and other normative notions.

To telegraph where we're going, what we'll call *bare inferential transitions* are a species of non-associative, rule-governed transitions between thoughts, and what we'll call *rich inferential transitions* also involve a richer form of 'taking' [Boghossian 2014: 3ff] in the form of explicit endorsement. Throughout, we assume that the constituents of inferential transitions are structured mental representations with both syntactic and semantic properties. We'll argue in §2 that the right way to understand inference is in contrast to association, leading to the notion of bare inferential transitions. In §3, we'll develop the account further, discussing how our view handles different sorts of inference, including misinference. Then, in §4, we'll provide an account of taking that leads to the notion of rich inferential transitions.

2. Inference and Association

2.1 Dual Systems

Boghossian [2014] begins by discussing the *dual-systems* approach to cognition. 'System 1' is a cognitive system responsible for quick, automatic, unconscious and associative processes, whereas 'System 2' is slow, voluntary, conscious, and rule-based [Evans and Stanovich 2013]. Boghossian says he's interested in 'System 1.5 and up' [2014: 2], meaning a system whose operations are conscious and voluntary but possibly also quick and automatic without necessarily being effortful and attention-demanding.

One putative difference between the two systems is that transitions in System 1 are *associative* while transitions in System 2 are *rule-based*. The dual-systems claim we care about here is that fast, automatic, unconscious transitions are associative, while genuinely inferential rule-based transitions are slow, reflective, and conscious. The tendency among philosophers to focus on reflective inference as 'the Platonic Form' of inference [Boghossian 2016: 48] can lead to the idea that unconscious or automatic transitions are merely associative, an idea that's bolstered by the dual-systems approach. We question the usefulness of both this tendency and of the dual-

systems approach. Our immediate question, then, is whether inferential transitions must be slow, reflective, or conscious.

From a commonsense perspective, the answer seems to be no. Suppose you're so engrossed in Descartes' *Meditations* that the door to your apartment opens and shuts without your consciously registering it. A moment later, when you emerge from contemplation, you find yourself with the belief, MY ROOMMATE IS HOME.[2] It seems reasonable to suppose that you *inferred* that your roommate is home from the fact that someone opened the door. Perhaps this is merely a case of association between the door opening and your roommate being home. But suppose that your roommate typically spent her time at her partner's apartment, and that in fact you more often heard your door opened by your nosey landlord than by your roommate; nonetheless, on this particular occasion, you knew that your landlord was out of the country. It seems plausible that you unconsciously access the information that, since your landlord is gone, anyone who opens the door must be your roommate, and then used that information to infer that your roommate is home. This interpretation is plausible even though neither the perception that triggered the inference nor the inferential process itself were conscious, slow, or reflective.

Stepping outside of common sense, consider the phenomenon of effort justification [Aronson and Mills 1959]. Marcus endures a harsh regimen of hazing and degradation to join a fraternity. Afterwards, he feels very positive about the fraternity—indeed, even more positive than Trevor, who joined a similar fraternity but did not suffer any hazing, feels about his. The cognitive-dissonance explanation of this case is that Marcus believes that he isn't stupid; he also believes that if he underwent hazing to join a group that was less than excellent, he'd be stupid; so he concludes that the group is excellent.

[2] We follow the convention of using small caps to denote structural descriptions of concepts.

This unconscious, unreflective process is genuinely inferential. If the processes in Marcus's mind were purely associative, then given that hazing triggers negative feelings and hazing becomes associated with the fraternity, you would expect negative feelings to be associated with the fraternity. It's only because Marcus is drawing inferences from his beliefs about himself and the effort he has expended that he ends up *increasing* his positive feelings toward the fraternity. Our reasoning isn't simply that association and inference are exhaustive, so anything non-associative is inferential. This transition isn't merely non-associative. It involves rational connections between Marcus's beliefs that lead him to acquire a new belief and that instantiate a paradigmatic form of inference (namely, *modus tollens*).[3]

Some final examples from the psychological literature on reasoning might help to drive the point home. There's evidence that people make deductions without conscious awareness when those deductions map onto certain inferential forms, particularly *modus ponens* [Reverberi et al. 2012]. Subjects who are given a major premise 'If *p* then *q*' supraliminally (for 2.5s) and then given the minor premise *p* subliminally (for 50ms, flanked by masks on both sides), have the conclusion *q* facilitated. However, subjects who also saw 'If p then q' supraliminally but saw *q* subliminally, fail to have the conclusion *p* facilitated. The latter subjects, those who encounter the affirming-the-consequent form of the argument, don't have the conclusion facilitated even though the relevant concepts (in *p* and *q*) have been primed from the major premise.

Similarly, in another study subjects read *modus ponens* arguments and were either instructed to say whether the conclusion logically followed from the premises, or whether it was 'believable' in light of background knowledge [Handley et al. 2011]. Logic-based judgments were quicker and more accurate than belief-based judgments. In cases where logical validity and believability diverged (say, the conclusion that a feather is heavy following from a valid

[3] And in fact, when subjects think they're stupid (or have lowered self-esteem) effort justification ceases to change attitudes [Glass 1964], just as one would predict if a premise in deductive chain of reasoning was deleted.

argument), the conflict hampered speed and accuracy for belief-based judgments while logic-based judgments were hardly affected at all. Thus logic-based judgments occurred automatically, without interference from other cognitive processes. *Contra* some versions of dual-systems theory, humans automatically and immediately run inferences from separate statements when those inferences satisfy certain mental logical rules (see also Lea [1995]). Again, these processes aren't merely non-associative—they instantiate paradigmatic inferential rules in taking thinkers from one propositional thought to the next.

