

How a neural correlate can function as an explanation of consciousness

*Evidence from the history of science regarding the
likely explanatory value of the NCC approach*

Ilya Farber

George Washington University / Singapore Management University

Abstract: *A frequent criticism of the neuroscientific approach to consciousness is that its theories describe only “correlates” or “analogues” of consciousness, and so fail to address the nature of consciousness itself. Despite its apparent logical simplicity, this criticism in fact relies on some substantive assumptions about the nature and evolution of scientific explanations. In particular, it is usually assumed that, in expressing **correlations**, neural correlate of consciousness (NCC) theories must fail to capture the **causal** structure relating brain and mind. Drawing on work in the history and philosophy of science, I argue that this assumption – along with the related claim that even a correct NCC theory would fail to **explain** consciousness – is grounded in an inadequate conception of the way in which scientific explanations develop. Examination of parallel developments in 20th century biology reveals that, under the right circumstances, seemingly crude correspondences can play an essential role in scientific discovery and can sometimes become central to our everyday understanding of the phenomena in question. A proper understanding of this process clarifies the value of NCC theories and sheds light on the standards by which they should be evaluated. In closing, I describe two specific criteria for evaluating NCC proposals: **intertheoretic bridge potential** and **detailed mapping**.*

1. Does a neural correlate of consciousness *explain* anything?

Neuroscientists who aspire to explain consciousness face a daunting array of methodological hurdles. Like all their peers in cognitive neuroscience, they must guard against criticism from above and below: neuroanatomists may accuse them of over-interpreting the data, hastily generalizing across species, and abandoning the realm of the empirical for that of the speculative; psychologists and philosophers may lodge charges of reductionism and determinism. This multi-front battle is challenging, but in fighting it the neuroscientist can fall back on the usual canons of scientific virtue: define your terms, gather lots of data, eliminate confounding factors, and develop tests to distinguish your hypothesis from each new competitor. This is a game that conceptually ambitious scientists in any discipline must learn to play.

There is a different sort of criticism, however, that seems to be the special burden of those who would propose a neural correlate of consciousness (NCC).¹ “That’s a very nice theory you have there,” their critics may say, “but you must realize that it doesn’t explain consciousness

¹ I have in mind here scientists such as Francis Crick, Christof Koch, Rodolfo Llinás, Gerald Edelman and Wolf Singer – people who have advanced specific proposals about the neural structures and activities that underlie conscious sensory awareness. The NCC terminology is due to Crick and Koch, who have explained it as “the minimal set of neuronal events that gives rise to a specific aspect of a conscious percept” (Crick and Koch 2003). Other theorists with similar methods and goals may use different terms, or may understand the term or the overall project somewhat differently, but nothing in this discussion should depend on those differences.

at all. We've known at least since Hume that correlation is not causation; you may have established what goes on in the brain *at the same time as* consciousness, but that tells us nothing about the nature of consciousness itself."

It's important to understand what a radical criticism this is. It doesn't allege that the theory is false; it says that, even if true, the theory is simply *irrelevant* to the fundamental problem of consciousness. What we want to know, the critics say, is what consciousness is and where it comes from; if indeed it comes from the brain, we want to know *how* and *why* it comes from the brain. It seems that NCC theories – however well-supported by evidence – cannot answer these questions, but merely recast them in narrower terms: "How and why does the brain produce consciousness?" becomes "How and why does gamma-band oscillation in visual cortex produce subjective visual experience?"

The logic of this argument seems simple, and many find it compelling; it lies at the core of some of the most influential and widely-cited critiques of the NCC project (e.g. Searle 1992, Chalmers 1995). At root, it seems to be a variant on the old (and eternally relevant) point that many different causal models can underlie an observed pattern. Deep Humean worries aside, statisticians have taught us to be wary of quick inferences from correlation to causation. An observed correlation between neural activity and conscious experience *could* result from the former causing the latter; but it is equally compatible with consciousness causing neural activity, with some third factor causing them both, or with a variety of other system-level relationships.

