Daniel C. Dennett

The Mystery of David Chalmers

1. Sounding the Alarm

'The Singularity' is a remarkable text, in ways that many readers may not appreciate. It is written in an admirably forthright and clear style, and is beautifully organized, gradually introducing its readers to the issues, sorting them carefully, dealing with them all fairly and with impressive scholarship, and presenting the whole as an exercise of sweet reasonableness, which in fact it is. But it is also a mystery story of sorts, a cunningly devised intellectual trap, a baffling puzzle that yields its solution — if that is what it is (and that is part of the mystery) — only at the very end. It is like a 'well made play' in which every word by every character counts, retrospectively, for something. Agatha Christie never concocted a tighter funnel of implications and suggestions. Bravo, Dave.

So what is going on in this essay? It purports to be about the prospects of the Singularity, and since I can count on readers of my essay to have read Chalmers, I needn't waste so much as a sentence on what that is or might be. See Chalmers (2010). I confess that I was initially repelled by the prospect of writing a commentary on this essay since I have heretofore viewed the Singularity as a dismal topic, involving reflections on a technological fantasy so far removed from actuality as to be an indulgence best resisted. Life is short, and there are many *serious* problems to think about. I said as much in an email to the editor only to get an email in response from Chalmers, urging me to reconsider:

Correspondence:
Email: daniel.dennett@tufts.edu

Journal of Consciousness Studies, **19**, No. 1–2, 2012, pp. 86–95

> hi dan,
>
> take a look at the paper. somehow i suspect that you'll have plenty to say. some of the core issues here concern the structure of intelligence/ design space, topics that you've thought pretty hard about.
>
> cheers,
>
> dave[1]

And since I respect Chalmers' judgment, I relented, and read the essay. My reactions to the first thirty-odd pages did not change my mind about the topic, aside from provoking the following judgment, perhaps worth passing along: thinking about the Singularity is a singularly *imprudent* pastime, in spite of its air of cautious foresight, since it deflects our attention away from a much, much more serious threat, which is already upon us, and shows no sign of being an idle fantasy: we are becoming, or have become, enslaved by something much less wonderful than the Singularity: the internet. It is not yet AI, let alone AI+ or AI++, but given our abject dependence on it, it might as well be. How many people, governments, companies, organizations, institutions, ... have a plan in place for how to conduct their most important activities should the internet crash? How would governments coordinate their multifarious activities? How would oil companies get fuel to their local distributors? How would political parties stay in touch with their members? How would banks conduct their transactions? How would hospitals update their records? How would news media acquire and transmit their news? How would the local movie house let its customers know what is playing that evening? The unsettling fact is that the internet, for all its decentralization and robust engineering (for which accolades are entirely justified), is fragile. It has become the planet's nervous system, and without it, we are all toast.

So endeth the sermon. And now to the rest of his essay, which does indeed touch on topics about which I have thought long and hard. All along, he scrupulously draws attention to the places where his argument is porous. Thus, when discussing the basic, enabling premise of the essay, he notes that 'there is logical space to resist the argument' (p. 21) in the form of doubts about whether an intelligence measure can be secured that permits it to be scaled ordinally (y is more intelligent than x, and z is more intelligent than y, so [?] z is more intelligent than x), and perhaps would in any case be better represented by a logarithmic scaling. An admirable attention to minutiae! But — I think he

[1] Personal correspondence (quoted with permission).

does slight this (at least logical) possibility — perhaps human intelligence is so remote in degree from all previous forms of intelligence in the natural world (dolphins, chimps, starfish, bacteria), that any scale we could contrive (think of IQ!) would be so anthropocentric as to be comically distorting of whatever reality it was called upon to measure. In any event, the inexorable march of all these stacked inferences leads us to worry about whether we human beings would be left out in the cold after the Singularity, and hence leads us to consider the prospects for 'uploading' ourselves into the AI+ world. This provides an interesting hypothetical motivation (for the first time, really) for taking some favorite philosophical puzzles seriously: 'the key question is: will I survive uploading?' (p. 42). While many philosophers and philosophy students have zestfully tackled the problems of personal identity and consciousness over the years, spurred on in some measure by Chalmers' own musings on the topics, the prospect of the Singularity probably provides a boost of self-interest, mounting even to alarm, in readers who would otherwise ignore these puzzles: if uploading is my only hope of surviving the Singularity, I had better take a good hard look at the idea, and not too breezily dismiss it as an amusing but idle philosophical fantasy or riddle! If, for instance, you never before found the debate between *further-fact* theorists and *closest continuer* exponents gripping your attention, maybe now you can be made to care deeply. Or maybe not, but it's a nice try, and it does frame the issues in a rather crisper setting than most earlier treatments.

