Carlos Muñoz-Suárez • Felipe De Brigard
Editors

Content and Consciousness Revisited
With Replies by Daniel Dennett
Chapter 11
Not Just a Fine Trip Down Memory Lane: Comments on the Essays on Content and Consciousness

Daniel Dennett

Abstract The current chapter contains commentaries and replies to all nine essays included in the present volume.

It has been more than a pleasure to read and reflect on these thoughtful, constructive essays, which have taught me a lot about my own work—always a bracing experience—and pointed to future directions well worth exploring further. I think what makes me most proud of my firstborn book (published when I was 27-years-old) is that 45 years later it can still provoke high caliber work like this. These essays are not backward-facing nostalgic reflections on an antique book of merely historical interest but forward-looking appropriation and exploitation of ideas that are useful, their authors think, on the cutting edge today. It is never gracious to say “I told you so,” but sometimes the urge to say it is strong, so it is particularly gratifying to have these excellent philosophers say it for me, in nine different ways. I don’t agree with everything they have to say, but where we still disagree, I may well be missing something they understand better than I do.

1
If Don Ross is right, it is possible to make a major scientific discovery without trying, and without recognizing that you’ve done so. I don’t view that conditional as an invitation to perform modus tollens. It is quite possible to stumble into something more important than you realize at the time, and Ross suggests as much:

It is instructive that of all the articulations of his theory of the mind that Dennett has produced over the course of his career, the one most strongly based in traditional philosophical analysis and argumentation got the science of the story right in all its essentials – as judged against both Dennett's later opinions and what has been implicitly endorsed by later scientific practice – while providing an unstable and unsatisfactory account of the metaphysics. (p. 37)
I “got the science right” almost by accident, while engaging in “traditional philosophical analysis and argumentation.” And the metaphysics has been “unstable.” Fair enough. My initial forays into the scientific literature were the efforts of an utter novice with no scientific training, but it does seem that I had a knack for finding the best scientific beacons—and interpreting them in a philosophically novel way. And, truth to tell, at the time I was working in very solitary fashion on my dissertation and later on its descendant, C&C, it did seem to me that I was opening up some productive new vistas, that I had found a way out of some perennial philosophical traps by taking seriously some ideas I had developed about how to understand the scientific project. But that hunch also often seemed too good to be true. I worried that I must be missing something that others understood, something that explained why what seemed to me like exciting new ideas were, in the end, forlorn. And to be sure, there are quite a few eminent philosophers who are utterly confident that I am missing something, and have been saying so for years, but at this point, I’m much more sure of my ground. From my perspective, C&C was a lucky strike of a mother lode of ideas that, with a little refining, form pieces that snap together in a very productive way. So I think Ross has nicely uncovered and “celebrated,” as he says, something of a discovery that I more-or-less made back in the 60s and have been trying to understand better and defend ever since. As he notes, I have not always had the best version of my position in focus, and over the years he has intervened on occasion to help me get back on the right track, as he sees it. I don’t always agree with his proposed improvements, some of which I may not understand, but since he’s often been dead right, I take all his suggestions seriously.

Ross speaks of the “over-reaction against behaviorism,” which was certainly a hallmark of the early days of cognitive science, largely inspired by Chomsky’s caricature of Skinner’s *Verbal Behavior* (1957). He points out, correctly, that I didn’t join in the funeral festivities for behaviorism. (I vividly remember a funny talk at an early cognitive science conference in Minneapolis where one of the psychologists—I wish I could remember his name—gave a talk modeled or the forced public “concessions” of falttering Maoists or Castroites: he admitted that he had “committed acts of behaviorism” in his errant youth. He was lampooning the anti-behaviorist fervor of most of the audience, whether or not they realized it—and many didn’t think it was funny at all.) Ross then laments my “Skinnerian” (1981a as backsliding. I protest: a close look at that paper, which I still endorse, shows that it is explicitly an attempt to “avoid the familiar brawl and do something diagnostic” (p. 54), exposing Skinner’s combination of doctrinaire overstatement and ineffective waffling, and isolating his—and Quine’s—distaste for intentional idioms. It should be borne in mind that the next chapter in *Brainstorms*, where “Skinner Skinned” was published, is “Why the Law of Effect Will Not Go Away.” I have earned my badge as a circumspect friend of behaviorism, as many philosophers of mind who still don’t get it gleefully insist. (For them, “behaviorist” is a term of abuse. We are well rid of the crude doctrine Jack Vickers once called “barefoot behaviorism,” but what these philosophers don’t realize, apparently, is that their headlong dash away from all hints of behaviorism lands them in the Nagel-Chalmers cul-de-sac.)

I will drag my feet on one point here. Ross is less than satisfied with my proposed role for philosophers in “negotiating the traffic back and forth” between the manifest and the scientific image. My sense of the point of this project is informed by Sellars’ famous definition of philosophy:

> The aim of philosophy, abstractly formulated, is to understand how things in the broadest possible sense of the term hang together in the broadest possible sense of the term. (1963, p. 1)

Ross says:

> whatever services reconciliation of the manifest and scientific images might render for political and economic support of science, it tends to interfere with the epistemic progress of science. It has this effect because it encourages proliferation of analogies between scientific and folk ontologies, which invariably ‘domesticates’ the former in the sense of blunting their most radical implications for further conceptual revisions that in turn open roads to new experiments and new mathematical and statistical tools. (p. 41)

Maybe so, on occasion, but I think Ross overestimates the scientists here; they have neither the philosophical prowess nor the invulnerability to massive confusion that this laissez-faire policy presumes. The exercise of finding explanatory paths between scientific and folk ontologies is not always just for the enlightenment of lay audiences; sometimes it is an important reality check (and I use the idiom advisedly) for the scientists and their schemes. We’re all in Neurath’s boat together, and we have to keep it floating.

2.

The main virtue of Felipe De Brigard’s essay, it seems to me, is a feature it shares with Wilfrid Sellars at his best—not so much the published Sellars as the lecturing Sellars. Wilfrid was a virtuoso blackboard artist, diagramming the logical geography of all the different available positions and showing how, if you adopted his perspective, you could see more clearly the bearing of, say, van Fraassen’s construcrive empiricism, and Millikan’s work, and Jackendoff’s, and Churchland’s, and Azzouni’s and... on the travails of the propositional attitude task force and my problematic part in it. I’ve learned a lot about the various battlefields I have traversed—and in some regards anticipated and steered clear of—from De Brigard’s synoptic but detail-rich account.

