Daniel Dennett: His philosophical work and legacy

BRIAN EPSTEIN

Presentation for *Symposium on AI and Consciousness Honoring the Legacy of Daniel Dennett,* October 14, 2025, Tufts University

I. Introduction

When Dean Thomas invited me to give this talk on Dan Dennett's philosophical work, I was delighted, and honored to kick off this wonderful day.

But I have to say I immediately had a second thought: I really hoped Dean Thomas had, in the scheduling of today's activities, allotted me maybe two, three hours for this talk. Because Dan's work? His impact? His influence? Well, let's just say, there's a lot of it.

I'm sorry to say, I was not allocated three hours.

To kick things off this morning, I want to trace just a few major themes of Dan's work—key things that motivated him, crucial bits of his methodology, insights and guideposts and cautions he set out for us, as we start this day celebrating his life in the way he would have appreciated most: engaging with exciting ideas and brilliant thinkers, on topics dear to his heart.

It's daunting to tackle Dan's body of work, because he was so energetic, so fascinated by everything. It was tremendous fun to talk to him, and it is equally fun to read his lively writings, packed with ideas, stories, and metaphors.

To read Dan is—as with the best philosophy—to have your own ideas sparked, page after page. It is to find him at turns insightful and infuriating, reasonable and provocative, fully persuasive and—let's be honest—sometimes dead wrong. In short, to read him is to have your imagination fired up. To read him is to spark in *yourself* interest in everything.

Philosophically, I've sometimes thought of Dan as a great debunker. As someone who refused to be at peace with mystery, with the ineffable—whether cognition, consciousness, free will, or theology. He was impatient with the idea that we should accept mysteries, as opposed to doing the hard work of figuring them out.

But though he reveled in debunking, that was not really what Dan was all about. Dan has this famous contrast between "cranes" and "skyhooks." A crane is a piece of honest machinery. When we see some complex phenomenon in the world—say beliefs, or consciousness—and do the work of explaining the mundane, technical, and developmental details that lead to that phenomenon, we've described a crane. A crane may be complex, may be intricate, but it's firmly planted on the ground.

A skyhook, on the other hand, is a miraculous lift—something that seems to bear weight but is attached to nothing; it has no purchase on the ground. When we see some complex phenomenon like beliefs or consciousness and say that they are their own stuff, or refuse to give an account in terms of the mundane and grounded details, we are just hooking things onto the air.

What a debunker does is point out that a skyhook is a skyhook. That is, to show that something is "bunk." But you can grant that something is bunk and still not really understand the phenomenon in question.

If you're going to claim that some phenomenon—like the mind—is held up by cranes, then you need to explain all the parts of the cranes. You

have to engage with the science. You have to get into the labs. You have to talk to the neuroscientists, the evolutionary biologists, the AI researchers. Only that way can you demonstrate that the cranes actually lift.

II. Dan's Projects

Dan's work is sprawling. There are so many ideas, so many examples and arguments, that it can be hard to see the structure.

I think one way of sorting out the method behind the madness is to notice that he was actually conducting several distinct projects at once, all running in parallel, all informing each other. But still somewhat distinguishable from one another.

(a) The method: heterophenomenology and fine-grained functionalism

One project you find in Dan's work is a broad methodological one. OK, grant that we're in the business of explaining the cranes that hold up complex phenomena like consciousness. What are the general methods for this work? What data can we use and how should we use it? And what bar does any given proposal need to clear, in order for it to have scientific traction?

There's a lot to Dan's views on this, but I just want to mention two big ideas that guide how, according to Dan, we ought to theorize about things like consciousness. One is his notion of "heterophenomenology," and the other is how he thinks about functionalism.

Heterophenomenology. While Dan was amazing at naming things, I have to confess that that word is a mouthful. But the idea's interesting, so it's worth unpacking. Commonly a distinction is made between taking a first-person perspective and taking a third-person perspective. Heterophenomenology is *not* the methodology of saying look, we have to *only* take a third-person perspective, because science has no access to the

subjective. But it does say that we can't take the subjective perspective as gospel, as if people have infallible access to truths about their own cognition. After all, as Dan delighted in pointing out, we constantly make mistakes about things like what we ourselves are perceiving. But on the other hand, and crucially, subjective experience is an enormous source of data. It would be insane for science to be forbidden to use it. It just can't be regarded as authoritative.

So Dan's procedure is this: Collect the subject's reports of their experiences. And treat those reports as *data*. You acknowledge the subjective perspective—"it seems to you that such-and-such"—while seeking an objective account of *why* their brain produces that report. You treat what they say as data to be explained, rather than as a theory-stopper.

So that's a crucial theme in the epistemology of cognitive science, in Dan's view—what kinds of data should be used and not used. Then there's a theme about the objects of cognitive science: which ones should get admitted and which ones don't deserve to. Here's where his version of functionalism comes in.