2. Uploading and Consciousness

Here is where the mystery begins to emerge. 'One central problem,' Chalmers tells us, 'is that consciousness seems to be a *further fact* about conscious systems' (p. 43) over and above all the facts about their structure, internal processes and hence behavioral competences and weaknesses. He is right, so long as we put the emphasis on 'seems'. There does *seem* to be a further fact to be determined, one way or another, about whether or not anybody is actually conscious or a perfect (philosopher's) zombie. This is what I have called the Zombic Hunch (Dennett, 2005). I can feel it just as vividly as anybody; I just don't credit it, any more than I credit the sometimes well-nigh irresistible hunch that the sun goes around the earth; it surely does *seem* to go around the earth. This makes me, in Chalmers' taxonomy, a 'type-A materialist' as contrasted with 'type-B materialists' such as Ned Block and 'property dualists' such as Chalmers himself. Chalmers thinks 'It is worth noting that the majority of

materialists (at least in philosophy) are type-B materialists and hold that there are epistemologically further facts' (fn 27, p. 43). He's probably right about this, too, more's the pity, but I think it tells us more about the discipline of philosophy than about the likely truth. I suspect that he doesn't give us any percentages of allegiance for non-philosophers because he just can't get non-philosophers to pay attention long enough to be sure they understand all the philosophical fine points that distinguish the options arrayed for their selection.

We are now ready for the posing of the mystery. *Why is Chalmers not a type-A materialist?* He gives very good arguments for type-A materialism, and finds no flaws in them. He also sides with type-A materialism against type-B materialism. And on the subsidiary issue of the distinction between *biological* and *functional* theories of consciousness — and the disagreement here is 'crucial' since an implication of biological theories is that uploads cannot be conscious (alas!) — he sides with me (against Block and Searle, for instance): 'My own view is that functionalist theories are closer to the truth here.' I am not entirely happy with his way of putting it:

> It is true that we have no idea how a nonbiological system, such as a silicon computational system, could be conscious. But the fact is that we also have no idea how a biological system, such as a neural system, could be conscious. (p. 44)

I think we *do* have lots of ideas about how such systems, biological or silicon, could be conscious, but I agree that it is just as hard to see this when staring at neurons as when staring at circuit boards. He goes on in any case to support this view 'with further reasoning', as he says, in both the main text and in footnotes. Considering a variation on Searle's (1992) thought experiment about the possible outcomes of 'gradual uploading', he sides with me again (see Dennett, 1993) noting that a gradual fading of consciousness in such a case 'seems implausible' (p. 46).

> We can imagine that at a certain point partial uploads become common, and that many people have had their brains partly replaced by silicon computational circuits. On the sudden disappearance [of consciousness] view, ...[p]eople in these states may have consciousness constantly flickering in and out, or at least might under total zombification with a tiny change. On the fading view, these people will be wandering around with a highly degraded consciousness, although they will be functioning as always and swearing that nothing has changed. In practice, both hypotheses will be difficult to take seriously.
>
> So I think that by far the most plausible hypothesis is that full consciousness will stay present throughout. (p. 47)