After all, folk psychology is just another theory—unrefined if you want, and operational over a slightly different domain than scientific psychology—but a theory none-the-less. (p. 65)

I wish De Brigard hadn’t said that. Or rather, I wish he’d also highlighted the prospect, long urged by me, that part of the continuing confusion on these topics might be the still ubiquitous and unrefeective practice of thinking of folk psychology as a *theory*—an assumption perfectly articulated in this sentence. It is this tempting assumption (“What else could it be?”) that enables the whole Churchland vs. Fodor dispute so aptly diagnosed here, as well as the theory-theory vs. simulation theory literature, and much else. Sellars’ (1956) Jones, the mythical deviser of a theory of mental involving variables, probably put the theory-theory idea on the stage, but it was I who introduced the term, “folk psychology” in “Three Kinds of Intentional Psychology” (1981b), and from the outset I was on the warpath against treating it as just like a scientific theory. (Well, then, what is it? It’s a strategy of interpretation.)
De Brigard shows how the passages he quotes from C&C prefigure that campaign; I specifically warned against jumping to ontological conclusions before we tackle the scientific questions, and this caution has been a theme of mine—most effectively, I think, in “Real Patterns” (1991c)—ever since.

In C&C, in Chap. 2, “Intentionality,” I rather unconvincingly sided by the problems that De Brigard exposes. At that time, I had been persuaded by the opening gambits of Quine and Chisholm that theoretical clarity could be achieved by re-expressing casual mentalistic talk in propositional-attitude formulae. As the years rolled by, it became more and more obvious to me that this was a forlorn quest. Steve Stich and I, on Fullbright together to Bristol in 1978, decided to write a co-authored paper diagnosing the systematic pathologies of the propositional attitude bandwagon, then in full swing, but we couldn’t agree on where to take this shared conviction, so we went our separate ways. I gave it my best shot by writing “Beyond Belief” (1982), by far the hardest philosophical task I have ever completed, and he wrote a whole book, From Folk Psychology to Cognitive Science: The Case Against Belief (1983). I considered myself to have paid my dues and shifted my attention to other topics. (Put Churchland encouraged this decision by dubbing my long, long essay, “Beyond Belief and Past Caring.”) Not surprisingly, in retrospect, the propositional attitude task force shrugged off both Stich’s diagnosis and mine, and turned back to their research program, which apparently persisted to this day. Every decade or so I return briefly to the literature to see if any progress has been made, and De Brigard’s critique of the current state of play convinces me, yet again, that they are still spinning the same old wheels. Of course it may be that they saw back in 1981 that my objections were so off target that there was no call to refute them, but over 30 years with no consensus results makes me suspect that they just didn’t want to stop playing the game they had mastered.

The dismantling by De Brigard of Paul Churchland’s arguments against folk psychology has a few innovations worth noting.

Blaming the entire apparatus of folk psychology on the basis of just one failure seems a bit exaggerated. For one, I can provide an explanation of the failure in terms of the very same theory: if you hadn’t forgotten the date, my prediction would have worked just fine. Secondly, it is true that similar extrapolations have proved successful in the past (last Wednesday—remember?—you did actually make it to our appointment). (p. 55)

As De Brigard goes on to note, folk psychology has the resources to identify what sorts of interventions would make forgetting more or less likely. Even if we concede for the moment that folk psychology is enough like a theory to be called a theory, there is nothing circular or vacuous about a theory being able to explain and predict the circumstances under which its predictions tend to be false. In practice, in the actual time-pressured world of likelihoods, we not only tolerate but welcome probabilistic predictions (of the weather, of the effects of medicine on our bodies, of weekend traffic jams, and many other things that matter). In such cases, it goes without saying that had the theorist gathered much more data a more accurate prediction could have been made. Ceteris paribus clauses abound, as De Brigard notes, and we don’t in general disparage theories that rely on them. I don’t recall this point being made before.

Folk psychology is necessary, as De Brigard notes, for devising and executing research programs—as I show in detail in my account of heterophenomenology (1987a, 1991, and elsewhere). In C&C I didn’t feel the need (in the decadent, waning years of ordinary language philosophy) to support the claim that we can’t do without intentional idioms (and the personal/sub-personal distinction—see Frankish, Wilkinson and Roth and my comments on them). Several waves of subsequent eliminativism have tried to persuade us that we are well rid of them, but De Brigard show what a draconian program that would be.

3

Keith Frankish looks at the chapters of C&C on thinking and reasoning, and finds previews there of the current enthusiasm in psychology for dual-process theories of thinking, now a major research industry in psychology, most recently made famous outside the academy by Danny Kahneman’s (2011) distinction between Type 1 and Type 2 thinking.

I appreciate the restraint of Frankish’s constructive and sympathetic account. He doesn’t claim that, if only I’d switched my awareness subscriptions, it would be obvious that I was the father of dual process theory. As he notes, dual-process theories have been independently invented over and over, a Good Trick with a large basin of attraction, and my particular version was underdeveloped, a philosopher’s semi-informed surmise rather than a specifically worked out theory, but at least my reflections were carrying me in the right direction, which is more than can be said for many rival gestures in the direction of theory from other philosophers. What is much more interesting than any priority claim is whether my work, in C&C and later, has anything specific to offer to current theory, and here Frankish points to my sketch of an account of how Type 2 thinking gets installed in human brains. That is a big question today, and I do have an inchoate account—the outlines of which Frankish discerns nicely—and am currently working on a considerably more ambitious, detailed, empirically supported, theory.

There is a big problem, which psychologists have been uneven in discerning and more uneven in tackling:

Type 2 processing seems capable of some prodigious intellectual feats. Indeed, it seems to occupy the role of something rather like a central executive, which can override instinctive, automatic, and emotional responses with rational thoughts and decisions. Now, the posit-

ing of such an executive system is, of course, a move which Dennett opposes, as being both unexplainatory and neurologically implausible—a central theme of Consciousness Explained (Dennett 1991a). (p. 76)

How can you account for the powers of Type 2 thinking without installing an ominously clever res cogitans to do the symbol-manipulating? My answer, in short: by recognizing that Type 2 thinking is a learned, culturally borne, personal-level activity. As Frankish notes, another problem facing dual-process theorists is to explain how it evolved so swiftly, and, perhaps, only in H. sapiens. And yet another is to explain its relationship to Type 1 thinking. How does Type 2 thinking exploit the resources of the mammalian (or more particularly, primate) brain? Frankish presents his answers to these questions, with the help of his interpretation of my own views. This is the best kind of value-added criticism, and I agree right down the
line with his suggestions, and will just highlight a few points here that strike me as deserving extra emphasis.

