This work originally appeared in:


This is Daniel C. Dennett’s final draft before publication. It has been modified to reflect the pagination of the published version of the work.
Dennett, Daniel C. In my opinion, the two main topics in the philosophy of mind are CONTENT and CONSCIOUSNESS. As the title of my first book, *Content and Consciousness* (1969) suggested, that is the order in which they must be addressed: first, a theory of content or INTENTIONALITY - a phenomenon more fundamental than consciousness - and then, building on that foundation, a theory of consciousness. Over the years I have found myself recapitulating this basic structure twice, partly in order to respond to various philosophical objections, but more importantly, because my research on foundational issues in cognitive science led me into different aspects of the problems. The articles in the first half of *Brainstorms* (1978a) composed in effect a more detailed theory of content, and the articles in the second half were concerned with specific problems of consciousness. The second re-capitulation has just been completed, with a separate volume devoted to each half: The *Intentional Stance* (1987a) is all and only about content; *Consciousness Explained* (1991a) presupposes the theory of content in that volume and builds an expanded theory of consciousness.

BEGINNINGS AND SOURCES

Although quite a few philosophers agree that content and consciousness are the two main issues confronting the philosophy of mind, many - perhaps most - follow tradition in favouring the opposite order: consciousness, they think, is the fundamental phenomenon, upon which all intentionality ultimately depends. This difference of perspective is fundamental, infecting the intuitions with which all theorizing must begin, and it is thus the source of some of the deepest and most persistent disagreements in the field. It is clear to me how I came by my renegade vision of the order of dependence: as a graduate student at Oxford, I developed a deep distrust of the methods I saw other philosophers employing, and decided that before I could trust any of my intuitions about the mind, I had to figure out how the brain could possibly accomplish the mind’s work. I knew next to nothing about the relevant science, but I had always been fascinated with how things worked - clocks, engines, magic tricks. (In fact, had I not been raised in a dyed-in-the-wool 'arts and humanities' academic family, I probably would have become an engineer, but this option would never have occurred to anyone in our family.) So I began educating myself, always with an eye to the curious question of how the mechanical responses of 'stupid' neurons could be knit into a fabric of activity that actually discriminated meanings. Somehow it had to be possible, I assumed. Since it was obvious to me that DUALISM was a last resort, to be postponed indefinitely.

So from the outset I worked from the 'third-person point of view' of science, and took my task to be building - or rather
sketching the outlines of a physical structure that could be seen to accomplish the puzzling legerdemain of the mind. At the time - the mid-60s - no one else in philosophy was attempting to build that structure, so it was a rather lonely enterprise, and most of the illumination and encouragement I could find came from the work of a few visionaries in science and engineering: Warren McCulloch, Donald MacKay, Donald Hebb, Ross Ashby, Allen Newell, Herbert Simon, and J. Z. Young come to mind. Miller, Galanter and Pribram's 1960 classic, *Plans and the Structure of Behaviour*, was a dimly understood but much appreciated beacon, and Michael Arbib's 1964 primer, *Brains, Machines and Mathematics*, was very helpful in clearing away some of the fog.

Given my lack of formal training in any science, this was a dubious enterprise, but I was usually forgiven my naïveté by those who helped me into their disciplines, and although at the time I considered myself driven by (indeed defined by) my disagreements with my philosophical mentors, QUINE and RYLE, in retrospect it is clear that my deep agreement with both of them about the nature of philosophy - so deep as to be utterly unexamined and tacit - was the primary source of such intellectual security as I had.