It would appear, then, that transitions between thoughts can be rule-based and non-associative—and hence genuinely inferential—without being slow, reflective, or conscious. Boghossian's 'System 1.5' is thus too high a place to start. We need a notion of inference that's rule-based and non-associative but without necessarily being conscious or voluntary. Put another way: What is the difference between fast, unconscious, automatic *associative* transitions and fast, unconscious, automatic *inferential* transitions?

2.2 Logic and Inference

Using basic principles of associationist psychology as a starting point, we rely on two key differences between inferential and associative transitions. The first difference, already mentioned, is that the former are constitutively rule-based and logic-obeying, while the latter are not. The second is that inference is reason-responsive and thus can be modified by evidence, while associations don't respond to reason and instead can only be modified in certain arational ways. We'll examine each of these differences in turn.[4]

[4] A further difference is that associative links are ideally symmetric: *ceteris paribus*, activating one associate will activate the other and vice versa. We lack space to develop this point here (although see Mandelbaum [2017]).

The first key difference is that inferential transitions are rule-based, and obey some kind of logic (whether that logic is normatively respectable, however, is an open question).[5] The inference from IT IS RAINING and IF IT IS RAINING THEN THE STREETS ARE WET to THE STREETS ARE WET operates in accordance with some logical rule, namely, *modus ponens*. When two representations are associated, however, the transition from one to the other isn't dependent on any logic. You might, perhaps through participation in some bizarre psychology experiment, come to associate the thought DONALD TRUMP IS THE AMERICAN PRESIDENT with the thought THE SUN WILL ONE DAY EXPLODE, even though there's no logical connection between these two thoughts. The transition is simply an artefact of an associative process and isn't due to the semantic or syntactic properties of the thoughts.

One way this difference can be cashed out empirically is how these kinds of transitions can be changed, which leads to the second difference: inference is responsive to evidence while association is responsive only to forms of conditioning. If you learn that bananas are in fact not yellow but red, and you have been the victim of an elaborate prank in your past experience with them, then you will, *ceteris paribus*, cease to infer that something is yellow from its being a banana. Associative transitions, by contrast, are not amenable to reason. Donald Trump's presidency bears no interesting relation to the inevitable explosion of our sun. But being rationally convinced of that fact won't suffice to break a preexisting associative link between those two thoughts. Associative links are modulated through counterconditioning and extinction. Roughly, if S associates A and B, one breaks that association not by giving good reason not to associate A and B, but by introducing A without B and B without A. A link between two concepts that

[5] Whether the logic is good old-fashioned classical logic, some non-classical logic, or rather a proprietary mental logic [Braine and O'Brien 1998] is immaterial—although we suspect that mental logic is indeed proprietary and inconsistent with wide swaths of classical logic (e.g. people don't seem to reason by the principle of explosion).

cannot be affected by any amount of extinction or counterconditioning is *ipso facto* not an associative link.[6]

Note that it's irrelevant whether there actually happen to be rich metaphysical relations between the states of affairs targeted by the associated representations; what matters is whether the structural or semantic relations between the representations is causally operative in the transitions. Suppose, for instance, that you eat a banana every morning, and spend evenings contemplating tomorrow's breakfast. After a while, you come to associate I WILL EAT A BANANA TOMORROW with I WILL EAT SOMETHING YELLOW TOMORROW, such that thinking the former causes a noninferential associative transition to the latter. Given that bananas are yellow, there's a non-arbitrary relationship between those contents, and it's originally responsible for the association. Once the associative tie is established, however, it operates independently of logic.

Finally, we say 'logic-obeying' in addition to 'rule-based' because not every rule is a logical one. One might construct a representational system that moves from IT IS RAINING to THE STREETS ARE WET as a built-in routine. This transition may be rule-based and, if it's not modifiable through extinction or counterconditioning, it's not associative. But the rule, *If IT IS RAINING then THE STREETS ARE WET*, isn't a logical rule. What distinguishes logical rules from other sorts of rules is that they abstract away from specific non-logical semantic contents and instead describe formal properties. It's famously difficult to say precisely what formality consists in (Beall and Restall [2006]: 18–26), but it seems to be a matter of the structure of representations irrespective of what things, properties, or relations the representations are about.

[6] 'Association' is a technical term whose meaning is understood against the background of associationist psychology going back to Pavlov (Mandelbaum [2017]). It also has a very loose meaning in ordinary language, which is best avoided in a careful discussion of the difference between inference, association, and other sorts of transitions. For example, we might wonder whether there's an innate 'association' between spiders and fear. But if this innate link between SPIDER and the activation of fear cannot be modulated by counterconditioning or extinction, it simply isn't an association in the sense that figures in psychology. In fact, these innate associations (such as taste aversions) were the first empirical counterexamples to associationism (e.g. Garcia and Koelling [1966]). If it turned out that no structures were modulable by counterconditioning and extinction, then we'd have to conclude that there were no associative structures.

Sticking to *modus ponens* as our paradigm case of an inferential rule, subjects infer according to it regardless of content. Thus one of our footholds into characterizing inferential transitions is not only that they're rule-based, but also that the relevant rules are logical. There are rule-based non-associative processes in early perceptual systems, for example, which aren't genuine inferences. Accounts like Kornblith's, which take 'transitions involving the interaction among representational states on the basis of their content' [2012: 55] to be inferential, thus miss out on a joint in nature between mere rule-based transitions between representations and logical-rule-based inferential transitions between thoughts.