Dennett’s main concern in this chapter is [...] to argue that talk of thinking or reasoning is often simply an idealized intentional characterization of sub-personal information processing operations of which we have no conscious awareness. (p. 79)

Here’s one more way of thinking about it: don’t make the mistake of treating Type 2 thinking as providing a process model for Type 1 thinking—for the same reason we shouldn’t view the Type 2 thinking by engineer/designers as the process model for the processes of natural selection! (See my commentary on Dub) Type 1 thinking is fast, parallel, etc., etc. and very good at homing in on excellent results—it gets clever animals through their challenging lives with grace and reliability. When we adopt the intentional stance towards animals in order to explain and predict their behavior, we often treat them as if they were Type 2 thinkers, but that is a crust for the imagination that should not be seen as committing us to a process model of their Type 1 thinking. Type 2 thinking is a recent add-on, dependent on language, mainly because talking to others and talking to yourself are personal-level actions that play essential roles in installing Type 2 thinking in each of us. Much of Type 2 thinking is in fact talking to yourself—in your native language, not in Mentalese or a “language of thought”—but not all of it is. There is wordless, imagistic (auditory, proprioceptive, tactile, not just visual) exploration that is accomplished by a sort of auto-Socratic method: posing “questions” and seeing what your Type 1 resources can come up with for responses. It is a kind of self-stimulation, in other words, that becomes as “second-nature” as, well, talking to others, some of which is deeply purposed and monitored and some of which is just idle yakking, as noted in Mose Allison’s wonderful song, “Your mind is on vacation but your mouth is working overtime.” Our internal personal-level activity (cf. Ryle’s attempt to answer the question “What is Rodin’s Thinker doing?”) includes not just the hard work of Type 2 thinking but every less disciplined variety of daydreaming and woolgathering.

The Socratic method is in effect an externalization of the private practice of reflection, an exercise in group reflection. In a wonderful passage in the Theaetetus, Plato draws attention to a deep epistemological problem:

Socrates: Now consider whether knowledge is a thing you can possess in that way without having it about you, like a man who has caught some wild birds—pigeons or what not—and keeps them in an aviary he has made for them at home. In a sense, of course, we might say he “has” them all the time inasmuch as he possesses them, mightn’t we?

Theaetetus: Yes.

Socrates: But in another sense he “has” none of them, though he has got control of them, now that he has made them captive in an enclosure of his own; he can take and have hold of them whenever he likes by catching any bird he chooses, and let them go again; and it is open to him to do that as often as he pleases. [Theaetetus, trans. Francis M. Cornford (New York: Macmillan, 1957), 197 C D]

Plato saw that merely possessing knowledge (like birds in an aviary) is not enough; you must be able to get the right birds in your aviary to come when you call. Techniques of self-stimulation designed (unwittingly) to give you access to your own (Type 1-embedded and Type 2) knowledge are the great innovations of Type 2 thinking, and they are thinking tools that must, in the main, be installed—though the installation process, like the acquisition of one’s native tongue, has been made easier over the generations by a Baldwin Effect interaction between fast-evolving cultural items and more slowly evolving genetically transmitted design improvements that enhance our ability to use these tools.

It is not just that installed thinking tools give you access to your own embedded knowledge; their installation is what creates the phenomenon of access in the first place, by creating the problem of access—and its solution! Type 1 thinking happens automatically, for better or worse, and whatever tracts in the brain get activated do—or fail to do—the right thing. There is no issue of trying to get the right birds to come when you call. But once one has acquired the habit of auto-Socratic exploration there is always the prospect of learning how better to remind yourself of what turns out, on reflection, to have been important. “Next time, it would help if I accessed what I know about X before I decide.”

What could guide and control such a process? Frankish has some ideas about this: Similarly, self-directed speech acts might be generated pandemonium-style, without antecedent calculation of their structure or likely effects. It is true that, if they are to count as intentional, self-stimulations must be susceptible to some intentional characterization, but this need not be in terms of desires for specific cognitive and behavioural effects and beliefs about how to achieve them. The motivating states might simply be a desire to solve some problem and the instrumental belief that doing this (uttering the words that spring to one’s lips) may help.

But could pandemonium processes generate the subtle self-stimulations required to support executive control, abstract problem solving, and hypothetical thinking? Where does the intelligence in these acts come from? (p. 84)

He answers his own question. First, many self-stimulations are not particularly intelligent—“chance associations, whimsy, free-wheeling speculations, and so on”—just the sort of hopeful rubbish a pandemonium process would often generate; second, "Self-generated speech and other imagery may not only stimulate cognitive and affective responses, but also trigger further acts of self-stimulation, shaped by those responses” creating “cycles of self-stimulation” that are themselves creative, and

Third, self-stimulation may be guided by knowledge imparted by culture. Cultural processes may disseminate, not only the trick of self-stimulation itself, but specific applications of it to particular problems. (p. 84)

I think this third point is the key that unlocks the mystery of the “prodigious” power of human thought: in the same way that genetic evolution by natural selection copies and copies the tiny design improvements discerned in each generation, cultural evolution by natural selection (over the last few hundred thousand years) has bench-tested and approved hundreds or thousands of thinking habits and disseminated them widely, in turn creating a huge selection pressure at the genetic level for brains that are good at installing and using these habits. Type 2 thinking is a product of meme-gene coevolution in much the same way lactose tolerance in adulthood is. Dairying is a culturally transmitted practice that creates
selection pressure for lactose tolerance, and auto-Socratic exploration is a culturally transmitted practice that creates selection pressure for brain structures and dispositions that can make the most of these habits.

Once personal-level Type 2 thinking has established itself as the prevailing activity among human beings, further cultural evolution can create whole new phenomena, unknown in the animal world, to exploit these habits. Humor is one of the most distinctive, and it depends on controlling the timing of conscious access (Hurley et al. 2011); a punch line telegraphed loses its punch.

4 Richard Dub baffle me at first, in a very useful way. How could he know the relevant literature so well (not just my work, but Davidson, Lewis, Quine, Ryle, Stich, ...) and still not "get it"? Since he lays out the issues better than anyone I’ve ever encountered, I conclude that I am probably the one who is missing something. What? I’m resisting the temptation to just "say it again, louder," and have been castigating for a new way to make my points.

My first hunch is that Dub has underestimated how radical my claim about the rationality assumption is—perhaps out of misplaced charity since he alludes to the incredulity that some philosophers have expressed. His distinction between individual ascription and scientific ascription allows him to contrast the time-pressured quick-and-dirty attributions of folk psychology with the measured, theoretical posit of scientific psychology, and this allows him to turn my constraint-on-attribute into an empirical discovery. This, he thinks, saves the best features of the intentional stance minus the incredible rationality constraint.

If we interpret agents as rational because we are led to do so by scientific norms of predictiveness, systematization, and empirical adequacy, then rationality need not be a constraint on interpretation, nor need it play any sort of role on the input side of psychological theory-building. It could be an outcome, or finding, of (current) psychology that agents are (largely) rational. (p. 104–105)

I, too, want to provide room for empirical discoveries about just how rational we are, and for scientific theorizing about the sub-personal neural mechanisms that subserve (a usefully vague term of art) the phenomena of cognition by successful agents. But I want to demonstrate that the power of folk psychology is due to its daring idealization, and that the bold extension of this folk-psychological power to other domains (computer chess programs, Martians, the R&D of natural selection, …) works precisely because of its presumption of rationality—and not because, say, the underlying processes of natural selection strongly resemble the processes that occur in believers’ brains.