Consider a more mundane example. If a correlation were found between coffee consumption and mean annual risk of death by heart failure, one *might* conclude that the coffee is causing the heart attacks, but one would have to conduct further investigations to test that causal hypothesis ($C \rightarrow H$) against others, such as the possibility that patients with ailing hearts feel sluggish and drink more coffee to compensate ($H \rightarrow C$) or that "type A" genes or a stressful lifestyle predispose one both to coffee drinking and to heart attacks ($A \rightarrow (C \& H)$). From this perspective, an NCC theorist who claims to have explained the causal basis of consciousness seems to be making an elementary logical error.

It's even fairly easy to see why this problem plagues NCC research in particular. The gold standard for resolving causal ambiguity is the elucidation of a *mechanism* that produces the observed correlation. Medicine in particular has become very sensitive to this issue: epidemiologists may use population data to guide their initial search, but once a risk factor has been identified, they turn to other forms of research to establish a plausible mechanism. To truly settle the question an explanatory chain must be constructed between the putative cause and the observed disorder, identifying a series of intermediate steps which may pass through the domains of sociology, immunology, genetics, biochemistry, and other disciplines. In the coffee case, researchers arguing for the $C \rightarrow H$ model would try to elucidate the precise physiochemical pathway from coffee to cardiac arrest. As it turns out, one can draw several direct connections: coffee contains caffeine, which blocks the effects of adenosine, leading to an increase in both heart rate and blood pressure. Caffeine also increases calcium availability in heart muscle tissue by opening membrane channels in the sarcoplasmic reticulum, an intracellular storage structure;

calcium plays a profound role in all nervous and contractile tissue, and increased calcium levels in heart tissue can speed and destabilize the cardiac rhythm.²

This detailed mechanistic explanation is possible because the entities involved all belong to the same approximate domain: call it the domain of “middle-sized organic objects and their biochemical constituents.” Even though no specific scientific discipline encompasses all of the entities involved, there are several “chains” of overlapping disciplines connecting coffee and heart function (for example, organic chemistry to biochemistry to physiology to cardiology), and each step in the causal story can be tested using data and methods from the relevant fields.

It’s hard to imagine where one would even begin constructing such a bridge between (say) synchronous oscillations in sensory cortex and subjective perceptual experience. “The neurons oscillate together, which in turn” ... what? There are no – or at least, too few – intermediate disciplines to bridge the gap between neurons and experience. And if, as a consequence, we can’t fill out the steps of the causal story, then it seems that even our best neural models won’t be able to give us the sort of satisfying causal explanation we want; they won’t be able to tell us *how* or *why* the observed psychophysical correlations hold.³

But at this point, one must ask: is the NCC theorist really in the same position as an epidemiologist studying population variables? Unlike the epidemiologist, the neuroscientist is in a position to intervene directly in the systems under study (or, in humans, to examine cases where nature or accidents have “intervened” to cause brain damage). The experiments that underlie modern NCC theories generally involve *manipulating* perceptual stimuli, in realtime, and observing the resultant changes in neural activity; recent technological advances have also made it increasingly possible to run things in the opposite direction, manipulating the neural activity (e.g. via genetic knockouts or transient magnetic stimulation) and measuring the resultant changes in perceptual behavior or subjectively reported experience. This method of establishing causal linkage is at least as old and well-entrenched as the explication-of-mechanism method; it’s fallible, to be sure, but so are they all. So it seems that the neuroscientist has a fairly strong answer to the charge of mistaking correlation for causation: when you can manipulate the brain (via stimulation or ablation) and reliably produce predictable changes in consciousness, *and* you can manipulate consciousness (via perceptual stimulation or direct instruction) and reliably produce predictable changes in the brain, the worry about non-causal correlation seems vacuous, even in the absence of an articulated causal story. Unless and until somebody can make a useful model out of preestablished harmony or epiphenomenalism, there doesn’t seem to be any important alternative to the hypothesis that there is some direct causal relationship between neural activity and subjective experience.

