Daniel C. Dennett[1]

The User-Illusion of Consciousness

About thirty years ago I attended a summer 'retreat' of the Harvard Medical School MD/PhD programme in neuroscience and I vividly remember one lab director's opening remarks in the first session. 'In our lab, if you work on one neuron, that's neuroscience; if you work on two neurons, that's psychology' — a term of abuse in his corner of the world. This bottom-up approach to neuroscience is still advocated in many quarters, and there are not a few neuroscientists who dismiss the 'cognitive' neurosciences as mere headline-grabbing speculation, but while the brute empirical researchers have provided a wealth of hard-won data in recent years, they have had next to nothing of importance to say about the mind or consciousness. This is not surprising when you confront the fact that the brain has literally trillions of moving parts. Billions of individual cells, each a complicated and rather autonomous micro-agent with an agenda, and no two exactly alike, are *somehow* coordinated to produce impressively accurate intelligence on the world outside the skulls they labour in, generating appropriate behaviour under most circumstances. How can anybody think responsibly and creatively about such a complicated organ? Clearly, one needs a model at a higher level which can systematize and rationalize the astronomical number of transactions and interactions between the parts.

Fortunately, we have one example of a physical phenomenon whose trillions of micro-actions *can* be understood: digital computers. Digital computers, unlike brains, have blissfully regular structures, composed

Correspondence:
Email: daniel.dennett@tufts.edu

[1] Center for Cognitive Studies, Tufts University, Medford, MA 02155, USA.

Journal of Consciousness Studies, **28**, No. 11–12, 2021, pp. 167–77
DOI: 10.53765/20512201.28.11.167

of billions of elements — registers and flipflops in particular — that are almost atom-for-atom duplicates of each other, and that can be relied on to execute *exactly the same process* millions of times in a row if called upon to do so. (Their digital design allows the systems to absorb atom-level variations instead of propagating them.) This gives digital computers and the models programmed on them a remarkable epistemological transparency: when they work, we can know to a moral certainty that no unimagined, unsuspected, undetectable forces or fields or processes contribute to the work being done. We may not be able to explain *just how* AlphaGo developed its ability to beat the world's best Go players, but we know it did it without the help of any 'irreducible' properties that require anything like a revolution in physics to explain. This is the chief — if seldom heralded — appeal of *functionalism.* It is the *information* embodied in the software that accounts for the 'magical' competence of AlphaGo, and we are just beginning to plumb the depths of what such information processing can do. Functionalism is, after all, a *species* of behaviourism, the conservative default position of all the physical sciences: once you've explained all the external and internal *behaviour* of something (volcanoes, the weather, the rotation of galaxies, the ribosome's ability to read DNA, the folding of proteins), you get to declare victory, since there is nothing left to explain.

Mark Solms, in his boldly imagined, empirically well-grounded book, *The Hidden Spring*, understands this attraction of functionalism, and sees that David Chalmers' 'hard problem' targets it directly: according to Chalmers, once all the functions of consciousness are explained by functionalism, we will still have the 'what-it-is-like-ness' of consciousness unexplained; we can readily imagine (can't we?) that all those functions could be performed by a philosopher's zombie that it isn't like anything to be. A philosopher's zombie acts just like a normal person, laughing and loving and (apparently) day-dreaming and solving problems, but 'there's nobody home'. This idea, which I have called (2005) the Zombic Hunch, with deliberate disrespect, has been strangely attractive to several decades of theorists, in spite of the fact that it is a bald appeal to an untestable variety of conceivability that typically is taken to be too obvious to need detailed defence. Solms thinks he has a novel response to it and in some ways he does. By taking up the challenge to provide at least some of the details of how *the functions of consciousness* are embodied in the brain, and by addressing the confusions underlying the Zombic Hunch, he exposes its negligible claim on our credence.

So I come to this book as a partisan, and interestingly, my partisanship initially led me to misread it badly, an error Solms graciously corrected when I sent him a draft. Reflecting on what misled me, I have come to see that I jumped on certain phrases and terms Solms uses that carried connotations that he didn't intend. One in particular was his choice of 'feelings' as his term for the key elements of consciousness, a term freighted for me by my decades of debate with my dear friend Stevan Harnad, who uses the term as his name for what zombies don't have, placing him firmly on the side of Nagel and Chalmers and the others with allegiance to the Zombic Hunch. But it turns out that, for Solms, *feelings* are *functional.* They are what they are because of what they *do* to the whole operation of the brain/computer. This is one of the key ideas in the book, and while, as he acknowledges, he is not the first to press this position, he puts some novel twists on it that promise to make some real progress.