Perhaps a little dramatization can bring this out.

Curious biologist: I’m baffled by the apparent extravagance of the design of this macromolecule I find in abundance in every bacterium I investigate. I’d like to explain why it has the properties it has.

DCD: My advice to you is to ask what reason Nature could have had for devising and protecting such an expensive bit of machinery.

Curious: Are you suggesting I treat natural selection as if it were a rational agent designing the inners of bacteria?

Biologists don’t need this advice; they already do this every day. They try (usually unsuccessfully) to refrain from using mentialistic terms in their sober research articles, but they nevertheless use the intentional stance as an imagination-prosthesis to generate hypotheses to test, and sure enough, they discover again and again that evolution is a quite reliably brilliant designer of organisms. They aren’t trying to prove Orgel’s Second Law; they are trying to discover the rationale of the designs of the devices in nature, the better to generate still more hypotheses to test about what how and why these things work the way they do.

Is this just a trick? It’s a good trick, an extension of the design stance that was always latent there. Remember that using the design stance involves taking on the simplifying assumption that the parts will work as advertised, that they are good springs and cogs and axles and bearings. The question that almost goes without saying in every such inquiry is “what could this bit be good for?” In the case of artifact hermeneutics, there is almost always a (good) reason, moreover, why the parts are arranged as they are, because designers are intelligent, that is to say, rational, and the same holds in the case of evolved entities because natural selection is that good. Of course we don’t need a “psychological theory” of natural selection; we already understand the underlying Darwinian algorithms that do all the work. We are just looking for a reasonable rationale for the work that they have done in this case.

The same “panglossian” idealization works well in folk psychology because people, and animals, and in some regards plants and even bacteria have been well-designed to protect themselves and further their interests. Folk psychology permits folk to make highly reliable predictions with just about zero knowledge of the underlying cognitive mechanisms. (As I have said, the intentional stance taken by itself is vacuous as a psychological theory; it presupposes only that the machinery in our heads is well-designed.) Over the millennia we folk have used informal introspection and “intuition” to develop a rather fanciful mythology about what the inner machinery is—desires dukit out, beliefs generated by perceptions piling up in the belief box, images being constructed and perused, intentions being endorsed, urges being suppressed—and some of these folk categories may prove to carve some of neuroscientific nature at the joints, but the utility of folk psychology as a portable sense-maker and hypothesis generator provides scant evidence for this hope, especially given the utility of the same strategy in evolutionary biology (and chess playing computers) where we already know that the processes behind the actions we are predicting/explaining are not much, hardly at all, like brain processes—except for the fact that they extract information and put it to adaptive use.

Dub gets close to this with his discussion of electrons. Notice that he doesn’t say that we discovered that electrons had negative charge. He said that we discovered subatomic particles with negative charge (and we call them electrons, identifiable or
distinguishable by their negative charge). Similarly, I am claiming, we discovered that intelligent animals have lots and lots of cohering information that they use to guide (appropriately) their actions. We call this information beliefs. We wouldn't call states of a person beliefs that didn't have this delightful property. Electrons are "by definition" negatively charged and beliefs are "by definition" rationally maintained.

Duh suppose we might invent the "schmiententional stance" which found a different way of systematizing the data and deriving reliable predictions. Of course this is possible, and I view cognitive science as engaged in just such an enterprise, finding new categories and states undreamt of in folk psychology. And whenever we encounter likely candidates for such theoretical innovations, we will have to confront the diplomatic/pedagogical (as opposed to metaphysical) question of whether to identify these items as none other than the beliefs and desires of folk psychology, or to claim that these items replace those obsolete categories. I join Duh in applauding the innovative proposals of Gendler, Schwitzgebel, Egan and Frankish, and don't see my view on the intentional stance as an impediment, aprioristic or otherwise, to such explorations.

I suspect that Duh has some lingering allegiance to the popular idea that we know that there are beliefs and are just trying to find the right theory of them, while I have been proposing that we consider belief-talk to be a strategy that works well, in spite of all the noisy 'counter-examples' around the edges. As Duh shows, both of my mentors, Quine and Ryle, have contributed to my confidence that this is the wise way to proceed, and I speculate that the residue of disagreement and/or misunderstanding between Duh and me is largely due to his not entirely sharing my enthusiasm for their insights.

Three relatively minor corrections for the record: I think Duh misinterprets Cherniak: it is the believer's finitary predicament, not the attributor's, that leads to the minimal rationality constraint. And Duh misses what I was trying to say about Quine's insouciance about projectionist and normative approaches: Quine noted that even the most ardent projectionist cleans up and normalizes the projections made from his own case, so on any realistic view there is not much room for the different views to yield different attributions. Duh is also mistaken when he says that "opinions were introduced in order to preserve rationality." Not so. They were introduced to distinguish language-incorporating cognitive states—bets on the truth of sentences, in my sketchy formulation—from other information-bearing cognitive states such as the beliefs (if that is what they are) of animals. (See Frankish, this volume, and my comments on it.)

5

Sam Wilkinson's essay exposes some questions I should have answered long ago, so with the help of his insights, I will try to answer them now: What is the relation between personal level intelligibility and predictability? What is the relationship between the personal level and the "free-floating rationales" of behaviors and structures of living things? Wilkinson shows me that there is a deeper connection between my early thinking and my more recent thinking than I had realized, and this will permit me to recast some points in what I hope are more persuasive terms. He has done an excellent job articulating my thinking in 1969 about the personal/subpersonal distinction, and he is right that I drifted away from (but didn't explicitly abandon) two central, Rylean parts of my original claim once I focused my attention on the three stances: I stopped stressing that personal level explanations were non-mechanistic, and favored predictability over intelligibility. I don't know if I ever clearly understood that—and why—I was shifting my emphases in these ways, but I do now, thanks to Wilkinson.

Yes, the personal level is non-mechanistic, and none the worse for that: it is this feature that makes it not just compatible with but congenial with whatever ultimately emerges as the correct mechanistic explanations of the phenomena a: the subpersonal level. (On this, see also my dialogue with the evolutionary biologist in my commentary on Duh.) As I would put it today (see Dennett 2014, "The Evolution of Reasons"), it supplements the (mechanistic) answers to the "how come?" questions with the (non-mechanistic) answers to the "what for?" questions. We will still need the personal level because of its role in anchoring intelligibility. Wilkinson also raises an interesting question: the intentional stance seems to render intelligible only a subset of the events that we address at the personal level. When someone says "Ouch!" in response to being kicked in the shin, we understand the meaning of this reaction, but it isn't that the action is shown to be the rational thing to do, given the subject's beliefs and desires. Nor is it that we understand this as a merely causal, mechanistic outcome, like the table leg buckling when somebody kicks it. How can I reconcile the intelligibility of these cases with the rationality-presupposition of the intentional stance?