This defense is valid, but at the same time it’s also a bit of a dodge. After all, the central worry raised by modern NCC skeptics isn’t that the mind might be causally unrelated to the brain; even the most committed modern dualist will admit that consciousness can be altered by modifications of the brain. The real worry is that neural theories describe a relationship without

² The reader may be relieved to learn that in this case, despite the presence of a plausible mechanism, the correlation itself has not been observed: even fairly high levels of coffee consumption do not significantly increase the risk of heart failure in normal subjects.

³ John Searle exemplifies this concern in *The Mystery of Consciousness* (1997): he surveys a number of scientific models of consciousness in separate essays, often with some excitement, but at the end of each one he winds up concluding more or less the same thing: that they’ve completely failed to address the fundamental question of “how the brain produces consciousness,” and hence that “the mystery remains.”

explaining it: they may tell us *that* certain states or modifications of consciousness arise from certain states or modifications of the brain, but they don't tell us *why*. In other words, the question isn't whether NCC theories are accurate, but whether they are *explanatory*.

To address this more worrisome challenge, we will need to look a little more closely at the structure of the theories themselves. Two questions in particular need answering: what exactly do theorists *mean* by the claim that some neural structure or process is the neural correlate of consciousness? And how, even in principle, could a proposal of this sort do the work of scientific explanation?

First off, it's important to realize that there's something misleading about the "correlate" part of "neural correlate of consciousness." The term naturally focuses attention on correlational sources of evidence for NCCs, such as the widely-cited experiments in which Nikos Logothetis found single cells whose activity covaried with monkeys' behaviorally reported percepts (Logothetis 1998). To treat the theories themselves as having this structure, however, is to mistake noun for verb. An NCC is *a correlate*, a *thing which corresponds* to consciousness; moment-to-moment psychophysical correlation is just one of many elements in this correspondence. All of the major NCC theories in fact draw multiple parallels, based on functional anatomy and pathology, on inter-species comparisons, and on interactions between the hypothesized NCC and other neural mechanisms which underlie related phenomena such as memory, dreaming and emotion. The relation which these theories establish between the neural and phenomenal aspects of consciousness is thus something much richer and more complex than mere correlation; it is a type of *isomorphism*, a multidimensional mapping between entities, structures and dynamics in the twin domains of mind and brain.

Once this is recognized, it becomes apparent just how badly the "correlation vs. causation" rubric distorts the debate, since neither option accurately represents the ambitions of NCC theorists. The goal of the NCC project is not to produce a causal model on which consciousness stands apart as a product of the brain, but rather to find the patterns of consciousness *within* the structure and dynamics of the brain. The methodology for pursuing this goal has already been charted out by researchers studying memory and perception: in roughest outline, it involves functionally decomposing the cognitive process in question, functionally and physically decomposing the brain, and trying to find matching patterns amidst the bits on each side.⁴ This process is fundamentally *analogical* rather than correlational, and the relation that it attempts to establish is not one of causal interaction but one of *identity*.

Though clear enough in practice, the idea that NCC theories are identity theories rather than causal-interaction theories has tended to get lost in the attendant philosophical discussions.⁵ Awareness of this point seems to be increasing, however, with one likely cause being the recent

⁴ By "functionally decomposing" I mean nothing more than analysis in the broad sense – that is, studying something complex by looking at its parts and their interactions. This implies no commitment to a philosophically "functionalist" account of mind, since functionalism incorporates additional restrictions on the form of analysis (e.g. that it be independent of physical substrate) which are incompatible with the NCC approach.