Cognitive scientists in general agree that the brain is a *sort* of computer; it isn't a radiator for cooling the blood and it isn't a dynamo. It is an *information processing system* of tremendous power that accomplishes its primary task — controlling the body in ways that enhance its chances of surviving to produce offspring — by extracting patterns from the torrent of 'input' signals it receives from transducers, patterns that can guide its 'output', which is another torrent of signals, effector or trigger signals, that can contract muscles or release hundreds of different chemical modulators, including many that create recursive cycles that refine the information available and the uses to which it is put. Is it a *digital* computer? Nobody knows, but even if it is, at some level, a digital computer, its architecture, and the parts it is made of, are profoundly unlike the architectures and parts of the digital computers we understand so well. This is what opens the door to romantic surmises about how the brain might — or must — escape the explanatory net of functionalism. Solms and I want to close that door, not by fiat, but by showing how the brain harnesses affect to get the many jobs done.

The brain is a massively parallel processor, millions of channels wide, but in its normal operation it creates something rather like the 'von Neumann bottleneck', a temporary dedication of large parts of the hardware to tightly focused *serial* processing on specific topics of current importance, the 'stream of consciousness' that confronts any theorist. *Why* does the brain go into this serial mode? Solms' answer is that it is (always) *uncertainty about what to do next* that provokes this extravagant activity. How does that work? Nobody knows — yet. But

instead of jumping to the conclusion that the brain (or the mind) is inexplicable in computational terms, cognitive scientists weigh in with ever more empirically informed, if still speculative, models that might be the key. Baars, Changeux, Dehaene, and their colleagues describe a global workspace; Graziano and his colleagues point to an attention schema that helps focus the serial processing. Solms says that they and the other modellers are looking in the wrong place. Yes, the frontal lobes of the cortex in interaction with the thalamus play a major role in articulating conscious content, but the driving force behind the stream of consciousness is in the brainstem, and more specifically in the reticular activating system and the PAG — the periaqueductal grey matter. For decades the reticular activating system has been seen as something rather like an on/off switch for consciousness, but Solms sees it, in partnership with the PAG, as the locus of control, mediating all the competing *affects* (or feelings) that determine *what happens next* in the brain. If the brain is a sort of computer, the brainstem is where its operating system resides. That is not how Solms puts it, but I think it is the upshot of his central claim.

> The PAG is the final common path to affective output. It must, therefore, in a word, 'choose' what is to be done next, once the various affective circuits and their associated conditioned behaviours have made their contributions to action… It must make these choices by evaluating the residual error signals that are relayed to it by the affect systems. It must judge their competing bids by the ultimate biological imperatives to survive and reproduce, with each error signal communicating its component need to it. In short, the PAG must set priorities for the next action sequence. (p. 137)

The operating systems of digital computers are breathtakingly convincing examples of intelligent design. If you want to *know why* an operating system has the features it does, there are experts who can tell you because they designed those features to accomplish very specific and clearly defined tasks. They can even give you mathematical proofs that their designs work, as long as the hardware doesn't break, and they have tested these designs rigorously on the hardware to confirm empirically what they already knew theoretically. Operating systems are totalitarian dictators of what happens next at all levels of digital computation. Such regimentation is what enables programmers to design software that runs as expected. But brains evolved by trial-and-error processes of natural selection, with no pre-planning or proof-of-concept demonstrations, so we should anticipate that control is achieved in brains by other processes, in particular by opponent

processes, not traffic cops or other autocrats. Solms, following in the pathfinding steps of Antonio Damasio and Jaak Panksepp, puts affect or emotional reaction in the principal controlling role. Hurley, Adams, and I were driven to much the same conclusion in *Inside Jokes: Using Humor to Reverse Engineer the Mind* (2011):

> We endorse the challenge of designing what we will call *emotional algorithms*… Our notion of emotional algorithms implies a control structure that relies on emotional states in competition and collaboration for inducing state changes in the systems to drive both its bodily *and cognitive* behavior, not algorithms that compute emotional content as if it were simply an output. (p. 86)

Solms and I, then, are in general agreement about the central role of affect or emotion in controlling the brain's operating system and I am happy to be informed by his claims about how the reticular activating system and PAG are the places where the affects get resolved and lead to action — which may not be musculoskeletal motion but further rounds of evaluation, redirection of attention, and so forth. Reflecting, and thinking (like Rodin's *Thinker*), is action too, and it needs to be motivated by the affective system just as fleeing or foraging does.

One unpersuasive feature of Solms' attempt to put affect in the central functional role is his insistence that affect cannot do its crucial work without being conscious.