So, in short, we need a notion of inferential transitions that captures the fact that they constitutively obey some logic and respond to reasons. These conditions can be met if we take inferential transitions to be transitions that are sensitive to the *constituent structure* of representations.

Constituent structure is an essential property of representations with a *discursive* representational format, as opposed to an *iconic* format [Fodor 2007; Quilty-Dunn 2016]. Consider the contrast between the sentence, 'Bananas are often yellow', and a picture of a yellow banana. The sentence is composed of four words; those words are the atomic semantic units of the whole representation, since they aren't themselves composed of meaningful representations. 'Bananas are often yellow' has at least one *canonical decomposition*, or right way of carving it into parts: 'Bananas', 'are', 'often yellow' (which in turn decomposes into 'often' and 'yellow'). There are many wrong ways to carve the sentence too—say, 'Bana', 'nas a', 're oft', 'en ye', 'llow'. A picture of a yellow banana, by contrast, can be carved up any way you like, and the parts that are separated will still be meaningful iconic representations. Every part of the picture represents some part of the scene, and as a result, the representation lacks a canonical decomposition. The *constituents* of the sentence are the parts that are individuated in its canonical decomposition; the picture, on the other hand, has parts but no constituents. The constituent structure of the

sentence is the structure it has in virtue of the structural relations in which the constituents stand to each other.

Some argue that mental representations do not have constituent structures. We don't intend to get embroiled in this debate here (although see, Fodor [1975]; Mandelbaum [2016]; Goodman et al. [2015]; Quilty-Dunn and Mandelbaum [ms.]). For present purposes, however, we simply appeal to the explanatory virtues of our own account. There is, we argue, a logical rule-based character to inferential transitions, and an account in terms of the constituent structure of mental representations can explain this feature of inference. Even if one is not swayed by independent evidence in favour of the structured mental representation hypothesis, we aim to motivate the hypothesis by providing a successful account of inferential transitions that presupposes it.

If transitions between thoughts are sensitive to constituent structure, those transitions must obey some logic. This is true because a logical rule just is a kind of rule which is sensitive to constituent structures. For instance, suppose that a rule of mental logic is the following: *If X is an AN, then X is an N*. Suppose, further, that you token the thought BERTHA IS A BROWN COW. You will then, *ceteris paribus*, token the thought BERTHA IS A COW. This transition is *logical* because it occurs in virtue of the fact that the constituent structure of the input representation satisfies the antecedent of the rule, and the output is generated because its constituent structure satisfies the consequent of the rule. Transitions between discursive representations that are triggered because their constituent structures instantiate some rule of mental logic thus suffice to make those transitions rule-based and logic-obeying.

Furthermore, such transitions enable the reason-responsiveness of inference. Suppose a thinker has a belief that G(x). If she acquires evidence that leads to the belief that F(x) and that if F(x) then ~G(x), then the constituent structures of the acquired beliefs are such that they

logically entail ~G(x). Assuming her mind is constructed in such a way that she infers according to *modus ponens*, these facts explain how she'll come to revise her initial belief that G(x).

2.3 Bare Inferential Transitions

The above discussion presupposes the psychological reality of logical rules that pertain to the constituent structure of discursive mental representations. There are, however, familiar issues about how such rules are psychologically realized. It must not be the case that these rules are always *explicitly* represented (meaning that they're the contents of some mental representation in the system).[7] As Lewis Carroll [1895] showed, that condition would generate a vicious regress. Instead, we propose that the basic rules of mental logic are *built into the architecture*. A rule is built into the architecture of a representational system iff whenever a mental representation is tokened that satisfies the antecedent of the rule, then, *ceteris paribus*, the system will token a representation that satisfies the consequent of the rule.

The *ceteris paribus* clause is important because there will inevitably be cases in which the transition doesn't successfully occur. Given that the rules are built into the architecture, the *ceteris paribus* clause can only be violated in certain ways. Unlike explicit rules, for example, those built into the architecture cannot have an intervening intentional state be the reason why *ceteris* isn't *paribus*. Take the practical syllogism: if you desire that Q and believe that doing P will bring about Q, then *ceteris paribus* you'll do P. A *ceteris paribus* clause here may be invoked merely because one also believes that doing P will bring about R and one desires not-R. In contrast, however, no such intentional factors could shortcut processes built into the architecture. A rule built into the architecture specifying *If P then Q* won't be shorted because of a belief that Q will lead to R and a desire that not-R. Roughly speaking, the *ceteris paribus* clause of rules built into the architecture

[7] 'Explicit' as we use it doesn't mean 'conscious'. An explicit representation is just a concrete mental token, which may be conscious or unconscious.

will be invoked only by variables at a level or more 'below' intentional psychology (such as processing constraints on memory, architectural boundaries, or neurological snafus).

Rules that are built into the architecture, therefore, are propositions that accurately describe all transitions within the relevant system in possible worlds where the relevant *ceteris paribus* clause is not violated. Our notion of rule-following is thus a functionalist one: a system's following a rule is a matter of the truth of counterfactuals that specify transitions between mental representations. We do not presuppose any more robust notion of rule-following. Any system that is constructed such that the proposition that *p* accurately describes transitions in worlds where *ceteris paribus* clauses are not violated is a system that has the rule *p* built into its architecture. Below, we characterize performance errors in a way that allows a substantive characterization of relevant *ceteris paribus* clauses. For now, what matters is that the psychological reality of rules built into the architecture of a system is a matter of how the system would move from one representation to another in cases that do not involve performance errors or otherwise violate *ceteris paribus* clauses.