⁵ Searle has been the most tireless and influential proponent of the causal interpretation of the brain-mind relation, referring to consciousness throughout his work as a property of the brain which is caused by neural processes. In practice the brain plays little more than a decorative role in such accounts, and the sharp cause-effect dichotomy yields a purely nominal "naturalism" which puts no real pressure on the tradition of treating mind and brain as separate realms. Velmans (2002) is an example of how one can use causal language while remaining sensitive to these issues (see esp. his footnote 8); he poses questions about the (bidirectional) causal relations between specific neural and mental *events*, without treating *consciousness itself* as an "effect" of the brain.

efforts of Dan Dennett. His characterization of consciousness as “fame in the brain” is designed to emphasize the right sort of explanatory structure, and describes both the error and its alternative with characteristic precision:

Theorists are converging from quite different quarters on a version of the global neuronal workspace model of consciousness, but there are residual confusions to be dissolved. In particular, theorists must resist the temptation to see global accessibility as the *cause* of consciousness (as if consciousness were some other, further condition); rather, it *is* consciousness ... The proposed consensual thesis is not that this global availability *causes* some further effect of a different sort altogether – igniting the glow of conscious qualia, gaining entrance to the Cartesian Theater, or something like that – but that it *is*, all by itself, a conscious state. This is the hardest part of the thesis to understand and embrace. (Dennett 2001, pp. 221-223)

As scientific materialists, NCC theorists hold that mind *is* brain – not a “product” or “emergent property” of the brain, but *the same thing* described in different terms – and their theories are attempts to spell out how this is so.⁶

So much for identity; that point is in good hands, and I leave it there. My focus here will be on the other feature of the NCC methodology as described above: its analogical structure. As I see it, there are two barriers to widespread acceptance and understanding of this point. The first is that the role of analogical reasoning in science as a whole has been underappreciated, even by those who make frequent use of it, and so it has not yet found a place in the descriptive toolkit of scientists and philosophers of science. The second problem, perhaps more urgent for present purposes, is that such a method seems at first glance to be a poor choice for the project of constructing a metaphysically unified account of consciousness. If the goal is identity, why start with analogies rather than with precise, well-defined relationships (whether causal or definitional)?⁷

As it turns out, this is one of those questions that becomes self-answering once it’s seen from the right direction. To achieve this shift of perspective, however, it is first necessary to understand a set of advances that took place in the analysis of metaphor approximately thirty years ago.

⁶ Christof Koch does express reservations about physicalist identity theory, saying “I am not sure whether this sort of strong identity holds for the NCC and the associated percept ... The characters of brain states and of phenomenal states appear too different to be completely reducible to each other” (Koch 2004, p. 19). The point about reducibility I would wholeheartedly agree with, since I think the reduction concept does a very poor job of capturing the ways in which scientific domains are actually integrated (Farber 2000). Being “unsure” in the sense of modesty and fallibilism is also appropriate for a topic as vexing as this. Neither of these points, however, implies that NCC theory is *aiming at* something other than mind-brain identity. Unless and until someone discovers specific evidence for non-neural causal factors that stand between the brain and consciousness, identity will remain the more parsimonious and natural interpretation of NCC results.

⁷ A note on methodology: In this essay I will *not* be addressing the related but much broader questions that might occur to some readers at this point, such as why we should assume that a neural explanation of consciousness is possible at all or why we should favor such an approach over functionalism, dualism or what-have-you. That broader ground is already very well-trod (including in my own 1995, 2000 and 2001), and my intent here is to focus on refining our understanding of one particular approach rather than revisiting the foundational questions that dominate so much of the literature on consciousness.

2. Metaphors in science

The received view of metaphors⁸ – both in science and in language generally – had long been that they were convenient shortcuts, ways of compactly (if unreliably) expressing something that would require far more time and effort to express literally. It was believed that the role of scientific metaphors, such as “the atom is like a little solar system” or “planets curve space like balls on a rubber sheet,” was pedagogical: they assisted scientists in expressing complex technical ideas to students, or to layfolk who lacked the background required to understand the literal versions of the theories as expressed in definitions and mathematical laws. This paralleled the traditional analysis of metaphors in language as primarily poetic devices, tools whereby one broad statement (“war is hell”) could, given a particular context, stand in for one or more specific statements (“war is an environment of great suffering,” or perhaps “war is a realm full of people behaving immorally”).