Indeed. I consider this to be an impressive consideration in favor of type-A materialism. It is observations of just this sort, in fact, that have always persuaded me that any alternative to type-A materialism is forlorn. Chalmers manifestly understands the arguments; he has put them as well and as carefully as anybody ever has. (See also his chapter 7 of *The Conscious Mind*, which, as he notes in fn 30, develops the arguments in even more careful detail.) So what is holding him back? Why does he cling to the Zombic Hunch and 'property dualism'? He tells us, point blank: 'Of course it remains at least a logical possibility that this process will gradually or suddenly turn everyone into zombies.' (p. 47) A logical possibility. How seriously should we take this logical possibility? 'But once we are confronted with partial uploads, that hypothesis will seem akin to the hypothesis that people of different ethnicities or genders are zombies' (*cf.* Dennett, 1991, pp. 405–6). Chalmers is reminding us of just how negligible the philosophers' notion of logical possibility can be: it is *logically* possible that all women, or lefthanders, or people born under the sign of Capricorn are zombies; it is similarly logically possible that there isn't a drop of water in the Pacific Ocean (an omnipotent evil demon has replaced it all with hallucination-stuff that seems just like water). One wouldn't want to deflect one's theory of consciousness (or oceans) by honoring such a trivial scruple about a mere logical possibility, would one?

Does Chalmers offer anything in addition to this logical possibility in support of his continued allegiance to the Zombic Hunch? He turns to the topic of personal identity and whether uploading would — under any circumstances — amount to survival, and presents both optimistic and pessimistic arguments (since he declares himself unsure). These arguments develop in somewhat greater detail the considerations I explored in 'Where am I?' (1978; reprinted in Hofstadter and Dennett, 1981) and lead him, once again, to the conclusion I leapt to then (Chalmers never *leaps* to conclusions; he *oozes* to conclusions, checking off all the caveats and pitfalls and possible sources of error along the way with exemplary caution):

> At the very least, as in the case of consciousness, it seems that if gradual uploading happens, most people will become convinced that it is a form of survival. . . . I am reasonably confident that gradual uploading is a form of survival. So if at some point in the future I am faced with the choice between uploading and continuing in an increasingly slow biological embodiment, then as long as I have the option of gradual uploading, I will be happy to do so. (pp. 53–5)

What about 'reconstructive' uploading? He reinvents Hofstadter's thought experiment in 'A Conversation with Einstein's Brain' (Hofstadter and Dennett, 1981): 'If we reconstruct a functional isomorph of Einstein from records, will it be Einstein?' (p. 57) And once again, his oh-so tentative conclusion is that it doesn't differ substantially from the gradual uploading he has already endorsed as a valuable variety of survival. All this is in nice agreement with type-A materialism of the functionalist sort.

We're getting closer and closer to type-A materialism, and Chalmers' tantalizing intellectual strip-tease continues, confronting the *further-fact* view of survival in uploading that seems to be the last bulwark against type-A materialism, and although '[t]here is at least an intuition that complete knowledge of the physical and mental facts in a case of destructive uploading leaves an open question ... and there is an intuition that there are facts about which hypothesis is correct that we very much want to know' (p. 58), '... it is far from obvious that there really are facts about survival of the sort that the further-fact view claims are unsettled' (p. 60). So we're down to two intuitions (for whatever they are worth) and a logical possibility (for whatever that is worth) and *it is far from obvious* that there is so much as an issue here. 'I *do not know* whether such questions have objective answers .... But *it is not out of the question* that this value scheme should be revised.... I *am not sure* whether a further-fact view or a deflationary view is correct' (p. 62). And philosophers wonder why non-philosophers get impatient with them!

What, then, do I make of all this? Some years ago in conversation with Chalmers, after reaching an impasse of just the sort illustrated above, I thought I heard him say that there was no point in my presenting him with any more *reasons* in favour of my position since no argument could shake his brute intuition (the Zombic Hunch), and that was all there was to it. I decided to take him at his word, and refrain from further attempts at philosophical argument since he had assured me they would be fruitless. So I recommended that he seek therapy or perhaps a change in diet. Who knew what might dislodge an impenetrable intuition! He did not take kindly to my suggestion, and I resolved not to press the point further. But now I find myself puzzling once again. My spade is turned, as before, and this time he has provided me with yet more evidence that arguments really will not avail, since he has presented excellent versions of them himself, and failed to convince himself. I do not mind conceding that I could not have done as good a job, let alone a better job, of marshalling the grounds for type-A

materialism. I'd be bringing coals to Newcastle if I tried. So why does he cling like a limpet to his property dualism?