> Apparently alone among mental functions, feeling is necessarily conscious. Who ever heard of a feeling that has no subjective quality? What would the point be of a feeling if you did not feel it? (Solms, 2021, p. 86; see also the discussion pp. 99ff.)

I see a spectrum of affective states, from unignorable pains and ecstatic attention-capturing joys through barely noticeable nudges of regret, boredom, confusion, annoyance, and on to the undetectable — but still influential — affective associations we have with colours, sounds, bodily postures, and all manner of cognitive/conative states. Hurley's fundamental and revolutionary insight into humour, in my opinion, was his identification of the micro-emotion of *Uh-oh,* a tiny smidgen of anxiety a millisecond or two before its cancellation or resolution with *Ha-ha*, the relief that gives humour its volt of energy, and makes most of us humour addicts. It was theory, not introspection, that suggested the existence of these micro-events to Hurley. You can't get a joke unless you are conscious of it, but are these micro-emotions conscious affects? Maybe there are reasons for saying so, but what of the (unconscious?) pains that motivate us to roll out of

damaging postures while we sleep? To deny that these are proper affects because they are unconscious is to vacate one's theory. And to insist that in fact these subjective events rouse us momentarily into a few milliseconds of conscious experience that are then utterly forgotten sets one off down a mirror-image path of vacuity — this is the challenge I raise about whether some effects are *Orwellian* or *Stalinesque* in Dennett (1991).

Another cul-de-sac is the path leading to the currently popular forms of panpsychism: everything is conscious. If every grain of sand has its smidgen of consciousness, consciousness loses all its interesting and important powers. (The acceptance of the Zombic Hunch provides a breathtakingly bonkers argument for panpsychism: since consciousness isn't *for* anything — all those functions could be performed by a zombie, you see — it couldn't have evolved, so it must have been there all along!) It is better to see consciousness as not an all-or-nothing phenomenon, but as an array of somewhat different phenomena, the 'higher' or more complex forms building on the lesser forms of sensitivity or sentience. Some of Solms' discussion suggests that he is amenable to, and even advancing, such a view, but like many others he sometimes lapses into the all-or-nothing mode, where the metaphorical lightbulb of consciousness must be either on or off. He has his reasons, but they are mainly pedagogical: keep things as simple as possible, even if it means dichotomizing when a gradual blur would be more accurate:

> The fact that voluntary behaviour must be conscious reveals the deepest biological function of feeling: it guides our behaviour in conditions of uncertainty. (p. 101)

This glosses over the prospect of behaviours that are not entirely automatized, and not involuntary either, but rather on the boundary between things we knowingly (consciously) and deliberately are doing and things we are doing inattentively, without any reflection. This is a common expository shortcut and it can ease the task of explaining very complex issues, but it can breathe life into the due-for-extinction idea of consciousness as the Cosmic Divide between Mind and everything else. For instance, Ginsburg and Jablonka's (2019) excellent exploration of the gradual accrual of consciousness over evolutionary history also occasionally succumbs, unintentionally, to raising the question of whether *it* (consciousness) is present yet at whatever stage they are examining, so irresistible is the presumption that as we ascend from carbon atom to poet, there is an answer — however

unknowable — to the question of whether consciousness is present or absent. This encourages the forlorn idea that, as functions and complexities are added, there is a moment of kindling or a 'critical mass' of complexity that leads to lift-off into an entirely novel phenomenon.

Nagel's seductive 'what it is like' formula has fostered this kind of thinking for decades, but perhaps a few observations about that formula will tarnish its reputation as the best touchstone of consciousness. What is it like to be an iPhone? It can do a kazillion things that would convince a time-traveller from the seventeenth century that there was something it is like to be it, but we know better, don't we? The iPhone doesn't *know* what it's like to be an iPhone (there is nobody home in the iPhone to know what it's like). Let's suppose then that it *is* like something to be a frog but then ask whether a frog *knows* what it's like to be a frog. What would that mean? What leverage do we get on understanding the life of a frog from asking ourselves this question? A frog can be startled, and this will provoke it to leap to safety. Does it need to *know what it's like to be startled* in order to leap to safety? It would be one thing if frogs could compare notes and say things to their fellow frogs like: 'It seems I'm not as skittish as you are', or 'Do sudden sounds bother you more than suddenly looming shadows?' My suggestion is that the fact that we humans *can* compare notes, and can *say* things about what it's like to be us, shouldn't inspire us to *assume* that any of the subtleties of human consciousness that enable us to tell each other these things are also present in frog consciousness. This is an empirical question; we should look for higher-order noticings of their noticings. If there is no evidence for such sophistications, then if we concede that 'there is something it is like to be a frog' we should add that frogs don't know any more about that than they do about what it's like to be a bat or a bartender. Wittgenstein famously wrote: 'If a lion could talk, we couldn't understand him.' No. If a lion could talk, we'd understand him just fine. He just wouldn't help us understand anything about actual lions.