Within linguistics, this view was first seriously challenged by Max Black (1962). Black argued for what he called “the interaction view of metaphor,” as opposed to the “substitution view” which held (as above) that metaphors merely stand in for some set of literal statements. The interaction view has three principal components:

- “1. A metaphorical statement has two distinct subjects, to be identified as the ‘primary’ subject and the ‘secondary’ one ...”⁹
- “2. The secondary subject is to be regarded as a system rather than an individual thing ...”
- “3. The metaphorical utterance works by ‘projecting upon’ the primary subject a set of ‘associated implications,’ comprised in the implicative complex, that are predicable of the secondary subject ...” (Black 1979, p. 28)

In other words, metaphors work by projecting some aspects of the secondary system (hell) onto the primary system (war). An important feature of this view is that it assigns the secondary subject an ongoing and hence ineliminable role in the semantic process; rather than being a simple “code” for some longer statement, a metaphor establishes a complex, implicit, sometimes dynamical relationship between two domains. Later researchers extended this line of thought to show that metaphor plays a similarly ineliminable role not just in language but in thought more generally (Lakoff and Johnson 1980). According to these theorists, the very structure of much human thought is best thought of as metaphorical, in that it works by assimilating unknown systems to known ones and then exploring the extent and validity of the implicative projections.

Boyd (1979) took the “interaction view” into philosophy of science, in a way consistent with his emphasis on the historical and dialectical nature of science. The resultant model assigns metaphors a central and irreducible role in scientific discourse. According to Boyd, the open-

⁸ Note that in this context “metaphor” refers to a class of linguistic and cognitive structures, and not to any specific grammatical form. In particular, the distinction between metaphor and simile is not relevant. “Analogy” could serve just as well, and is the more popular choice in recent work on this topic; I here use “metaphor” because it was the term used in the historical work which is discussed in this section, and also because it seems to more starkly express the philosophical problem which is at issue.

⁹ In recent work on metaphor, especially within cognitive science, it is now more common to refer to these as the “target” and “source” domains, respectively. This clarifies the nature of the asymmetry: knowledge about the secondary/source domain is used to explain or illuminate the primary/target domain.

endedness of metaphors is a virtue, for at least two related reasons. First, it allows scientists to put a name and tentative structure to some set of phenomena that they only vaguely understand: for example, the “atom as solar system” metaphor makes it possible to begin discussing and investigating electron “orbits” without committing researchers to any particular story about how exactly electrons are moving about the nucleus. And second, it allows metaphors to suggest directions for research, since negotiating the meaning of the metaphor requires working out which aspects of the systems do and do not correspond. (One might imagine an early atomic theorist mining astronomy texts for interesting research projects: “Are electrons captured in the way that planets capture moons? Are electron orbits subject to precession (like planets) or decay (like satellites)? Are there elliptical electron orbits? Do electrons nearer the nucleus orbit faster?” Note that these would be *empirical* projects in atomic physics, not questions of analogical convention or definition.)

Boyd presents the use of digital computer metaphors in cognitive science as an example of how the analogical links can be more fundamental than any of the properties that would be used to describe them. These metaphors can be traced back even farther, to Turing’s original formulation of an analogical model for human computation (figure 1). For anyone familiar with mid- to late 20th century psychology, it seems impossible to deny either of Boyd’s central claims: that these metaphors provided the structure for most theorists’ conceptions of the mind, and that they did so largely in the absence of any literal definitions of the concepts in question.

What can we infer from this example about theoretical role of NCCs? One clear lesson is that an analogy can be an important tool for developing our understanding of an abstract, mysterious domain. Neural analogies can help us to see new structure and distinctions within the realm of consciousness: for example, in the wake of anatomical localizations it has become a commonplace that spatial reasoning and linguistic reasoning are separate mental faculties. In my own experience, I found that learning that the olfactory system codes stimuli in an unusual way – by specific chemical identity and by overall similarity to specific known smells, rather than by some combination of broad categorial properties such as frequency or sweetness – helped me to better conceptualize the way that olfaction interacted with the other senses and with memory.