Not every rule built into the architecture of some mental system is necessarily a rule of inference, nor need every rule instantiate some logical principle. We argued above that inferential transitions are essentially transitions between discursive representations in virtue of their constituent structure. Thus an inferential rule is built into the architecture iff whenever a mental representation is tokened whose constituent structure satisfies the antecedent of the rule, then, *ceteris paribus*, the system will token a representation whose constituent structure satisfies the consequent. As aforementioned, any cognitive system will exhibit ideal regularities such that token representations of one type lead, barring intervening factors, to token representations of another type. Our claim is that, when those representations have a discursive format and the regularities pertain solely to the constituent structure of those representations, then the transitions between them are inferential.

Finally, one might assume that the relevant representations must be not only discursive but fully propositional, given the reasonable assumption that inferences must operate over truth-apt (and hence propositional) representations. One might quarrel with this and argue that the transition from thinking BROWN COW to thinking COW counts as inferential despite lacking any propositional structure. We will sidestep this debate and simply use the term 'discursive', while assuming that at least the paradigm cases are fully propositional.

The foregoing furnishes us with a simple account of inferential transitions:

> **(1)** The transition from state A to state B is inferential iff (i) A and B are discursive, (ii) some rule is built into the architecture such that A satisfies its antecedent in virtue of A's constituent structure and B satisfies its consequent in virtue of B's constituent structure (*modulo* logical constants), and (iii) there is no intervening factor responsible for the transition from A to B.[8]

We'll call transitions as described in (1) *bare inferential transitions*, or BITs. We'll make three remarks on rules and BITs before moving on.

First, the notion of BITs doesn't rely on a primitivist notion of rule-following. We don't (*contra* Boghossian) take there to be some special, fundamental relation that thinkers stand in to a certain rule when inferring in line with it, nor do we think the rule must be the content of some intentional state, such as a belief. Because we reject the taking condition as necessary for BITs, it's enough for our purposes that the rule accurately describes *ceteris paribus* regularities of transitions in the cognitive system—in other words, that it's built into the architecture. BITs need only conform to the rule in a particular way. It's not sufficient for a transition to count as inferential that it can be loosely described as following some rule, since one could even

[8] We add '*modulo* logical constants' because BITs will be sensitive to elements of thoughts that aren't purely syntactic, such as IF and THEN in a conditional, negation, etc. Since logical constants can be given narrow identity conditions, and since the only semantics involved is that of logical constants, this condition doesn't undermine the formal computational character of BITs.

perversely describe associative spreading between two logically related thoughts as following a rule. Of course, this associative transition would not be inferential. It must be a feature of the cognitive system that putting in representations with one type of constituent structure will, all else equal, result in representations with another type of constituent structure. Neither the system, nor the thinker, nor any of the intentional states figuring in the transition need represent or otherwise follow the rule in some more robust sense. The rule needs only to be built into the architecture, such that representations in the system will *ceteris paribus* act in accordance with it.

Second, we'll put aside sceptical worries about rules, as canonically articulated by Kripke's Wittgenstein [Kripke 1982]. We don't see much cause for despair in the face of these worries. It's important for cognitive science to have some account of inferential transitions that are unconscious, fast, and automatic, for we know that people do have unconscious, fast, automatic transitions that aren't associative transitions but are sensitive to logical form. An account that analyses these transitions needs to be developed even if, ultimately, there will be some indeterminacy in the rules that are built into the architecture.[9]

Third, the notion of a rule being built into the architecture is not a mere appeal to dispositions. It is, crucially, a counterfactual notion: in a world in which there are no performance errors, the rule will accurately describe *every* transition within its scope. A mind can have such a rule built into it even if the rule accurately describes only a small percentage of the transitions that mind is disposed to make in the actual world due to systematic performance errors. A system can be disposed to make transitions in line with a rule without having that rule built into its architecture, and a system can have the rule built into its architecture without making transitions in line with a rule with any statistical regularity. The mere possession of the

[9] Note that there being *some* indeterminacy is much different from the more radical Kripkensteinian claim that there's no naturalistic way of delineating performance errors as opposed to following some bizarre rule.

disposition to accord with a rule is neither necessary nor sufficient for its being built into the architecture.[10]

3. Problem Cases: Semantic Entailment, Misinference, and Induction

3.1 Semantic Entailment

BITs are governed by structural features of the representations that figure in them and by the rules that are built into the architecture pertaining to those structural features. One might reasonably object that this story can only apply to *syntactic entailment*, or entailment in virtue of the syntactic (constituent) structure of premises and conclusions, not *semantic entailment*, or entailment in virtue of the contents of premises and conclusions.

An example of semantic entailment is inferring from the fact that an apple is red to the fact that it's coloured. Thus described, the transition is not due to structural features of the premises and conclusions. The structures involved are simply, X IS Y and X IS Z. The rule *If X is Y then X is Z* is clearly not built into the architecture—if it were, then predicating any property of something would cause you to predicate every other property you can represent of that thing. Instead, the contents of RED and COLOURED are semantically related such that the inference is valid even though the general schema is not. This sort of fact suggests a distinction between semantic entailment and syntactic entailment.