Traditionally in cognitive science, functionalism is understood as a view about mental phenomena being sort of coarse-grained: if a human and a computer and a dog and a Martian have machinery that does the same high-level operations, they all can be considered to be implementing the same cognitive processes. Dan's version of functionalism is more fine-grained. Take some proposed thing in some actual cognitive system. If that thing makes a difference anywhere in the entire web of activity of that system—if it alters discrimination, report, memory, or control—then it matters. If it never shows up *anywhere*—no difference to prediction, no difference to behavior, no difference to how other parts are organized—then it doesn't earn a place in the science. This difference between coarse-grained functionalism and Dan's fine-grained functionalism is really what makes him so interested in the details of the science. He thinks that aspects

of consciousness etc. arise from processes that only show up in very detailed features of actual cognitive systems, and that if we abstract from those details, we completely miss out on them. On the other hand, the only thing that makes something matter for science is the functional impacts it has. And so if we do have a different system that implements those same functions—say one in silicon rather than in neurons—then it also will manifest the mental phenomena.

Dan saw heterophenomenology and functionalism as two sides of the same coin.

(b) The debunking project

This broad methodological project is really the engine behind his debunking project. It shows how we can then dislodge what he takes to be unhelpful dogmas about the mind.

One major target was a notion he thought we'd be better off *not* using; namely the philosophical concept of "qualia." The idea, that is, that experiences have intrinsic, ineffable, private properties. The redness of red, for example; what it feels like to see red. Dan's point is not "nothing hurts" or "nothing looks red." It's that when people speak of qualia, they're not even clear about what they're talking about, and tend to use it in many different ways. And more importantly, qualia violate both of his methodological beacons: they are supposed to be exactly what heterophenomenology cannot give evidence for, and on the other side of the coin, they are also functionally inert, since they are supposed to be the aspects of mental phenomena that go beyond what makes a functional difference.

With regard to *consciousness,* Dan's debunking project is a little bit different: he's not saying that consciousness is unreal or incoherent or that it is used in so many ways that it should be tossed away. He doesn't like qualia because he takes them to be a notion designed by philosophers to

point at something mystical. But for consciousness, the aim should be just to strip it of mystery.

And that bring us to Dan's third project, which is to open the curtains and reveal the guidewires and tricks behind the magic.

(c) The scientific project

The third project we can see in Dan's work is to explain why mental life presents itself the way it does. Why it *feels* like there's a movie playing inside. Why we talk about "what it's like" as if it were a special substance.

Dan doesn't say that there's no such thing as consciousness, or that it's not real. But he does approve of Keith Frankish and others calling his view "illusionism." There's a sense in which consciousness is, according to Dan, an illusion.

It's the same sense in which magic tricks are illusions: you think you saw the lady get cut in half. You *do* see something—you're not hallucinating—but what you think you see, or what you infer you've seen, isn't what actually happened.

Similarly, according to Dan, for consciousness. What the science shows, he argues, is that there is a family of *layered mechanisms* that, together, produce the familiar profile of our experience.

It starts with the fact that the brain is not centralized. It's running many processes in parallel, constantly drafting and revising candidate contents all at once. There isn't a single "arrival time" when things become conscious. Instead, these different drafts compete for influence. Some stabilize in working memory, shape our actions, or make it into our verbal reports. Others fade.

This competition explains why our experience sometimes plays tricks on us. The brain integrates information over short windows of time before

anything becomes “official” for memory and action. This means later signals can actually shape earlier reports. That’s how you can experience things that seem impossible, like a color changing mid-flight in certain perceptual experiments.

Furthermore, our attention isn’t a neutral spotlight illuminating a pre-existing scene. It’s just a matter of what gets on top in the competition for scarce bandwidth.

For Dan, the supposed “hard problem of consciousness” is a halo effect cast by many easy-to-state, hard-to-engineer problems working in concert. The way you address it is to do the reverse engineering, showing how the parts produce the profile, detail by detail.

(d) When mental talk is earned

And then there’s a fourth project, which is really a project of traditional philosophy of mind, and which Dan’s work sometimes engages with but sometimes stands against. Namely, identifying criteria for mental phenomena we’re interested in. *When* are we entitled to talk about *belief*, *understanding*, and *consciousness* without scare quotes? Dan is not a fan of philosophical analysis of these sorts of notions. He’s a functionalist, but doesn’t offer us coarse-grained functional (or teleofunctional) analyses. He does give us a general answer, which lets us be “mild realists” about these things: we attribute these when those attributions track patterns in organization and control that you can’t replace with a cheaper description.

Dan doesn’t, though, give us specific answers or analysis of belief, understanding, or consciousness. There’s a reason for that: there is no distinctive kernel or essence to these. Nothing, for instance, that privileges one particular pretty-fine-grained package of layered tricks as the consciousness-package. As Dan says, for most systems—that is, particular organisms as they evolve—we find ourselves needing to use the “sorta” operator a lot. This organism is sorta-conscious, sorta not. Implements

these tricks, doesn't implement those. He's not in the business of insisting on philosophical analyses.

Dan's refusal to analyze means that he doesn't give us specific criteria for things like AI consciousness. What *exactly* does it take for some computational system to be conscious? His work suggests features that consciousness will tend to involve—some form of integrated control, wide access for winning contents, coupling of perception and action, etc.—but these are features that a good design discovers, not any kind of checklist.