This role alone would make parallelisms between brain and consciousness worth pursuing – but at the same time, one must suspect that something more is supposed to be at stake. After all, NCC theorists choose to investigate the brain because they assume that it is directly involved in consciousness, not just because its structure provides for especially handy metaphors. Shouldn’t the fact that the mind is *made of* the brain make some difference here?

The historical limitations of the computer analogy provide at least the beginnings of an answer to this question. For all its metaphorical power, the computationalist approach in cognitive science relied on an assumption that was not borne out: the assumption that evolved, wet brains and designed, silicon computers were functionally equivalent at all the important levels of analysis. As it turned out, the differences are just too deep and too pervasive to be ignored as matters of “implementation.”¹⁰ The more we learn about the brain, the deeper the disanalogies seem to run (figure 2). This does not mean that the original metaphor has lost its utility; rather, it seems that we’ve run up against the *limits* of that utility, and consequently there is now less confidence that new and interesting mental structure can be discovered by looking for

¹⁰ See Farber, Peterman and Churchland (2001) for evidence that at least some important cognitive functions are deeply implementation-dependent at *every* interesting level of analysis.

parallels with known structures in the realm of symbolic computation. With respect to consciousness in particular, the prospects of developing a full-fledged model in terms of computer-inspired concepts such as “executive control” and “monitoring” now seem quite dim.

But again, is there any reason to believe that brain-based analogies will do any better? Does the fact that we’re made of neurons rather than logic gates really make any difference in the sorts of metaphors that we should seek, especially for something as abstruse as consciousness?

I believe that it does. To see why, it will be necessary to make one more brief detour, through the history of another field that has made extensive use of metaphors: the science of heredity. One way of understanding the novelty of Mendel’s approach is to look at its metaphorical structure (figure 3). Mendel reconceptualized the organism as a container, and its properties or “factors” (such as seed color and shape) as independent objects within that container. This metaphor provided him with a simple and mathematically tractable “marbles in boxes” system whose dynamics mirrored the dynamics of heredity.

The advantages and disadvantages of Mendel’s metaphor strongly parallel those of the computer metaphor in cognitive science. It provided a way of formalizing heredity, opening the way for abstract, even mathematical treatments of the subject; it postulated a set of entities and dynamics which scientists could start looking for in real organisms; and it provided the impetus for a number of research projects that can be understood as attempts to check whether, and how, various implied properties of the source domain could be extended to the target domain. At the same time, however, the source domain was *so* abstract that it made implementation seem utterly mysterious: how does an organism “contain” its characteristics, and how do those characteristics get randomly assorted and copied (via intermediaries which do not exhibit them) into succeeding generations? In treating the organism as a static container, Mendel’s framework left insufficient room for questions about the relations between heredity and development.

T. H. Morgan remedied this defect by redefining “factor” as first and foremost a physical entity: a physical part of the organism, passed from parent to child, which guided the organism’s development. (Its precise physical identity was left open, though Morgan eventually accepted the identification of factors with loci on the chromosomes.) This allowed him to integrate Mendel’s organism-as-container metaphor with the organism-as-machine metaphor which had been gaining popularity among turn-of-the-century embryologists. Around the same time, Wilhelm Johannsen coined the term “gene” to refer to the other half of Mendel’s original factor concept, the heritable trait of the adult organism (figure 4).

This finer-grained metaphorical system was useful for posing questions about the physical basis of heredity and its relation to development, but there remained one major inadequacy. Nothing in the metaphor accounted for the way in which the hereditary material *guides* development so as to reliably achieve a particular adult form. In 1930 it was possible (just barely) to envision a machine that could add parts to itself, but it beggared imagination to ask how such a machine – a physical system without intrinsic purpose, without teleology – could accurately reconstruct its own missing parts, or produce a tiny new machine that would grow into a duplicate of itself.