The first stable conclusion I reached, after I discovered that my speculative forays always wandered to the same place, was that the only thing brains could do was to approximate the responsibility to meanings that we presuppose in our everyday mentalistic discourse. When mechanical push came to shove, a brain was always going to do what it was caused to do by current, local, mechanical circumstances, whatever it ought to do, whatever a God's-eye view might reveal about the actual meanings of its current states. But over the long haul, brains could be designed - by evolutionary processes - to do the right thing (from the point of view of meaning) with high reliability. This found its first published expression in *Content and Consciousness* (1969, §9, 'Function and Content') and it remains the foundation of everything I have done since then. As I put it in *Brainstorms* (1978a), brains are syntactic engines that can mimic the competence of semantic engines. (See also the thought experiment - a forerunner of Searle's Chinese Room - about being locked in the control room of a giant robot, in 1978b.) Note how this point forces the order of dependence of consciousness on intentionality. The appreciation of meanings - their discrimination and delectation - is central to our vision of consciousness, but this conviction that I, on the inside, deal directly with meanings turns out to be something rather like a benign 'user illusion'. What Descartes thought was most certain - his immediate introspective grasp of the items of consciousness - turns out to be not even quite true, but rather a metaphorical by-product of the way our brains do their approximating work. This vision tied in beautifully with a doctrine of Quine's that I had actually vehemently resisted as an undergraduate: the indeterminacy of radical translation. I could now see why, as Quine famously insisted, indeterminacy was 'of a piece with' Brentano's thesis of the irreducibility of the intentional, and why those irreducible intentional contexts were unavoidably a 'dramatic idiom' rather than an expression of unvarnished truth. I could also see how to re-interpret the two philosophical works on intentionality that had had the most influence on me, Anscombe's *Intention* (1957) and Taylor's *The Explanation of Behaviour* (1964).

If your initial allegiance is to the physical sciences and the third-person point of view, this disposition of the issues can seem not just intuitively acceptable, but inevitable, satisfying, natural. If on the other hand your starting point is the traditional philosophical allegiance to the mind and the deliverances of introspection, this vision can seem outrageous. Perhaps the clearest view of this watershed of intuitions can be obtained from an evolutionary perspective. There was a time, before life on earth, when there was neither intentionality nor consciousness, but eventually replication got under way and simple organisms emerged. 237
Suppose we ask of them: Were they conscious?
Did their states exhibit intentionality?
It all depends on what these key terms are taken to mean, of course, but underneath the strategic decisions one might make about pre-emptive definition of terms lies a fundamental difference of outlook.
One family of intuitions is comfortable declaring that while these earliest ancestors were unconscious automata, not metaphysically different from thermostats or simple robotic toys, some of their states were nevertheless semantically evaluable. These organisms were, in my terms, rudimentary intentional systems, and somewhere in the intervening ascent of complexity, a special subset of intentional systems has emerged: the subset of conscious beings. According to this vision, then, the intentionality of our unconscious ancestors was as real as intentionality ever gets; it was just rudimentary.
It is on this foundation of unconscious intentionality that the higher-order complexities developed that have culminated in what we call consciousness. The other family of intuitions declares that if these early organisms were mere unconscious automata, then their so-called intentionality was not the real thing. Some philosophers of this persuasion are tempted to insist that the earliest living organisms were conscious - they were alive, after all - and hence their rudimentary intentionality was genuine, while others suppose that somewhere higher on the scale of complexity, real consciousness, and hence real intentionality, emerges. There is widespread agreement in this camp, in any case, that although a robot might be what I have called an intentional system, and even a higher-order intentional system, it could not be conscious, and so it could have no genuine intentionality at all.

In my first book, I attempted to cut through this difference in intuitions by proposing a division of the concept of consciousness into awareness, the fancy sort of consciousness that we human beings enjoy, and awareness2, the mere capacity for appropriate responsivity to stimuli, a capacity enjoyed by honey bees and thermostats alike. The tactic did not work for many thinkers, who continued to harbour the hunch that I was leaving something out; there was, they thought, a special sort of sensitivity - we might call it animal consciousness - that no thermostat or fancy robot could enjoy, but that all mammals and birds (and perhaps all fish, reptiles, insects, molluscs …) shared. Since robotic devices of considerably greater behavioural and perceptual complexity than the simplest of these organisms are deemed unconscious by this school of thought, it amounts to some sort of latter-day vitalism. The more one learns about how simple organisms actually work, the more dubious this hunch about a special, organic sort of sensation becomes, but to those who refuse to look at the science, it is a traditional idea that is about as comfortable today as it was in the seventeenth century, when many were horrified by Descartes's claims about the mechanicity of (non-human) animals. In any event, definitional gambits are ineffective against it, so in later work I dropped the tactic and the nomenclature of 'aware1' and 'aware2' - but not the underlying intuitions.