I think there are (at least) seven possible answers to this puzzle, and I find myself unconvinced by all seven. (Chalmers' caution is infectious.) Still, I think that *a case can be made* for each of them, and while there is a *logical possibility* that none of them deserves to be called *the* explanation, *it is not out the question* that one or more of them deserves to be taken seriously, as seriously as any not merely logical possibility deserves to be taken. As luck would have it, all seven answers can be labelled — with a little procrustean tugging — with the same letter, much like the famous four Fs of animal options: fight, flee, feed and engage in sexual intercourse.

3. The Famous Seven Fs of the Mystery of Chalmers' Resistance to Type-A Materialism

1. Faith

Could it be that Chalmers, like Descartes, is attracted to dualism by a residual fondness for the Christian doctrine of an immortal, immaterial soul? I find this highly unlikely, but in the interests of something approaching exhaustion of possibilities, I must list it. The late Sir John Eccles, Nobel laureate neuroscientist and devout Catholic, certainly gave us an instance of the category, and it does give one pause that, coming from an entirely different quarter, Jerry Fodor (2008) has recently decided that the epithet that best describes his own view of wisdom about the mind is 'Cartesian', a label he is now proud to sport. But the central attraction of *property* dualism, I gather, is that it provides a stumbling block for the scientific study of the mind (the Hard Problem) *without* postulating an embarrassing substance, a miracle pearl of sorts, that might leave our bodies when we die.

2. Fame

Many years ago, over a few drinks, I offered up Uncle Dan's advice for how to become a famous philosopher: invent a new (short, punchy, but unsound) argument for dualism; publish a brief version of this in a philosophy journal and then watch it get snapped up by professors around the world looking for a head-snapping attention-grabber for their students, an argument that even the most callow undergraduates could be motivated to care about — and refute. It would migrate from a few syllabuses to many, and then be anthologized, rebutted, defended, analysed, translated, caricatured, and turned into a 'classic'. David Chalmers was not present on that occasion but somebody

who was — and who shall remain nameless — actually tried to take my advice, offering publishers 'a new argument for dualism' and getting a contract to write the book. Unfortunately this would-be famous philosopher had neglected to compose the novel argument in advance, and in spite of much searching and agonizing couldn't deliver. A different book was written and grudgingly accepted for publication. Fame eluded its author. This is not David Chalmers' story, but it is possible that the fame that has accrued to Chalmers and the so-called Hard Problem has something to do with his continued allegiance to the position. If so, he should reconsider: Frank Jackson has recanted his famous argument for dualism about Mary the Color Scientist without any loss — indeed with an increment — of fame and influence. (And no, Frank Jackson was neither the inspiration for, nor the one inspired by, my advice.)

3. Freud

Douglas Hofstadter is David Chalmers' *Doktorvater* (and I was an informal member of his dissertation committee). Neither Hofstadter nor I have expressed any support for Chalmers' brand of property dualism, and indeed have published quite a lot over the years expressly arguing for (what Chalmers calls) type-A materialism. Moreover, Hofstadter has been unusually frank in expressing his own conviction that the Singularity is an idea not worth serious consideration, calling it on one occasion a 'nutty technology-glorifying scenario' (http://tal.forum2.org/hofstadter_interview) and saying on another occasion that the discussion of it by Kurzweil and others was 'as if you took a lot of very good food and some dog excrement and blended it all up so that you can't possibly figure out what's good or bad' (http://www.americanscientist.org/bookshelf/pub/douglas-r-hofstadter). Could it be that Chalmers has gone to great lengths to distance himself from his early mentors, even going so far on this occasion as to ignore the versions of arguments, by Hofstadter (in *Gödel, Escher, Bach,* in *The Mind's I,* in *I Am a Strange Loop*) and me (in *The Mind's I,* in *Consciousness Explained* , in *Sweet Dreams*) that anticipate his own discussions? A farfetched hunch, but logically possible.