First, let me address the shift from intelligibility to predictability. In retrospect I can see that the rationale (possibly free-floating) was this: "intelligibility" has the flavor of Ordinary Language Anti-science Conservatism, an attitude I found insufferable and obtuse ("You scientists go have your fun with mechanistic accounts of bits and pieces of things; we Ordinary Language Philosophers are engaged in appreciating the meanings of acts and ideas, an utterly distinct world off-limits to science.") I wanted to bridge the chasm between meaning and mechanism, not defend it, and one key element of the bridge was the requirement that intelligibility must have some practical effects, some payoff, some leverage. If rendering some stretch of human activity intelligible didn't help us see what to do next, it was just some sort of pointless decoration we couldn't help but indulge in. My motto could have been No Intelligibility without Predictability. I wanted to demonstrate that the intentional stance could do things that the physical stance couldn't do (practically), beating the physical stance at its own game of prediction. That's why the personal stance is ineliminable; not for Wittgensteinian "reasons" (explanations have to stop somewhere) or Strawsonian "reasons" (we just are the sort of creatures who harbor resentment), but for an ultimately biological reason: the personal level, by making life somewhat predictable, helps us live safer, easier, more productive lives. If I'd made that point, it would perhaps have forestalled the common objection to the intentional stance along the lines of "but we're not interested in predicting our companion's every move; we're interested in understanding it." To which my reply is: you may not appreciate that you are engaged in prediction, but you are, automatically and involuntarily, and there would be no understanding without it. I have often said that the job of the brain is to "produce future;" a claim that is becoming more and more obvious as Bayesian approaches to cognition come to be appreciated.
Of course I also wanted to stress the continuity between our personal level attributions and the attributions of computer programmers, biologists, and other scientists working with designed systems. So I would decline Wilkinson’s suggestion that “there is nothing metaphorical” about attributing a belief to a person, in contrast with attributing a belief to a chess-playing computer—if that means there is a sharp line between the two practices. I see them, and have always seen them, as on a continuum, with more “literal” attributions at one end, and highly fanciful (but still explanatory and hence justifiable) attributions at the other.

Now what about the intelligibility of saying “Ouch” when in pain? Why do we shudder, wince, tremble, smirk, sigh, flinch, scream, groan,...? Emotional reactions are not intentional actions, but they do typically get explained with ineliminable appeal to the intentional stance, because they are often the involuntary (and typical) responses to beliefs and desires. In fact, some of the most secure clues to intentional stance attributions are involuntary emotional responses. The children shrieking with delight at the puppet show because they believe Punch believes that Judy is in the box is blue-chip evidence that they are capable of attributing false beliefs, no matter what they can or can’t say, intentionally, in response to adult’s questions. (See Dennett, “Beliefs about beliefs” 1978, and the Sally-Ann industry—false-beliefs-tasks—that arose from it.)

Some emotional reactions shade seamlessly into voluntary, deliberate, intentional responses to the same circumstances. Ducking intentionally is continuous with flinching; shuddering and trembling stand in between fainting or collapsing in despair and fleeing (intentionally) from something feared/believed harmful, etc. They are intelligible responses in part because they are familiar symptoms of typical beliefs and desires, which we effortlessly learn to rely on (unless we are autistic, for example) in our largely involuntary adoption of the intentional stance. A child who has not yet witnessed blushing embarrassment (or fury) may find the first few bright red faces unintelligible, but will soon catch on. But they are also often intelligible because in spite of not being intentional actions, they have free-floating rationales that we may vaguely appreciate (and sometimes we’re wrong about them).

The backbone of personal level intelligibility is the rationality assumption, without which body language and facial expression would be almost powerless as clues. It would be an interesting (and fun) experiment to take a film—a romantic comedy, let’s say—and re-edit it with the help of some computer graphics so that actions and facial expressions that made effortless sense in context were now utterly baffling because the intentional stance could get no purchase on what the characters were engaged in trying to do.1 I might go so far as to say that intelligibility just is predict-ability from the intentional stance, but it is important to recognize that the rationality assumption of the intentional stance also has work to do quite independently of the personal level. (See also my comments on Dub.) Many human and animal behaviors have free-floating rationales (Dennett 1983, 2013, 2014) that are not personal level explanations, though they are often sloppily described as if they were. When gazelles stort (make those amazing leaps while being chased by lions) they are signaling that they are healthier than average and hence harder to catch, and the lions “believe” them, and turn their attention to other gazelles that can’t stort. Neither the gazelles nor the lions need to understand these signals, but this is the free-floating rationale for this otherwise baffling behavior, well confirmed by both evidence and theory. We use the intentional stance to render the behavior intelligible (and it is manifest not a mechanistic explanation, since no agent—no gazelle, no lion, no intelligent designer—formulates the rationale in anything like a language of thought or Language of Divine Thought). When zoologists speak, loosely, of the gazelles signaling, it looks for all the world like a personal level attribution (cf. “She’s signaling to you that it’s time to leave the party,”) and it may give someRomantics the impression that the gazelles are being attributed great williness and appreciation of linc-pysychology. (Those clever gazelles! They appreciate how to bluff the lions into leaving them alone!). And when evolutionary psychologists speak, loosely, of women being “coy” because they have a greater “investment” in reproduction than men (in time—9 months—and precious eggs), this is the free-floating rationale for an undeniable asymmetry in the animal (and human) world but not at all a personal level explanation! Evolutionary psychologists are not claiming that women (in general or in particular) have a miserably attitude towards their precious ova and are disingenuously assaying men for their genetic fitness with every ploy. Failure to appreciate this is probably a major source of the otherwise bizarrely overwrought negative reaction to evolutionary psychology by many deploiring critics.

A personal level explanation is one that a person can acknowledge, report, appreciate, evaluate (or, of course, dissemble about—but you have to be aware of it to dissemble about it). In the case of the free-floating rationale for a preference, habit, tendency, or reflex reaction, for instance, it counts for nothing when a person claims not to have considered, or to understand, or to accept, it. It may still be the (non-mechanistic, rationality-presupposing) intentional explanation of the sub-personal arrangement that provides the how come explanation. Personal level explanations of various human features are notorious for bottoning out rather suddenly in vacuity, as in “We love jokes because they are so funny!” or “I like sweet things because they taste so nice!” “Her dance arouses me because it is so sexy!”—all true enough but uninformative as answers to the perfectly legitimate what for questions that remain untouched (Hurley et al. 2011). The line between personal level explanations and free-floating rationales is porous. In the case of non-human animals, it is particularly easy to see that a personal level attribution of belief or desire may seriously exaggerate the presumed understanding of the animal while still speaking truly of the underlying rationale of the behavior. That the dog wants to catch the squirrel and believes the squirrel is still