My accounts of content and consciousness have subsequently been revised in rather minor ways and elaborated in rather major ways. Some themes that figured heavily in Content and Consciousness lay dormant in my work through the 70s and early 80s, but were never abandoned, and are now re-emerging, in particular the theme of learning as evolution in the brain and the theme of content being anchored in distributed patterns of individually ambiguous nodes in networks of neurons. The truth is that while I can fairly claim to have seen the beauty, and indeed the inevitability, of these ideas in Content and Consciousness (see also Dennett. 1974), and to have sketched out their philosophical implications quite accurately, I simply couldn't see how to push them further in the scientific domain, and had to wait for others - not philosophers - to discover these ideas for themselves and push them in the new directions that have so properly captured recent philosophical attention. My own recent dis-
cussions of these two themes are to be found in Dennett (1986, 1987b, 1991a, 1991b, 1991c, 1992a).

CONTENT: PATTERNS VISIBLE FROM THE INTENTIONAL STANCE

My theory of content is functionalist (see FUNCTIONALISM): all attributions of content are founded on an appreciation of the functional roles of the items in question in the biological economy of the organism (or the engineering of the robot). This is a specifically 'teleological' notion of function (not the notion of a mathematical function or of a mere 'causal role', as suggested by David LEWIS and others). It is the concept of function that is ubiquitous in engineering, in the design of artefacts, but also in biology. (It is only slowly dawning on philosophers of science that biology is not a science like physics, in which one should strive to find 'laws of nature', but a species of engineering: the analysis, by 'reverse engineering', of the found artefacts of nature - which are composed of thousands of deliciously complicated gadgets, yoked together opportunistically but elegantly into robust, self-protective systems.) These themes were all present in Content and Consciousness, but they were clarified in 'Intentional Systems' (1971) when I introduced the idea that an intentional system was, by definition, anything that was amenable to analysis by a certain tactic, which I called the intentional stance. This is the tactic of interpreting an entity by adopting the presupposition that it is an approximation of the ideal of an optimally designed (i.e. rational) self-regarding agent. No attempt is made to confirm or disconfirm this presupposition, nor is it necessary to try to specify, in advance of specific analyses, wherein consists RATIONALITY. Rather, the presupposition provides leverage for generating specific predictions of behaviour, via defeasible hypotheses about the content of the control states of the entity.

My initial analysis of the intentional stance and its relation to the design stance and physical stance was addressed to a traditional philosophical issue - the problem of free will and the task of reconciling mechanism and responsibility (1973). The details, however, grew out of my reflections on practices and attitudes I observed to be ubiquitous in ARTIFICIAL INTELLIGENCE. Both Allen Newell (1982) and David Marr (1982) arrived at essentially the same breakdown of stances in their own reflections on the foundations of cognitive science. The concept of intentional systems (and particularly, higher-order intentional systems) has been successfully exploited in clinical and developmental psychology, ethology, and other domains of cognitive science, but philosophers have been reluctant to endorse the main metaphysical implications of the theory (see COGNITIVE PSYCHOLOGY; DEVELOPMENTAL PSYCHOLOGY).

In particular, I have held that since any attributions of function necessarily invoke optimality or rationality assumptions, the attributions of intentionality that depend on them are interpretations of the phenomena - a 'heuristic overlay' (1969), describing an inescapably idealized 'real pattern' (1991d). Like such abstracta as centres of gravity and parallelograms of force, the BELIEFS and DESIRES posited by the highest stance have no independent and concrete existence, and since this is the case, there would be no deeper facts that could settle the issue if - most improbably - rival intentional interpretations arose that did equally well at rationalizing the history of behaviour of an entity. Quine's thesis of the indeterminacy of radical translation carries all the way in, as the thesis of the indeterminacy of radical interpretation of mental states and processes.