We think cases of semantic entailment will branch into cases of syntactic entailment (which are which are genuine inferential transitions) and noninferential associative transitions; there's no category of transition called 'semantic entailment' that's both genuinely inferential and not due to constituent structure. There can clearly be transitions from APPLES ARE RED to

[10] There's a very loose reading of 'disposition' on which our account might be dispositional simply in virtue of our appeal to counterfactual support. But on that loose reading, even an arch-representationalist and anti-dispositionalist like Fodor [1975] provides a dispositional account. We have in mind the more robust notion of disposition employed in, for example, Schwitzgebel's [2002] account of belief (cf. Quilty-Dunn and Mandelbaum [ms.]).

APPLES ARE COLOURED that are noninferential. Just as semantically unrelated thoughts like DONALD TRUMP IS THE AMERICAN PRESIDENT and THE SUN WILL ONE DAY EXPLODE can become associated, thoughts that happen to be semantically related, like APPLES ARE RED and APPLES ARE COLOURED, could become associated. For example, one might have a stored association between RED and COLOURED, so thinking APPLES ARE RED will trigger an associative transition to APPLES ARE COLOURED without being mediated by inference [Mahon and Caramazza 2003].

The question at hand, then, is: What makes it the case that a transition from APPLES ARE RED to APPLES ARE COLOURED is genuinely inferential? Intuitively, you cannot infer that apples are coloured from the fact that apples are red *unless you know that red things are coloured*, and you employ that knowledge in the transition. This intuitive requirement on inference fits comfortably with (1). What it means for you to know that red things are coloured and employ that knowledge, we assume, is in part for you to have the thought IF X IS RED THEN X IS COLOURED. So, for this semantic entailment to count as an inference, you have to think both APPLES ARE RED and IF X IS RED, THEN X IS COLOURED. These premises provide an instance of the antecedent of the rule, *If F(x), and if F(x) then G(x), then G(x).* So, at the point when both thoughts are tokened simultaneously, the structure-sensitive architecture takes over, and the system delivers the thought APPLES ARE COLOURED. This transition is simply a BIT. We hypothesize that cases of semantic entailment that are genuinely inferential transitions will involve an explicit representation of the entailment, thereby satisfying the antecedent of some rule that's built into the architecture, resulting in a BIT. One may reply that we have simply stipulated that only the BIT cases of semantic entailment, and none of the associative cases, count as inferential. But we arrived at this position by seeing what independent way there is to distinguish inferential semantic entailments from associative ones; if our proposal is incorrect, there must be some alternate account of this distinction. In the absence of such an account, we

assume that a semantic entailment is inferential only if the thinker employs her knowledge of the entailment, and thus only if it's a BIT.[11]

Two competing intuitions conspire against our proposal for how to understand inferential transitions. The first is that some cases of semantic entailment are genuinely inferential—surely one can infer that apples are coloured from the fact that they're red. The second is that going from the thought that apples are red to the thought that they're coloured doesn't always involve forming the thought that if something is red, then it's coloured. We can accommodate these intuitions by denying that the cases overlap. Some cases of semantic entailment are inferential, and some cases don't involve explicitly representing the entailment, but no cases are both. Furthermore, common sense is on our side insofar as it's common sense to think that moving from APPLES ARE RED to APPLES ARE COLOURED only counts as inferential if the thinker employs her knowledge that red things are coloured. The intuition that we can infer without representing the entailment may arise from associative transitions that develop out of repeated inferences. The transition continues to feel like an inference even though, strictly speaking, it has become an association. One might have the intuition that some semantic entailments are neither BITs nor associative transitions. However, the matter need not rest on intuition. Our theory has clear empirical commitments: mental transitions that instantiate semantic entailments will either involve explicit representation of the intervening premise (and thus be BITs), or else they will be modulable through extinction (and thus be associations).

[11] One option between the propositional and associative poles holds that the concept APPLE functions as a pointer that enables access to various predicates, such that activating the propositional structure APPLES ARE RED facilitates access to the predicate COLORED via the pointer APPLE. We lack space to develop this intriguing possibility here (but see Green and Quilty-Dunn [forthcoming]).

3.2 Misinference

A theory of inference needs to distinguish inferences from misinferences and other types of transitions. We take misinference to be a *performance error*. A person with aphasia might have competence with English but be unable to produce a sentence, and so lack the ability to linguistically perform. Competence is a standing state of a given cognitive system, the state of being disposed to operate in accordance with certain rules (such as rules that are built into its architecture). In some types of aphasia (say, Broca's), the language faculty's competence is unharmed while neurological damage prevents the competence from being manifested in linguistic performance. We understand performance errors as follows: a performance error, relative to a given system S that exhibits a particular competence, is a behaviour or mental event caused either by an intervention by another system that interrupts the normal functioning of S, or by some factor one or more psychological levels down (perhaps down to a neural level).

Not every apparent mistake will count as a genuine misinference. For instance, if one thinks IF P THEN Q and ~Q, but one has an association between ~Q and P, one might associatively activate P, even though the *modus tollens* rule mandates an inferential transition to ~P. This case would be a logically problematic transition, but would simply be a case of association, not genuine misinference as we understand it.