---

1Woody Allen’s first film, What’s Up Tiger Lily? (1966) was the inspiration for this suggestion, but I’m proposing something different, and more radical. Allen bought the rights to a Japanese James-Bond-type film, threw away the dialogue and the plot and dubbed it with an entirely different story, dealing with a lost recipe for egg salad. It’s goofy and Dadaist and fun, but I’m imagining leaving the dialogue intact, but rearranging brief scenes in such a way as to make them unintelligible in spite of the fact that they were normal bodily responses to very particular circumstances. I predict the results would be striking and unsettling.
in the tree it is standing under is plausibly "personal level" for the dog, but the dog's belief that strangers are not to be trusted or desire for fatty acids in its diet can be seen as more properly free-floating rationales for tendencies of which the dog has scant—if any—understanding. In the case of human beings, I think that much human behavior we generally treat asrationally intended and well understood is at best only dimly (or retrospectively) understood. If you have ever made a move in a chess game the brilliance of which only later dawns on you, but claimed that you had the insight all along, you know how easy it is to fool others—and even yourself—about how intelligible in prospect your own behavior is to yourself. We still consider such a chess move a personal level action (unless a piece is accidentally nudged, say) because it occurs in a paradigmatic setting of rational agents in competition. Similarly, we treat "impulse purchases" and other responses to covert manipulations as personal level phenomena even when people are demonstrably confused or ignorant of the influences on their choices because these are transactions between "consenting adults" who are presumed to be rationally guiding their actions.²

When we look at the personal/subpersonal level distinction in the context of mental illness, as Wilkinson shows, we find a rather different porous boundary, created by the (ultimately mechanistic) pathology in the subpersonal systems of perception and control, leaving many attributions problematic at the personal level. Here I think Wilkinson slightly misstates the case to be made when he says that top-down and bottom-up theories, while both making the personal/subpersonal distinction, differ "about substantive, empirical facts about what is going on inside these patients." (p. 124) They may not differ on these (largely still unknown) facts, but only on what the threshold of understanding for personal level attributions should be. As Wilkinson points out, delusional patients typically fail to act on their delusional claims in ways that would tend to rescind the attribution of belief were it not for the sincere avowals of the sufferers. I also disagree with his claim that "when the chess computer malfunctions, it is so different from us that we would never ask to render its malfunctioning behavior intelligible, let alone expect to be able to do so." We wouldn't expect the computer to do so. (Do chess playing computers have a personal level? Not yet, I would say.) But we do often render their malfunctions intelligible by using the intentional stance. One of my favorite real life examples, often cited by me, was Rich Greenblatt's casual observation of a rival chess playing program that "it thinks it should get its queen out early." In a single stroke this comment rendered a great deal of that program's behavior intelligible, but I guess it should count as a free-floating rationale. (And note: even the intelligent designer of that program didn't contemplate or consider that proposition or attribution in the course of designing the program.)

²Felipe De Brigard, editing this passage, made a useful comment that deserves to be quoted: "Reasonable behavior need not be behavior that responds to reasons, or that is brought about in response to reasons. I like it. Many traditional philosophers will disagree, of course. To them I'd say, in the guise of Don Quixote, 'let the dogs bark, Sancho. It is a sign we are on track.'"

6 Martin Roth repairs "a conspicuous absence" in the current controversy over embodied cognition: the honoring of the personal/sub-personal distinction. I think he is right about its application, and in retrospect I am as surprised as he is that the combatants have ignored it, especially since both Andy Clark and Fred Adams have been interacting with me for decades on other topics. Unaccountably I never thought to propose my own distinction to them, but I expect Andy will endorse Roth's friendly amendment. It will be interesting to see if Adams and Aizawa have a response.

Roth's reconstruction of my disagreement with Fodor over Ryle is right on target. I wish I'd seen then as clearly as I do now, thanks to Roth's analysis, just what the core of our disagreement was. When I wrote C&D, Fodor was known to me only through his papers on meaning and linguistics with Jerry Katz, which were all the rage in Oxford and elsewhere, one of the opening salvos in the siege that pretty well extinguished ordinary language philosophy, but I didn't see myself as having a dog in that particular fight. (I tended to side with Fodor and Katz, as part of my growing interest in bringing science to bear on philosophical issues.) That explains why there are no references to Fodor in C&D. After I had sent my manuscript to Routledge & Kegan Paul, my colleagues at Irvine, Joe Lambert, Gordon Brittan, and Jack Vickers proposed a discussion group on Fodor's new book, Psychological Explanation (1968), which provoked us all, in different directions. I remember that we terrorized the graduate students who sat in on it with the vehemence of our attacks on each other's interpretations and arguments. We were all dear friends, but a mark of that friendship was our enthusiasm for blasting away at each other with abandon—offering hooting reduction, sarcastic "parody of reasoning" putdowns and all manner of scoffing and name-calling—philosophical debate comes to the locker room. The ideas in "Intentional Systems" (1971) were almost literally hammered out in that discussion group, so that was the first time I got lifted by Jerry, the human trampoline. As I have said before, if I can see farther than others it is because I've been jumping on Jerry.

What strikes me on reading Roth's essay is how well I'd anticipated Fodor's subsequent (mis)-reading of Ryle in my warnings. I met Fodor soon after moving to Tufts in 1971, and we soon got to thrash out the issues in person, on many occasions, in a discussion group of Boston-area philosophers, and on Jerry's sailboat, Insolvent, but that's a tale for another day.

Roth says

if Adams and Aizawa are correct that "Undervived content arises from conditions that do not require the independent or prior existence of other content, representations, or intentional agents" (2010, p. 32), it will turn out that the intentional contents and processes of people—brain-bound or extended—are derived. (p. 141)

I agree, but perhaps in a somewhat different sense of "derived" than Adams and Aizawa intend. In my debate with Searle on original and derived intentionality, I point out how, on an evolutionary account of the birth of content, all the content in the nervous systems of organisms turns out to be just as "derived" as a written
shopping list (The Intentional Stance, “Evolution, Error and Intentionality” (1987b)).

I suspect that for many people in the field, this appears to be a bridge too far, but I think it is the key insight needed to break away from what Quine called the museum myth of meaning.  (See my “With a little help from my friends,” in Ross and Brook, eds., 2000, and “Radical Translation and a Quinian Crossword Puzzle,” in Intuition Pumps, 2013). If you try to be a more staunch realist about content than this, you inevitably find yourself drifting down Scarpé’s stream to an ultimately mysterious view of original intentionality.

7

Ellen Fridland’s essay clarifies constructively what I said about intelligence and learning, clearing out a few clouds and sharpening the focus. I find nothing substantial to disagree with, but will avail myself of the opportunity to build a few more wrinkles into her account, and, first, correct a factual error of mine that she has innocently propagated.