The fact that cases of radical indeterminacy, though possible in principle, are vanishingly unlikely ever to confront us is small solace, apparently. This idea is deeply counterintuitive to many philosophers, who have hankered for more 'realistic' doctrines. There are two different strands of 'realism' that I have tried to undermine:

(1) realism about the entities purportedly described by our everyday mentalistic
discourse, what I dubbed FOLKPSYCHOLOGY (1981) - such as beliefs, desires, pains, the self; realism about content itself - the idea that there have to be events or entities that really have intentionality (as opposed to the events and entities that only behave as if they had intentionality).

Against (1), I have wielded various arguments, analogies, parables. Consider what we should tell the benighted community of people who speak of 'having fatigues' where we speak of being tired, exhausted, etc. (1978a). They want us to tell them what fatigues are, what bodily states or events they are identical with, and so forth. This is a confusion that calls for diplomacy, not philosophical discovery; the choice between an 'eliminative materialism' (see ELIMINATIVISM) and an 'IDENTITY THEORY' of fatigues is not a matter of which 'ism' is right, but of which way of speaking is most apt to wean these people of a misbegotten feature of their conceptual scheme.

Against (2), my attack has been more indirect. I view the philosophers' demand for content realism as an instance of a common philosophical mistake: philosophers often manoeuvre themselves into a position from which they can see only two alternatives: infinite regress versus some sort of 'intrinsic' foundation - a prime mover of one sort or another. For instance, it has seemed obvious that for some things to be valuable as means, other things must be intrinsically valuable - ends in themselves - otherwise we'd be stuck with a vicious regress (or circle) of things valuable only as means. It has seemed similarly obvious that although some intentionality is 'derived' (the aboutness of the pencil marks composing a shopping list is derived from the intentions of the person whose list it is), unless some intentionality is original and underived, there could be no derived intentionality. There is always another alternative, which naturalistic philosophers should look on with favour: a finite regress that peters out without marked foundations or thresholds or essences. Here is an easily avoided paradox: every mammal has a mammal for a mother - but this implies an infinite genealogy of mammals, which cannot be the case. The solution is not to search for an essence of mammalhood that would permit us in principle to identify the Prime Mammal, but rather to tolerate a finite regress that connects mammals to their non-mammalian ancestors by a sequence that can only be partitioned arbitrarily. The reality of today's mammals is secure without foundations.

The best known instance of this theme in my work is the idea that the way to explain the miraculous-seeming powers of an intelligent intentional system is to decompose it into hierarchically structured teams of ever more stupid intentional systems, ultimately discharging all intelligence-debts in a fabric of stupid mechanisms (1971, 1974, 1978a, 1991a). Lycan (1981) has called this view homuncular functionalism. One may be tempted to ask: are the subpersonal components real intentional systems? At what point in the diminution of prowess as we descend to simple neurons does real intentionality disappear? Don't ask. The reasons for regarding an individual neuron (or a thermostat) as an intentional system are unimpressive, but not zero, and the security of our intentional attributions at the highest levels does not depend on our identifying a lowest-level of real intentionality. Another exploitation of the same idea is found in Elbow Room (1984): at what point in evolutionary history did real reason-appreciators, real selves, make their appearance? Don't ask - for the same reason. Here is yet another, more fundamental, version: at what point in the early days of evolution can we speak of genuine function, genuine selection-for and not mere fortuitous preservation of entities that happen to have some self-replicative capacity? Don't ask. Many of the most interesting and important features of our world have emerged, gradually, from a world that initially lacked them - function, intentionality, consciousness, morality, value - and it is a fool's errand to
try to identify a first or most-simple instance of the 'real' thing. It is for the same reason a mistake to suppose that real differences in the world must exist to answer all the questions our systems of content attribution permit us to ask. Tom says he has an older brother living in Cleveland and that he is an only child (1975b). What does he really believe? Could he really believe that he had a brother if he also believed he was an only child? What is the real content of his mental state? There is no reason to suppose there is a principled answer.