This notion of misinference rules out certain exotic candidate rules from being built into the architecture (cf. Kripke [1982]). For example, suppose that you typically infer in line with *modus ponens*, but being struck on the head at the right time causes you to move from thoughts of the form F(X) and IF F(X) THEN G(X) to BOB DYLAN IS JESUS. A sceptic might ask whether the rule *If F(X) and IF F(X) THEN G(X) then either G(X) or, if hit on the head in way W, then BOB DYLAN IS JESUS* is built into the architecture. In this case, the intervening factor is not explicable in psychological terms. Instead, something a level down—the neural state induced by being hit on the head a certain way—is responsible for the transition. This case thus constitutes a performance error and

thereby violates the *ceteris paribus* clause of psychological-level rules built into the architecture of the relevant cognitive system.

3.3 Induction

Inductive inferences aren't simply performance errors. So, if they aren't BITs and they aren't performance errors, how should we model them? With respect to induction and other forms of probabilistic reasoning, one possibility is that they're really BITs after all, just probabilistic ones. For what it's worth, this take appears to be the one advocated by the few descriptive models of Bayesianism out there. Take Goodman et al. [2015], who attempt to provide a descriptive account of Bayesian mental processing. Their account explicitly characterizes the role of mental representations in probabilistic reasoning. According to them, probabilistic reasoning relies on a discursive language of thought in which probability operators play a crucial role. Their account thus seems wholly compatible with our theory of inferential transitions.

We're committed to the feasibility of some account such as Goodman et al.'s on which probabilistic inference such as induction is at bottom a matter of sensitivity to constituent structure and involves some additional psychological apparatus for evaluating and computing probability in cases of probabilistic reasoning and induction. On such accounts, computations of probability operators run alongside inferential transitions between the constituent structures of propositional thoughts those operators attach to.

4. The Taking Condition

Our account thus far has focused on BITs, which are structure-sensitive transitions between discursive representations. BITs may not suffice for inference in the richer sense required by Boghossian [2014] and his interlocutors (e.g. Broome [2014]). We think BITs are the kinds of inferential transitions that matter for cognitive-scientific theorizing, but perhaps don't satisfy a

richer philosophical notion of reasoning [Harman 1986] as, among other things, a quasi-epistemological, quasi-psychological act consciously performed by a rational agent. Boghossian says genuine inference satisfies the Taking Condition, according to which a thinker must *take* the premises of his inference to support the conclusion [2014: 5]. This matters in part because, according to Boghossian, understanding inference in this richer sense requires positing a metaphysically fundamental relation of *rule-following*, 'an unanalyzable primitive' which cannot be naturalistically accounted for [2014: 17]. If one already accepts primitive rule-following, one might argue that it's all we need to account for inferential transitions generally, and so the notion of BITs won't do any important theoretical work. For our account of inferential transitions to integrate with inference in the richer sense, we have to sketch an account of the richer kind of inference and how BITs interact with it.

At the heart of the richer sense of inference is 'taking', and BITs seem to do a substantial chunk of the descriptive work that taking is invoked to do. If a rule is built into someone's mental architecture, they needn't explicitly represent the rule nor stand in any attitudinal relation to it. That's just as well, for even according to Boghossian taking doesn't involve explicit representation [2014: 14ff]. Perhaps taking involves some *implicit* appreciation of the rule that's nonetheless more robust than mere accordance with the rule, and is thus more robust than mere mental 'jogging' [Broome 2013: 225] from premises to conclusions. A rule's being built into the architecture of a thinker's central cognitive system seems sufficient for that thinker to implicitly appreciate that rule. A BIT that instantiates *modus ponens* is not a case of mere accordance. One might associatively move from two thoughts to a third thought, and the thoughts might happen to be of the form P, IF P THEN Q, and Q, respectively. But in that case, the *modus ponens* rule was not really involved; the associative transition at most merely accorded with it. In the case of the BIT between the same thoughts, it's precisely because *modus ponens* is built into the architecture of the thinker's cognitive system that she draws the conclusion. The rule thus plays a direct role

in producing the conclusion. We aren't sure whether this suffices for full-blooded 'rule following'. BITs nevertheless explain 'how such a rule could *guide* a person's inferences' [Boghossian 2014: 13; emphasis his] while explaining why inferential rules don't guide associations. The core descriptive function of taking—allowing rules to guide inferences directly and thus distinguish them from noninferential transitions like associations—is performed by BITs.[12]

Other psychological factors considered in the recent inference literature seem more obviously to go beyond mere BITs. In particular, there does seem to be a difference between transitions that merely happen and those one endorses. Imagine a logician who propounds a radical logic on which *modus ponens* is invalid. Suppose, nonetheless, that the *modus ponens* rule is still built into the architecture of her central cognitive system. Thus, when she tokens the thought that apples are red, and the thought that if something is red then it's coloured, she'll trigger a BIT to the thought that apples are coloured. While she might endorse the premises and the conclusion, she'd reject the transition itself because of her views about logic. There seems to be a real psychological difference between her case and the case of someone who undergoes the same transition while accepting the validity of *modus ponens*. The difference isn't in their premises or conclusions (and we can stipulate that both transitions are unconscious). The difference seems to be that the second person is disposed to endorse *the inferential transition itself*, while the radical logician is not. This seems like a meaningful and important sense in which the one person takes the premises to support the conclusion while the other doesn't. Satisfying the Taking Condition in a richer sense, then, is being disposed to endorse the inferential transition.