Both Doug Hofstadter and I were struck by the Wooldridge passage she quotes, and Doug was inspired to coin the very useful term “sphexishness” in honor of these wasps, but:

We have recently learned that Wooldridge gave us—as popular science writers so often do—an oversimplified sketch of the phenomenon. The psychologist Lars Chittka wrote to me, quoting from the work of Jean-Henri Fabre (1879), which had apparently been the source for Wooldridge, who, if he had read on in Fabre, would have found that in fact only some Sphex wasps are sphexish! In fact, Fabre was eager to make the point. If at first blush you thought Sphex was clever, and at second blush you thought Sphex was stupid, try third blush, and find that some Sphex are not so sphexish after all. Chittka sent me the German translation of Fabre (I still haven’t located the French), which includes the following sentence: “Nach zweier oder drei Malen, packt ihre Füßer mit den Kieferzangen und schleift sie in die Höhle. Wer war nun der Dummkopf?” (“After two or three times,... she grabbed her [the prey’s] antennae with her pincers and slid it into the hole. Now who’s the dummy?”) (Dennett 2013, p. 398)

So now we can all go on using the term “sphexishness” with clear consciences, knowing that it is something of a misnomer, but too well established to abandon.

Fridland rightly highlights the normativity that brings flexibility and manipulability into the picture (since intelligence isn’t magic), but these features thereby also frustrate—predictably, I would say—any attempt to capture intelligence inside any fixed definitional fence. For instance, an intelligent agent has the intelligence to adjust its interests, so what is in the ‘best interests’ of an agent can change almost indefinitely: suicidal projects are not ruled out, for instance, if they further the highest goals of the agent. As I put it in Breaking the Spell (2006):

Whenever an agent—an intentional system, in my terminology—makes a decision about the best course of action, all things considered, we can ask from whose perspective this optimality is being judged. A more or less standard default assumption, at least in the Western world, and especially among economists, is to treat each human agent as a sort of isolated and individualistic locus of wellbeing. What’s in it for me? Rational self-interest. But although there has to be something in the role of the self—something that answers the cui bono? question for the decision-maker under examination—there is no necessity in this default treatment, common as it is. A self-as-ultimate-beneficiary can in principle be
dindicately distributed in space and time. I can care for others, or for a larger social structure, for instance. There is nothing that restricts me to a me as contrasted to an us. I can still take my task to be looking out for Number One while including, under Number One, not just myself, and not just my family, but also Islam, or Oxfam, or the Chicago Bulls! The possibility, opened up by cultural evolution, of installing such novel perspectives in our brains is what gives our species, and only our species, the capacity for moral—and immoral—thinking. (p. 170)

Her definition of learning nicely incorporates the tremendous changes that are wrought in us by cultural inculcation, and she notes that it is when we are what I have called Gregorian creatures that we particularly surpass in intelligence all other learning agents on the planet. What gets in the door by this route are a lot of thinking tools from which we can benefit without entirely understanding. As Andy Clark (1997) puts it, “We use intelligence to structure our environment so that we can succeed with less intelligence. Our brains make the world smart so we can be dumb in peace!” This can sometimes appear to be “cheating” when we consider (or measure) intelligence. Can you bring your pocket calculator or laptop to the exam? It depends on many factors. Senator Ted Kennedy was intelligent enough to realize that he was not intelligent enough on his own to make good decisions on many issues so he appointed the smartest advisors he could find and listened to them. Now that’s smart! But what are the limits?

As we offload more and more of our opportunity-generation-and-assessment chores to handheld electronic thinking tools (or trusted human advisors, for that matter), are we heading into transhuman sphexishness? There is no easy answer to that question, which I have been pondering for decades. It was one of the issues that inspired me, along with my colleague George Smith, to create the Curricular Software Studio at Tufts back in 1985. We had a metaphor: there were two ways of improving human muscle power: the bulldozer way and the Nautilus machine way. The first way lets you move mountains but you may still be a weakling; the second uses technology to build up your personal strength. We set out to create Nautilus machines for the mind, “imagination prostheses” that could enhance your understanding, not just give you the right answers. It is possible, and desirable, because—use it or lose it—if you delegate the hard questions to your tools, you’ll have no way of knowing if the answers you get are right.

Another point of Fridland’s I want to enlarge upon is her observation that an implication of the manipulability requirement is that

intelligence becomes a personal-level phenomenon. This is because manipulability requires global, integrated, centralized, hierarchical processes that are not available to subpersonal systems. That is, to be manipulated, a state must be targeted by higher-order states or mechanisms. The requirement that intelligent states are personal-level accords nicely with our intuitions about intelligence since, at the very least, the requirement that behaviors, processes, or representations be manipulable puts intelligence in the same realm as, for example, rationality and knowledge. (p. 151)

More pointedly, it is this feature that gives consciousness real work to do (if you hold a sane view of consciousness, ignoring zombies and the so-called Hard Problem). Meta-representation is a core strategy of greater intelligence (Fridland
aptly cites Clark and Karmiloff-Smith on this, and for a much more detailed examination of the issue see Staisilas Dehaene's recent book, *Consciousness and the Brain: Deciphering how the Brain Codes our Thoughts* (2014), which does an excellent job of ushering the "consciousness as mere epiphenomenon" view off the stage.) The personal level is constituted by what persons can share and discuss about what they are doing, and that practice is, quite obviously, the key to human intelligence. Einstein, forced to grow up alone on a desert island, without language, would be profoundly disabled, cognitively, however "gifted" at birth by his genetic endowment.

Finally, I particularly applaud her footnote 3 on her methodology, which is also mine.

The notion of intelligence that I am pursuing is a scientific notion. As such, my methodology will not be conceptual analysis. In this kind of endeavor, if various counterintuitive consequences result from my account, these will not immediately count as a reductio of the position. After all, science is often counterintuitive. Still, I hope to illustrate that what we think of as intelligence is already, to a large extent, in line with the claims that I am making here. As such, I would like the notions of learning and intelligence that I put forward to correspond to ordinary intuitions as much as possible. However, I do not insist that if ordinary intuitions conflict with the account I am offering, then the account is wrong. On my approach, it may turn out that we have empirical or methodological reasons that trump our ordinary intuitions. Intuitions ought to be considered, but they ought not to be the final arbiters. (p. 144, fn. 3)

This strikes me as just obvious good sense in the twenty-first century, but I find a surprising number of philosophers who resist it. An example of her methodology at work in this essay is her recognition of the role of higher-order or meta-representational states. I don’t think this would emerge from any pure “conceptual analysis” of the concepts of intelligence or learning. Once noted, it is quite intuitive, as she observes, but its warrant arises from empirical work in psychology, neuroscience and related fields, not from armchair reflection.