4. Fiction

On this hypothesis, it is a mistake to read this essay as what it appears to be on its surface: a serious philosophical essay. It is rather, like Borges's *faux*-erudite reviews of non-existent books (in *Labyrinths*, 1962, for instance), a parody of academic scholarship, or philosophy,

or both. It is designed to take in academic philosophers of the analytic school in much the way Alan Sokal's hoax took in the postmodernist editors and readers of *Social Text*. Such a subtle project is much more difficult than Sokal's, I think, and it is a credit to Chalmers' talent that he has managed to convince so many people that this is earnest philosophy, not a practical joke. (Chalmers is not averse to such capers; he once spread the rumor on his website that I had recanted and embraced dualism.) But I am not persuaded that this is the case, since there is an alternative that has more plausibility (to me).

5. Filosofia

(This is the procrustean tug I warned about — I have to switch to Italian to preserve my alliteration scheme.) This is not witting parody; this is unwitting parody. This is a philosopher performing the following speech act: *I am a philosopher and this is what philosophers do.* We no longer debate how many angels can dance on a pinhead, but we do pursue exhaustively nuanced analyses of our intuitions and the (logically) possible implications of them.

6. Fun

There is some textual evidence in the essay for the hypothesis that concern about the impending Singularity is really just a pretext, intended to 'motivate' the clever exploration of a set of delicious puzzles where you get to display your intellectual agility. The first forty pages seem designed to protect the prophecy from all varieties of killjoy skepticism that would spoil the game, so that we are licensed to consider the prospects of uploading as something more important than idle fantasy. This is a gambit not unknown among philosophers. Much of the contemporary literature on free will, for instance, is saturated with discussion of the tactics of argumentation, and meta-comments on the strengths and weaknesses of various *moves*, to the point where the reader may begin to suspect that the combatants would hate to see a resolution to the controversy since it would bring their sport to an end. As the late great linguist Jim McCawley once quipped, in answer to the question of how you tell the philosophers from the linguists: 'The philosopher is the one who will contribute a paper on the hangman paradox to a symposium on capital punishment.'

7. Fear

Finally, there is the possibility that Chalmers is motivated, as he hints at the end, by fear of death. But then wouldn't he cling to type-A

materialism, which is the view that holds out the best promise of continued survival indefinitely (see *Consciousness Explained*, p. 430)? Perhaps his motivation is more subtle. He dare not hope too openly, but must plump relentlessly for the worst, most dismal option, thereby damping the blow of bitter disappointment with reasoned anticipation. As he says in closing, 'My own strategy is to write about the singularity and about uploading. Perhaps this will encourage our successors to reconstruct me, if only to prove me wrong' (p. 63).

References

Borges, J.L. (1962) *Labyrinths: Selected Stories and Other Writings*, New York: New Directions. [La Biblioteca de Babel, 1941, in El jardin de los senderos que se bifurcan, published with another in *Ficciones*, 1956, Emece Editores, S. A., Buenos Aires.]

Chalmers, D.J. (2010) The singularity: A philosophical analysis, *Journal of Consciousness Studies*, **17** (9–10), pp. 7–65.

Dennett, D.C. (1978) *Brainstorms: Philosophical Essays on Mind and Psychology*, Montgomery, VT: Bradford Books.

Dennett, D.C. (1991) *Consciousness Explained*, Boston, MA: Little, Brown, and London: Allen Lane.

Dennett, D.C. (1993) Review of John Searle, *The Rediscovery of the Mind*, *Journal of Philosophy*, **60** (4), pp. 193–205.

Dennett, D.C. (2005), *Sweet Dreams: Philosophical Obstacles to a Science of Consciousness*, Cambridge, MA: MIT Press.

Fodor, J. (2008) *LOT 2: The Language of Thought Revisited*, Oxford: Oxford University Press.

Hofstadter, D.R. (1979) *Gödel, Escher, Bach: An Eternal Golden Braid*, New York: Basic Books.

Hofstadter, D.R. (2007) *I am a Strange Loop*, New York: Basic Books.

Hofstadter, D.R. & Dennett, D.C. (1981) *The Mind's I: Fantasies and Reflections on Self and Soul*, New York: Basic Books.

Searle, J. (1992) *The Rediscovery of the Mind*, Cambridge, MA: MIT Press.