[12] In addition to the question of whether the rules involved in inferential transitions are followed, there's the question of whether the rules are followed *by the thinker*, or are followed by the cognitive system. We intend talk of rules being built into the architecture to be neutral on whether the rule is thereby followed by the thinker or just a subsystem of the thinker (or even, as noted, whether it's 'followed' in some more robust sense at all). Perhaps, since thinking is something the agent does even when it's involuntary (e.g. *you* are involuntarily thinking about polar bears now that we've mentioned them), BITs are things agents do despite their being architectural. We lack the space to pursue this question here.

We think that the best way to understand the disposition to endorse an inferential transition is as follows:

(2) A person is disposed to endorse an inferential transition from F(x) and If F(x), then G(x) to G(x) iff she is disposed to form the thought F(X) THEREFORE G(X).

What's doing the heavy lifting in (2) is the THEREFORE concept (see also Neta [2013]). We think there must be some concept that represents a relation of support between facts without being merely conditional. The difference between thinking F(X) THEREFORE G(X) and thinking IF F(X) THEN G(X) is that the former isn't neutral on the truth of the individual propositions. It's represented as being the case that F(x), and that its being the case that F(x) provides support (of some kind) for its being the case that G(x). One might think that if this apple is red, then it's coloured; but one might also think that the apple is red, so it's coloured. Presumably people have a THEREFORE concept, which yokes together facts by representing one as implying the other. The 'therefore' thought entails the conditional thought, but differs from it in not being merely hypothetical. Call inferential transitions that the person is disposed to endorse (in line with (2)) *rich inferential transitions* (RITs). While the taking afforded by BITs is implicit taking, the richer form of taking afforded by RITs is *explicit* taking.

Note that neither BITs nor RITs need be conscious. We see this as a virtue of the present account. For one thing, if RITs do special epistemic work, then the fact that they need not be conscious makes them compatible with epistemological theories that allow unconscious mental states and processes to do epistemic work (e.g. Siegel [2017]). And while some aspects of inference are surely intentional—deciding what to think about, how much attention to devote, whether to allow your mind to wander—our account entails that the actual inferential transitions from premise thoughts to conclusion thoughts are determined by the cognitive architecture, and as such aren't intentional acts. This seems to us to be intuitively the right answer. It's hard to imagine a mind like ours that could activate thoughts of the form *p* and *if* p *then* q, attend to them

without distraction (and lack relevant neuropathology, resource depletion, emotional manipulation etc.), and yet decide not to draw the conclusion that *q*.

5. Conclusion

The notion of inference is foundational in cognitive science and philosophy, so it's refreshing to see a revival of philosophical analyses of inference. However, we think that recent analyses of inference have been overly intellectualized, and thus obscure the vast role that inference plays in our cognitive economy (particularly, in unconscious cognition). We have tried to do justice to the notion of inference by highlighting the dual roles it plays. On the one hand, it's a central and distinctive way we move between thoughts, which is often completely unconscious, involuntary, and understood in contrast to associative transitions. On the other hand, it can be understood as a process that involves explicitly taking a set of premises as reasons for believing some conclusion.

We haven't accounted for inferential transmission of warrant. We believe, however, that any such story should proceed on the basis of a naturalistic descriptive account of psychological reality rather than allowing *a priori* epistemology to dictate philosophy of mind. There's a deep methodological divide in philosophical theorizing between those who take normative factors to be primary and those who think descriptive questions must be answered before we can see which norms actually constrain us and which turn out to be beyond our limited capacity. Fodor asks, in a similar vein, how one could be 'bound by norms that one is, in point of nomological necessity, unable to satisfy?' [Fodor 2007: 115]. We don't make the bold statement here that internalist theories of inferential warrant transmission are false or incompatible with our theory. Perhaps there's a way of squaring BITs with internalism, or perhaps there are other mental capacities that fit better with internalist construals of epistemic norms. But we don't think any descriptive account, particularly one committed to naturalistic intelligibility and explanatory

continuity with cognitive science, should be thrown out simply on grounds of lack of fit with *a priori* epistemological theories.

Descriptively speaking, there are many aspects of conscious reasoning that our account doesn't explain. Thinking through a problem, even in mundane circumstances, involves many complex operations, such as: searching through vast stores of thoughts and concepts, evaluating probabilities, drawing analogies, imagining various states of affairs in both propositional and sensory imagination, activating some thoughts and concepts and not others, integrating desires and goals with beliefs, activating episodic memories, and doubtless other kinds of mental processes including some yet to be named. Some of these processes are partly understood, and others aren't understood at all. Some may be conscious intentional acts, and others may be unconscious automated processes. Conscious human reasoning is a variegated 'mental chaos' [Siegel 2017: 99] that integrates many distinct processes in ways that are opaque to introspection. Providing a complete account of conscious human reasoning is an extremely ambitious project of which an analysis of inference is only one part. Nonetheless, inferential transitions between thoughts are a central facet of human thinking and, we suspect, a central part of what makes for rational thought. Despite our failing to fully capture the mental chaos of conscious reasoning, we have given an account of the core of inference, one that separates inference from other sorts of transitions and respects its essentially logical nature.

Most of all, we think we have accurately described the phenomenon while also responding to a deep worry Boghossian poses. He writes, 'If the present account of reasoning is along the right lines, it opens up the possibility that reasoning poses as much of a challenge to a naturalistic worldview as does consciousness. It makes it difficult to see what naturalistic process inference could consist in' [Boghossian 2014: 18]. By introducing the BIT notion of inferential transitions, we have aimed to secure the groundwork for a naturalistic theory of inference, one

which helps illuminate how reasoning works without adding to our hard problems.[13,14]

REFERENCES

Aronson, E. and J. Mills 1959. The Effect of Severity of Initiation on Liking for a Group, *Journal of Abnormal and Social Psychology* 59/2: 177–81.