There is one idea that has recently been moving out of the category *unthinkable* into the category of not just thinkable but promising. It has several importantly different variants. While Chris Frith (2007) and Anil Seth (2021) and a few other neuroscientists like to talk about conscious experience as 'controlled hallucination', some philosophers, among them Keith Frankish (2016) and I (2016), prefer to defend *illusionism*: our brains are designed (by evolutionary processes) to take advantage of a tightly controlled user-illusion that simplifies our

restless efforts to satisfy our many needs. In many quarters this proposal is brusquely — one might well say thoughtlessly — dismissed as an obviously incoherent, self-contradictory non-starter as a theory of consciousness or as an explanation of the hard problem (Chalmers, 1995). Solms, however, actually comes close to endorsing illusionism,[2] or maybe he does: 'Within the predictive coding framework, odd as it seems, what we perceive is a virtual reality constructed from the mind's own building materials' (p. 213). I hope that bringing illusionism into clearer focus will send a wake-up call to those who just can't take it seriously.

Solms tackles the traditional but dubious assumption of epistemic intimacy (which sounds more and more like papal infallibility as it is brandished by contemporary philosophers like Chalmers and Galen Strawson), and points to a mistake in Chalmers' thinking:

> So, Chalmers thinks the dual aspects of 'the phenomenal features of the world' and 'the physical' are in information itself — at its source — not in the equipment of the participant observer. This is like thinking that sweetness is intrinsic to the molecular structure of glucose. (p. 260)

His footnote following this quote cites the analogy with glucose drawn from Hurley, Dennett and Adams (2011), but he doesn't quite follow through all the way to the user-illusion. Solms nicely picks out 'the equipment of the participant observer'. It is *there* that the *illusion* of 'phenomenality' is to be found, in just the same way the sweetness of sugar is *not* to be found in the structure of glucose but in the dispositional structures activated by the detection of glucose. Do not try to sprinkle subjective honey on those informational structures, and do not try to attribute 'phenomenal' visual properties to the informational structures of the visual cortex ('mental paint' or *figment* as I have called it). What one is *motivated to* do by the occurrence of these states is what explains their 'phenomenality' and we have no privileged access to the machinery that thus motivates us. In the case of pains, this motivational feature is easy to recognize; in the case of sweetness, it is almost as easy; in the case of bright red, it is more subtle, but still there. Nicholas Humphrey (e.g. 1992; 2006; 2017) has been finding ways of saying this for decades, but it is still a hard point to get one's head around.

Solms finds a metaphorical way of evading the issue in his précis:

---

2 Michael Graziano and his colleagues provide another close encounter with illusionism. See Graziano *et al.* (2020) and Dennett (2020) for a discussion.

> The needs of complex organisms which can act differentially, in flexible ways, in variable contexts, are therefore 'colour-coded' or 'flavoured'. This provides at least one mechanistic imperative for qualia.

Yes, 'colour-coded' without any colours, 'flavoured' without any flavours. There is no second transduction, no *restoration* of the coded colours and flavours and sounds in a new medium in the brain, as if consciousness were a kind of super-duper smello-television with 'phenomenal' properties being appreciated by the audience, the inner witness. Philosophers who insist that their 'first-person point of view' gives them direct access to such properties are just wrong; they have *no idea* what physical features of their neural processes cause their confident judgments about their 'qualia'.

In a good passage Solms is duly cautious about this:

> As far as I know, I am never awake and responsive but phenomenally unconscious. The two things go together. When I wake up, I become aware of things. In fact, from where I am sitting, my inner consciousness feels like it *causes* my outer responsiveness, at least to some degree. It is typically when I notice things — that is, when I become conscious of them — that I respond to them intentionally. Presumably you are the same. (p. 66)

Yes, it 'feels like' the presence of a bright red *quale* in my mind *causes* me to believe I am seeing something bright red but this is a theorist's illusion about a user-illusion. I *do* believe that I am seeing something bright red, and you would have a very hard time persuading me that I am not seeing something bright red, but I have no privileged access to how it happens to be that I have this well-nigh unshakable conviction. Here is a question the doubters might ask themselves: couldn't evolution have found some clever ways of installing beneficial *user-illusions* in organisms that would enable them to respond under time pressure to patterns, to environmental challenges and opportunities of all sorts, dealing with a macroscopic behavioural world that was tracked by simplifications of their evolved imaginations, not a metaphysically or scientifically accurate depiction of how physical things really are? They would be the beneficiaries of these arrangements without having to understand them at all. Sellars' (1963) *manifest image*, the world we live in, is not presented to us as clouds of colourless particles, but as clumps of coloured solids, wet liquids, and invisible gusts of air and other gases, for instance. The colours that exist in that world, then, are a sort of illusion. We are *indirectly* but robustly acquainted with those properties of things in the manifest

image, but the conviction that we are *directly* acquainted with 'phenomenal properties' is a confused theorists' illusion about the benign first illusion.