The most sweeping conclusion I have drawn from this theory of content is that the large and well-regarded literature on PROPOSITIONAL ATTITUDES (especially the debates over wide versus narrow content, 'de re versus de dicto' attributions, and what Pierre believes about London) is largely a disciplinary artefact of no long-term importance whatever, except perhaps as history's most slowly unwinding unintended reductio ad absurdum. By and large the disagreements explored in that literature cannot even be given an initial expression unless one takes on the assumptions I have argued are fundamentally unsound (see especially 1975b, 1978a, 1982, 1987b, 1991d): strong realism about content, and its constant companion, the idea of a LANGUAGE OF THOUGHT, a system of mental representation that is decomposable into elements rather like terms, and larger elements rather like sentences. The illusion that this is plausible, or even inevitable, is particularly fostered by the philosophers' normal tactic of working from examples of 'believing-that-p' that focus attention on mental states that are directly or indirectly language-infected, such as believing that the shortest spy is a spy, or believing that snow is white. (Do polar bears believe that snow is white? In the way we do?) There are such states - in language-using human beings - but they are not exemplary or foundational states of belief; needing a term for them, I call them opinions (How to Change your Mind', in 1978a; see also 1991c). Opinions play a large, perhaps even decisive, role in our concept of a person, but they are not paradigms of the sort of cognitive element to which one can assign content in the first instance. If one starts, as one should, with the cognitive states and events occurring in non-human animals, and uses these as the foundation on which to build theories of human cognition, the language-infected states are more readily seen to be derived, less directly implicated in the explanation of behaviour, and the chief but illicit source of plausibility of the doctrine of a language of thought. Postulating a language of thought is in any event a postponement of the central problems of content ascription, not a necessary first step. (Although a few philosophers - especially Millikan, Robert STALNAKER, Stephen White - have agreed with me about large parts of this sweeping criticism, they have sought less radical accommodations with the prevailing literature.)

CONSCIOUSNESS AS A VIRTUAL MACHINE

My theory of consciousness has undergone more revisions over the years than my theory of content. In Content and Consciousness the theory concentrated on the role of language in constituting the peculiar but definitive characteristics of human consciousness, and while I continue to argue for a crucial role of natural language in generating the central features of consciousness (our kind), my first version overstated the case in several regards. For instance, I went slightly too far in my dismissal of mental imagery (see the corrections in 1978a, 1991a), and I went slightly too fast - but not too far! - in my treatment of colour vision, which was unconvincing at the time, even though it made all the right moves, as recent philosophical work on colour has confirmed, in my opinion. But my biggest mistake in Content and Consciousness was positing a watershed somewhere in the brain, the 'awareness line', with the following property: revisions of content that occurred prior to crossing the awareness line changed the content of consciousness; later revisions (or errors) counted as postexperiential tamperings; all adjustments of 241
content, veridical or not, could be located, in principle, on one side or the other of this postulated line. The first breach of this intuitive but ultimately indefensible doctrine occurred in 'Are Dreams Experiences?' (1975a), in which I argued that the distinction between proper and improper entry into memory (and thence into introspective report, for instance) could not be sustained in close quarters. Related arguments appeared in 'Two Approaches to Mental Imagery' (in 1978a) and 'Quining Qualia' (1988), but only in *Consciousness Explained* (1991a) and 'Time and the Observer' (Dennett and Kinsbourne, 1992) was an alternative positive model of consciousness sketched in any detail, the Multiple Drafts model.

The best way to understand this model is in contrast to the traditional model, which I call the Cartesian Theatre. The fundamental work done by any observer can be characterized as confronting something 'given' and taking it - responding to it with one interpretive judgment or another. This corner must be turned somehow and somewhere in any model of consciousness. On the traditional view, all the taking is deferred until the raw given, the raw materials of stimulation, have been processed in various ways and sent to central headquarters. Once each bit is 'finished' it can enter consciousness and be appreciated for the first time. As C. S. Sherrington (1934) put it:

> The mental action lies buried in the brain, and in that part most deeply recessed from outside world that is furthest from input and output.