What, then, does his work imply for things like AI and AI consciousness?

III. Dennett on Strong AI and Today's LLMs

Dan was never shy about *Strong AI*. If minds are *organized achievements*—if what matters is the web of mechanisms by which contents win influence, behavior gets controlled, and errors get corrected—then, in principle, you could build a mind in a different substrate. There is no magical ingredient you must smuggle in.

But "in principle" is the easy part. That still leaves open the question of *what kind of organization would actually do it*, and *how we would know* when we see it.

Inasmuch as there are criteria for AI, he favors diagnostic ones. Dan was all along a big advocate of the Turing Test—so long as it is done properly. Not a five-minute imitation game, but the open-ended performance of a life: sustained, wide-ranging conversation that weathers months of interaction, memory for past encounters, the ability to bring things *seen, done, and learned* to bear, and to revise in the face of correction. He thought a true pass at that bar would *force* the right cranes—perception, action, memory, integrated control—because a creature that can do all that will have needed them along the way.

Still, Dan isn't proposing that those features, or the bag of tricks that underlie our consciousness, is either necessary or sufficient. He just thinks that the Turing test is aligned with heterophenomenology—both discipline us to rely on what can be manifested in behavior, report, and control over time.

On today's LLMs

With that in view, how should we understand his stance toward today's large language models? I think it's best seen as a complex and balanced approach. Balance between openness to the future and caution about the present.

On the one hand, he was steadfastly open to the possibility of conscious AI. It is not crazy to think that near-future systems—systems augmented with tools, memory, sensors, actuators, and better forms of control—might have all the functional characteristics needed. His picture never needed a ghost.

But on the other hand, he was deeply cautious about our current situation. Present systems are extraordinarily good at *eliciting* our stance-taking. They talk like us, so our social machinery fills in beliefs and motives they simply haven't earned.

That is why he wrote a warning in the Atlantic a couple of years ago about *"counterfeit people."* This isn't an insult to the engineering; it's a warning about the fragility of our social trust. We have to *label* what we're dealing with, require *provenance*, and restrict *impersonation* where personhood conveys authority. We must do this while the research community does the slower work of actually building the cranes.

A system that has some reasoning-like functions, and that passes a quick imitation game, has nothing like the layers-upon-layers that give rise to phenomena like consciousness. People are generous in taking the

intentional stance, and even to ascribing consciousness: we even talk to our labubus, or cabbage patch kids, or whatever. We need to be on the lookout for that, without then concluding that AI consciousness is ruled out.

IV. Legacy and Conclusion

Part of Dan's legacy is his criticism: the skyhooks he made it unfashionable to appeal to. But the greater part, by far, is positive.

Dan shifted the intellectual landscape in ways that are now almost invisible. Many of the assumptions common across philosophy, psychology, neuroscience, and AI bear his fingerprints, even when people don't realize it.

It is now routine, in serious conversations about the mind, to treat first-person reports as data to be explained rather than as vetoes. It is routine to ask whether a level of description pays its way in prediction and control. It is routine to talk about competition for influence instead of inner theaters. And it is routine for researchers to use the intentional stance as a disciplined modeling strategy, not a metaphysical confession.

Today, many of these moves feel ordinary, things we take for granted. That is a mark of a deep impact.

Second, he left behind *research programs.* In philosophy, the "illusionists" extend his pressure on consciousness. In neuroscience, models like the Global Workspace Theory and Predictive Processing resonate deeply with his anti-theater insistence. In human–robot interaction, researchers are actively developing ways of measuring when and why we attribute minds to artifacts.

Third, and most important, is the legacy of people. Dan built networks. He connected researchers across disciplines who might otherwise never have spoken. He mentored countless students, demanding clarity and

intellectual courage. He was a tireless correspondent and an enthusiastic champion of good work, wherever he found it. The community gathered here today is a testament to that.

Before kicking off the day, and getting to the people you *actually* came to hear, I want to end with a couple of personal notes. First, I want to say how much we as the Tufts community owe to Dan. He was deeply engaged in building Tufts into a world-class institution at exactly the time Tufts needed it, making it the premier place it is—a destination for cognitive science, a destination for philosophy, and a destination for intellectual achievement. He also had an enormous and lasting impact on students, training them in the ways of research and intellectual honesty.

For my part, he was a crucial supporter. He encouraged me to take over his big Language and Mind course early in my time here as an assistant professor, and then I had the pleasure of co-teaching Philosophical Foundations of Cognitive Science with him before taking that course over as well. Dan, this giant in cognitive science, insisted that I make our joint syllabus mine, rather than just re-teaching his seminar. And when we taught together, we disagreed and debated about nearly every philosophical point. I agreed with Dan about a lot, but disagreed about more. And that was just the way he liked it.

The day ahead of us is exactly the sort of day Dan would have relished. We have a sparkling lineup of speakers. They represent exactly the kind of rigorous, interdisciplinary engagement he championed, spanning neuroscience, philosophy, robotics, bioengineering, and cognitive science. We're grateful to have them all here, and grateful to have you all here, joining together to celebrate Dan.