Beall, J.C. and G. Restall 2006. *Logical Pluralism*, Oxford: Oxford University Press.

Boghossian, P. 2014. What Is Inference? *Philosophical Studies* 169/1: 1–18.

Boghossian, P. 2016. Reasoning and Reflection: A Reply to Kornblith, *Analysis* 76/1: 41–54.

Braine, M.D.S. and D.P. O'Brien, eds, 1998. *Mental Logic*, Mahwah, NJ: Erlbaum.

Broome, J. 2013. *Rationality through Reasoning*, Sussex: Wiley Blackwell.

Broome, J. 2014. Comments on Boghossian, *Philosophical Studies* 1691: 19–25.

Carroll, L. 1895. What the Tortoise Said to Achilles, *Mind* 4/14: 278–80.

De Houwer, J. 2009. The Propositional Approach to Associative Learning as an Alternative for Association Formation Models, *Learning & Behavior* 37/1: 1–20.

Evans, J.St.B.T. and K.E. Stanovich 2013. Dual-process Theories of Higher Cognition: Advancing the Debate, *Perspectives on Psychological Science* 8/3: 223–41.

Fodor, J. 1975. *The Language of Thought*, Cambridge, MA: MIT Press.

Fodor, J. 2007. The Revenge of the Given, in *Contemporary Debates in Philosophy of Mind*, ed. B. McLaughlin and J. Cohen, Oxford: Blackwell: 105–16.

Garcia, J. and R.A. Koelling 1966. Relation of Cue to Consequence in Avoidance Learning, *Psychonomic Science 4*/1: 123–4.

[13] Helpful comments and criticisms were offered by Tim Bayne, Joseph Bendaña, Jake Berger, Paul Boghossian, Ryan DeChant, Zoe Jenkin, David Papineau, Jesse Rappaport, David Ripley, Susanna Siegel, and the NYU Consciousness discussion group. Special thanks to the editor and referees at AJP for their useful and open-minded comments.

[14] This paper was conceived and executed while JQD and EM were on EM's PSC-CUNY Award 67331-00 45. PSC CUNY is hereby thanked for their largesse.

Glass, D.C. 1964. Changes in Liking as a Means of Reducing Cognitive Discrepancies between Self-esteem and Aggression, *Journal of Personality* 32/4: 531–49.

Goodman, N., Tenenbaum, J. and T. Gerstenberg 2015. Concepts in a Probabilistic Language of Thought, in *The Conceptual Mind: New Directions in the Study of Concepts*, ed. E. Margolis and S. Laurence, Cambridge, MA: MIT Press: 623–54.

Green, E.J. and J. Quilty-Dunn Forthcoming. What Is an Object File? *British Journal for the Philosophy of Science*.

Handley, S., Newstead, S. and D. Trippas 2011. Logic, Beliefs, and Instruction: A Test of the Default Interventionist Account of Belief Bias, *Journal of Experimental Psychology: Learning, Memory, and Cognition* 37/1: 28–43.

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

Kripke, S. 1982. *Wittgenstein on Rules and Private Language*, Cambridge, MA: Harvard University Press.

Kornblith, H. 2012. *On Reflection*, Oxford: Oxford University Press.

Lea, R.B. 1995. On-line Evidence for Elaborative Logical Inferences in Text, *Journal of Experimental Psychology: Learning, Memory, and Cognition* 21/6: 1469–82.

Mandelbaum, E. 2017. Associationist Theories of Thought, *The Stanford Encyclopedia of Philosophy* (Summer 2017 edition), ed. E.N. Zalta, URL = <https://plato.stanford.edu/archives/sum2017/entries/associationist-thought/>.

Mandelbaum, E. 2016. Attitude, Inference, and Association: On the Propositional Structure of Implicit Bias, *Noûs* 50/3: 629–58.

Mahon, B. and A. Caramazza 2003. Constraining Questions about the Organisation and Representation of Conceptual Knowledge, *Cognitive Neuropsychology* 20: 433–50.

Neta, R. 2013. What Is an Inference? *Philosophical Issues* 23/1: 388–407.

Quilty-Dunn, J. 2016. Iconicity and the Format of Perception, *Journal of Consciousness Studies*

23/3–4: 255–63.

Quilty-Dunn, J. and E. Mandelbaum ms. Against Dispositionalism: Belief in Cognitive Science.

Reverberi, C., Pischedda, D., Burigo, M. and P. Cherubini 2012. Deduction Without Awareness, *Acta Psychologica* 139/1: 244–53.

Richard, M. Forthcoming. Is Reasoning a Form of Agency? in *Reasoning: Essays on Theoretical and Practical Thinking*, ed. M. Balcerak-Jackson and B. Balcerak-Jackson, Oxford: Oxford University Press.

Schwitzgebel, E. 2002. A Phenomenal, Dispositional Account of Belief, *Noûs* 36/2: 249–75.

Siegel, S. 2017. *The Rationality of Perception*, New York: Oxford University Press.

Valaris, M. 2014. Reasoning and Regress, *Mind* 123/489: 101–27.

Wittgenstein, L. 1953 (1958). *Philosophical Investigations*, tr. G.E.M. Anscombe, 2nd edn, Oxford: Basil Blackwell.