In the Multiple Drafts model, this single unified taking is broken up in cerebral space and real time; the judgmental tasks are fragmented into many distributed moments of micro-taking (Dennett and Kinsbourne, 1992). Since there is no place where 'it all comes together', no line the crossing of which is definitive of the end of preconscious processing and the beginning of conscious appreciation, many of the familiar philosophical assumptions about the denizens of human phenomenology turn out to be simply wrong, in spite of their traditional obviousness.

For instance, from the perspective provided by this model one can see more clearly the incoherence of the absolutist assumptions that make QUALIA seem like a good theoretical idea. It follows from the Multiple Drafts model that 'inverted spectrum' and 'absent qualia' thought experiments, like the thought experiments encountered in the propositional attitude literature (Twin Earth, what Pierre believes, beliefs about the shortest spy), are fundamentally misbegotten, and for a similar reason: the 'common sense' assumption of 'realism' with regard to the mental items in question - beliefs, in the first instance, qualia, in the second - is too strong.

**OVERVIEW**

The intermediate ontological position I recommend - I call it 'mild realism' - might be viewed as my attempt at a friendly amendment to Ryle's (1949) tantalizing but unpersuasive claims about category mistakes and different senses of 'exist' (see especially 1969, ch. 1; 1991d). What do you get when you cross a Quine with a Ryle? A Dennett, apparently. But there is a novel texture to my work, and an attitude, which grows primarily, I think, from my paying attention to the actual details or the sciences of the mind - and asking philosophical questions about those details. This base camp in the sciences has permitted me to launch a host of differently posed arguments, drawing on overlooked considerations. These arguments do not simply add another round to the cycle of debate, but have some hope of dislodging the traditional intuitions with which philosophers previously had to start. For instance, from this vantage point one can see the importance of evolutionary models (1969, 1974, 1978a, 1983, 1984a, 1990b, 1991f) and, concomitantly, the perspective of cognitive science as reverse engineering (1989, 1991, 1992a), which goes a long way to over-
coming the conservative mindset of pure philosophy. The idea that a mind could be a contraption composed of hundreds or thousands of gadgets takes us a big step away from the overly familiar mind presupposed by essentially all philosophers from Descartes to the present.

Something else of mine that owes a debt to Quine and Ryle is my philosophical style. No sentence from Quine or Ryle is ever dull, and their work always exhibits the importance of addressing an audience of non-philosophers. even when they know that philosophers will be perhaps 95% of their actual and sought-for audience. They also both embody a healthy scepticism about the traditional methods and presuppositions of our so-called discipline, an attitude to which I have always resonated. I have amplified these points, attempting to follow their example in my own writing. But I have also been self-conscious about philosophical methods and their fruits, and presented my reflections in various meta-level digressions, in particular about the role of intuition pumps in philosophy (1980, 1984a, 1991a). and about the besetting foible of philosophers: mistaking failures of imagination for insights into necessity.

My insistence on the need for philosophers to stoke up on the relevant science before holding forth, and my refusal to conduct my investigations by the traditional method of definition and formal argument, have made me a distinctly impure philosopher of mind. Moreover, on both main topics, content and consciousness, I maintain a 'radical' position, which in a rather lonely and implausible fashion declares that much of the work at the presumed cutting edge is beyond salvage. I thus cut myself off from some of the controversies that capture the imaginations of others in the field, but the philosophical problems that arise directly in non-philosophical research in cognitive science strike me as much more interesting, challenging, and substantive. So I concentrate on them: the frame problem (1984b, 1991e), problems about mental imagery and 'filling in' (1992b), the binding problem and the problem of temporal anomalies (1991a; Dennett and Kinsbourne, 1992). I take these to be the real, as opposed to artefactual, problems of mental representation, and I encourage philosophers of mind to contribute to their solution.

BIBLIOGRAPHY


Miller, G., Galanter, E., and Pribram, K. 1960. 244

DANIEL C. DENNETT