What was missing was the idea of *encoding*, the notion that form can be stored in matter without being literally expressed. The only familiar concept that had something like this structure was writing, and so early thought on this subject was dominated by the metaphors of written

language and the “book of life” (Kay 2000). With the discovery of DNA and the elucidation of the transcription process,¹¹ the metaphor shifted a bit to become one of *instructions* or *blueprints* to be read and followed by the organism-as-machine.

Despite its complexity, this extended metaphor – of a physical part containing a plan for the organism, a plan which is read and carried out by other physical parts – is perhaps the most successful scientific metaphor of all time. It has made accessible, even to the lay imagination, a system whose literal description is both extremely daunting and still far from complete. Anyone with a grade-school education can use the metaphor to discuss the phenomena of heredity – fallibly, to be sure, but with a reasonable chance of capturing at least some of the properties and dynamics that are central to genetics. And in fact, the metaphor is now so fundamental to our thought about heredity that it has, in some respects, ceased to *be* a metaphor. Genes aren’t *like* instructions, they *are* instructions; the body isn’t *like* a machine, it *is* a machine (albeit a very complex and squishy one).

It’s important to understand how this state of affairs developed. In its early days, the metaphor had to be a metaphor because the relevant literal concepts weren’t sufficiently developed. Even if one could formulate the abstract idea of a self-assembling machine, there was no model or technical vocabulary for describing how such thing might actually work. One could speak vaguely of organismic development as the following of instructions by a machine, but there was no body of theory explaining how such machines might operate or what the instructions might look like. The conceptual resources for describing such things did eventually appear, but they came half a century *after* the metaphors that first picked out the relevant dynamics and provided structure to scientific communication and investigation.

3. The special promise of neural theories of consciousness

This pattern has direct consequences for the evaluation of theories of consciousness. It seems likely that a mature theory of consciousness will require at least as much new conceptual apparatus as did our mature theory of heredity – and probably much more. If we accept (following Boyd) that metaphorical theories should be evaluated pragmatically, it becomes apparent that one very important pragmatic criterion should be the extent to which a theory is likely to drive the development of new conceptual resources. Theories that assimilate consciousness to another mystery (like the soul) are unlikely to do this; so, too, are theories which begin by declaring that consciousness is irreducible and basic, since this blocks the “decompose-analogize-reconceptualize” strategy which has proven to be such a powerful engine of conceptual development. A promising metaphor will be one that hints at finer-grained structure *within* consciousness, structure that can be explored and developed within the framework provided by the metaphor.

And this gets back, finally, to the “why the brain?” question. The problem with analogies like Turing’s and Mendel’s is that, in decomposing the target system functionally rather than physically, they sacrifice much of the intrinsic open-endedness of metaphor. A part *defined in terms of* its function doesn’t encourage the development of new conceptual resources, because

¹¹ Along with the development of information theory, itself a profoundly metaphorical discipline. The evolution of the concept of “information,” in both its formal and informal senses, has had a great impact on the central metaphors of the life sciences. For a brief discussion of this with respect to consciousness, see Farber (2000); Kay (2000) is an in-depth examination of its role in genetics.

it's already bound by its very definition into the old system of concepts. By contrast, a *physical* part can be picked out long before one has any good idea of what it does, and this provides a powerful stimulus for research.

Chromosomes played just such a role in the study of heredity. In the mid-19th century biologists using the best available microscopes reported that, during cell division, the nucleus could be seen to disappear and a mass of what came to be called “spindles” would appear in its place and then divide into two masses which became the sites of the new nuclei. By the 1880s Walther Flemming was able, by fixing and staining cells at all stages of division, to show that there were persistent nuclear structures (the chromosomes) which were divided and distributed during this process. This strongly suggested that they played a role in heredity, and once Mendel's theory gained currency scientists were quick to note the parallels between the behavior of chromosomes during cell division and the Mendelian assortment of factors. A turning point in Morgan's work was when he concluded, around 1915, that the statistical patterns he was observing in his fruit fly studies could be well-explained by the assumption that factors were laid out linearly along the chromosomes. This assumption made possible the revolutionary technique of linkage analysis, which used phenotypic data to deduce the number and functional organization of the chromosomes. In this way, identifying the movements of chromosomes with the assortment of traits led to important advances in scientific theory at both levels.