8

John Michael’s proposal of a developmental loop that sustains and refines the intentional stance as a predictive/explanatory strategy usefully builds on the earlier insights of McGee and Mameli, and I welcome this as an enlargement and improvement of my account of the intentional stance. His survey of the empirical literature (Gergely, Csibra, et al, …) is right on target, and as he notes, it is all at least consistent with, if not directly supporting, the idea that concept-acquisition or concept-mastery is itself a gradual, approximating phenomenon. And what defines the gradient up which this competence marches? Rationality, in the neutral sense of cognitive competence, whatever that comes to. Here are two apparently very different ways of putting the claim:

in Bayesian terms: children come to have ever more accurate, reliable, high-fidelity expectations.

in propositional-attitude terms: children see more and more of the implications of the propositions we are boldly attributing to them.

The intentional stance takes “propositional attitudes” from folk psychology as a way of alluding to what is learned, while the Bayesian also takes something like propositions for granted without going into the details (Expectations? Just what is an expectation, and how many of them can you distinguish?…) In either case it is important to recognize that propositions are idealizations on their own, as I argued in C&C in my example of the child who says, “Daddy is a doctor” (p. 183). (See also my discussion of the sorta operator, in *Intuition Pumps.* Just what, exactly, does the child believe at time t? For convenience we can choose one or another sentence as the best expression of what we are getting at, and plug it into our Bayesian formula, or into our propositional attitude attribution, but except, maybe, for rare artificially sharp-bounded categories, these are always idealizations.

Notice, by the way, that as soon as we permit ourselves to talk, as Michael does, of infants who “partially master” concepts, we have left Fodor and many of the “propositional task force” operatives on a distant shore. For them, Fregan grasping is all or nothing—you either have the concept HORSE (or the concept SCHMORSE) or you don’t. They are taking a bold idealization as if it were a description of brute facts and trying to theorize about it.

Michael’s idea about the developmental loop permits us to put the two ways of thinking about rationality and intentionality together: children start Bayesian, like all young animals, and are gradually intellectualized by language, so that propositional-attitude talk, always idealized, begins to do more and more justice to their psychological states. They come to have opinions, but these bets on the truth of sentences are themselves a gradually blooming and refining matter, as the child’s opinion expressed as “Daddy is a doctor” reveals.

The upshot of Michael’s developmental loop is his suggestion that as it recursively feeds on its own outcomes,

this entanglement of pattern recognition, on the one hand, and pattern etiology on the other, provides an additional justification for the belief that those patterns indeed exist, because our recognition of the patterns enables us to further embed them in their respective target systems. (p. 177)

Yes, indeed, but this passage raises the ominous specter of a community-wide delusion that is innocently supported by the new initiatives. This probably has happened. Perhaps belief in witches is like this. When everybody has the category, and knows what the defining marks of witches are, group consensus is achievable about not just the existence of witches as a general proposition but also about the identification of particular people as witches. If you’ve been raised to look out for witches, you’ll soon be pretty good—consilient with your peers—at the task. Might the intentional stance be nothing more than a persistent communal delusion then? I expect Paul Churchland and other eliminativists would be tempted to accept this

1 I agree with Michael’s suggestion that the “causal realist option put on the table here remains compatible with the general spirit of Dennett’s theory,” (p. 179) so there is no difficulty maintaining continuity of reference as both attributors and attributes revise their TOMs (though I continue to dislike the use of “theory” in this context).
little gift from me, and declare that I had finally seen the light! But, I reply, there is a striking difference: the patterns discovered and highlighted by the intentional stance are prodigiously predictive of not only those who have been enculturated to adopt the stance, but of animals, well-designed robots and chess-playing computers, and indeed of natural selection, the blind watchmaker. So I must disagree with one final passage:

A further insight generated by the developmental perspective is that it is perhaps not an assumption of ideal rationality that constitutes the core of the intentional stance as an interpretive strategy but an assumption of culture-specific imperfect rationality. (p. 180)

I am still going to resist this, since I think that culture-specific imperfections are largely elaborations of, rather than alternatives to, ideal rationality. When you learn about witches, your expectations are flavored by witch-categories, but your neuronal system is still engaged in optimizing its expectations in these terms.

9

I am grateful to Pete Mandlik for doing such a good job policing the HOT topic, a task I decided to leave to others in the twenty-first century. (My last of three or four responses to Rosenthal was Dennett 2000. See earlier 1991b, 1993a, b) I see that in the meantime there has been a lively debate, but on Mandlik's reading, it seems—to my relief—to be about to land back on my playing field after all. First person operationalism is a not so strange attractor, and it has accumulated not only allies in cognitive neuroscience, but an impressive and growing bounty of experimental results. For the latest, see Dehaene (2013). In CE, I made it clear that I thought David Rosenthal's HOT theory was right about something—and something important. His categories were, and still are, resolutely folk-psychological—"thoughts" and "beliefs"—but he has been driven, appropriately, to stretch their sense: unconscious thoughts are not just acceptable but required by Rosenthal's HOT theory, and a non-relational reading of higher-orderness is, well, in order. What he had found was an almost-folk-psychological way of expressing something important about the relation between (human) consciousness and communication: since we can report our conscious experiences, we must have thoughts (occurrent or episodic beliefs, if you prefer) to be expressed by those reports. Hence to have a conscious experience is ipso facto to have a thought to the effect that you are having it. I endorsed the strong tie to reportability in human consciousness (we'll consider non-human consciousness later), while finding Rosenthal's way of putting it still too Cartesian, depending as it does on an unanalyzed res cogitans, the thinker of those thoughts, "at the top."

Once we acknowledge unconscious higher-order thoughts, why should being the object of just any old higher-order thought secure the elevated status of consciousness? Rosenthal must be supposing there is a privileged variety of higher-order thoughts, unconscious but somehow central or dominant, that secure this status. It's like the difference, I suggested, between being famous or influential and being known by the King (Dennett 2000). In some countries being known by the King is both sufficient and necessary for being influential. In some more democratic or even anarchic regimes, there is no King whose cognizance is obligatory for influence. It would be better, I urged, to capture Rosenthal's points in explicitly sub-personal terms: fame in the brain, for a start, which doesn't depend on being known by the Emperor, because there is no Emperor. Ray Jackendoff's (1987, 2011)—and now Jesse Prinz's (2012)—vision of consciousness as an intermediate-level cognitive phenomenon with "higher" but unconscious processes doing much of the interpretive and reactive work, are elaborations of this idea with much to recommend them (see Dennett 2015)

I have always treasured Voorhees' outrage on this score:

Daniel Dennett is the Devil. [...] There is no internal witness, no central recognizer of meaning, and no self other than an abstract 'Center of Narrative Gravity' which is itself nothing but a convenient fiction.... For Dennett, it is not a case of the Emperor having no clothes. It is rather that the clothes have no Emperor. (Voorhees 2000, pp. 55–56)

Of course the clothes have no Emperor—it wouldn't be a theory of consciousness if the Emperor were still there, witnessing and reacting to all the goings on in the Cartesian Theater.

References