I see Solms as an ally in the campaign to turn our backs on the 'hard problem' and get on with the extremely difficult, but not systematically impossible, set of problems about how the brain's affective operating system maintains order in the neural computers we are endowed with. One of the impediments to progress on this task of imaginative theorizing is a conviction that, as John Haugeland liked to say, 'computers don't give a damn' (e.g. 1998). He was speaking about digital computers with their autocratic operating systems, which are designed to churn away forever, with no room for caring about anything. He was right, but his observation does not mean that systems that *do* give a damn cannot be constructed on computers that don't. Since what matters is in the organizational structure — the information and how it is exploited — of the system modelled on the computer, we can use digital computers, in all their transparency, to compose models of more biologically realistic computers, in which affects run the operating systems, in which opponent processes composed of cellular-level agents that have agendas of their own (they give a damn — their lives depend on it) vie to keep the larger agent they compose safe and sound. Whether we keep them 'alive' in virtual environments or install them in real-world robots that have to fend for themselves, they will allow us to explore a different area in computational space, the area where consciousness is apt to be found — but there won't be anything like a spectral glow, or the emergence of 'qualia' that tells us we have succeeded. It will be the vulnerability of the agents we create, and their various abilities to track the things that matter to them and to fend for themselves, with anticipations and evaluations that improve their chances. I think that is what Friston is getting at with his free energy principle: it is a central part of the 'specs' for a conscious agent, not a model of how the agent manages to fulfil them.

References

Chalmers, D.J. (1995) Facing up to the hard problem of consciousness, *Journal of Consciousness Studies*, **2** (3), pp. 200–219.

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

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

Dennett, D.C. (2016) Illusionism as the obvious default theory of consciousness, *Journal of Consciousness Studies*, **23** (11–12), pp. 65–72. Reprinted in Frankish,

K. (ed.) (2017) *Illusionism as a Theory of Consciousness*, Exeter: Imprint Academic.
Dennett, D.C. (2020) On track to a standard model, *Cognitive Neuropsychology*, **37** (3–4), pp. 173–175. doi: 10.1080/02643294.2020.1731443.
Frankish, K. (2016) Illusionism as a theory of consciousness, *Journal of Consciousness Studies*, **23** (11–12), pp. 11–39. Reprinted in Frankish, K. (ed.) (2017) *Illusionism as a Theory of Consciousness*, Exeter: Imprint Academic.
Frith, C. (2007) *Making Up the Mind: How the Brain Creates our Mental World*, Oxford: Wiley-Blackwell.
Ginsburg, S. & Jablonka, E. (2019) *The Evolution of the Sensitive Soul: Learning and the Origins of Consciousness*, Cambridge, MA: MIT Press.
Graziano, M.S.A., Guterstam, A., Bio, B.J. & Wilterson, A.I. (2020) Toward a standard model of consciousness: Reconciling the attention schema, global workspace, higher-order thought, and illusionist theories, *Cognitive Neuropsychology*, **37** (3–4).
Haugeland, J. (1998) *Having Thought*, Cambridge, MA: Harvard University Press.
Humphrey, N. (1992) *A History of the Mind*, London: Chatto and Windus.
Humphrey, N. (2006) *Seeing Red: A Study in Consciousness*, Cambridge, MA: Harvard University Press.
Humphrey, N. (2017) The invention of consciousness, *Topoi*, 9 August, pp. 1–9. doi: 10.1007/s11245-017-9498-0.
Hurley, M., Dennett, D.C. & Adams, R. (2011) *Inside Jokes: Using Humor to Reverse Engineer the Mind*, Cambridge, MA: MIT Press.
Sellars, W. (1963) *Science, Perception and Reality*, London: Routledge & Kegan Paul.
Seth, A. (2021) *Being You: A New Science of Consciousness*, Boston, MA: Dutton.
Solms, M. (2021) *The Hidden Spring: A Journey to the Source of Consciousness*, New York: Norton.