It seems likely that neurons will play a similar role in the study of consciousness. We know they're there, we know they're intimately related to mental activity, and we can observe their individual behavior in detail; yet as any honest neuroscientist will admit, we have precious little understanding of their large-scale dynamics, and this ignorance mirrors our ignorance about the fine structure of consciousness. As a result, NCC theorists are often in the position of *simultaneously* proposing novel analyses of consciousness and developing novel accounts of neural structure and function to underlie them.¹² While certainly difficult, this methodology provides a crucial alternative to the common assumption that some sort of revolutionary insight will allow us to map our existing theory of consciousness onto our existing understanding of the brain.

Let me briefly describe what I believe to be the most important current instance of this co-developmental pattern. Since at least the early 1990s, a major theme in consciousness research has been the need to explain the neural mechanism(s) underlying the mental process known as “binding” or perceptual unity. (See for example the articles by Singer, Crick and Koch, and Llinás and Ribary in Koch and Davis 1994). Before much was known about the architecture of the sensory systems, it was regarded as relatively unproblematic that we could perceive whole objects *as* whole objects. There were Kantian questions about the types of structure that we impose on our perceptions, but the very *imposition* of structure wasn't seen as an especially challenging step in perception. As scientists worked out the structure of the brain's sensory systems, though, a puzzle began to emerge: as it turns out, information about different aspects of the perceptual world (e.g. shape, color, movement, even the presence of faces) is extracted in different parts of the brain and *never converges in any single area or small group of areas*;¹³ in

¹² For a review of several such models, including a more technical discussion of the ones mentioned below, see Banks and Farber (2003).

¹³ Note the important difference between converging *in* and converging *on*. While there are areas – in, for example, the thalamus and frontal cortex – which receive convergent projections from many parts of sensory cortex, these areas do not themselves represent the sensory information, but rather make use of it to perform other tasks.

other words, there is no “Cartesian theater,” no central viewing screen in the brain. A question thus arises: how is it that we can put all these properties back together so as to have a unitary experience?

Christoph von der Malsburg dubbed this puzzle “the binding problem,” and it has been a substantial goad to empirical and theoretical progress. Though there is not yet a well-elaborated consensus solution, attention has focused on the notion that percepts may be represented by “clusters” of individual feature representations which are somehow dynamically linked or bound to each other, possibly in virtue of synchronized oscillations in their activity (figure 5). The core idea here is that a *subjectively* integrated percept occurs when populations of neurons in different areas of the brain become *functionally* integrated. The appeal of these models derives in large part from the possibility of showing that activation clusters of this sort would produce dynamics which mirror those of consciousness in important ways. For example, it can be demonstrated computationally that synchronized groups of neurons will (all else equal) have an enhanced impact on associative processes such as memory, emotion and motor response, paralleling the way in which a conscious percept is more likely to drive such associations than is an object which is perceived but not consciously attended to (Edelman and Tononi 2000). Limitations on the number of available frequencies for synchronous oscillation mean that only a small number of objects may be simultaneously represented in this way, though each object may have many features associated with it (Singer 1994); this mirrors the severe limitations on our ability to consciously attend to multiple separate objects without grouping them into larger, aggregate objects.

In some cases, such work can even yield testable predictions about the properties of subjective consciousness. For example, if binding is achieved via synchrony, there is reason to believe that conscious perception should be temporally “quantized” – meaning that what appears to be a continuous smooth perceptual experience is in fact composed of a series of discrete, fixed-duration chunks or “instants.” An interesting version of this hypothesis was proposed by Rodolfo Llinás. As a theoretical implication of his particular model of synchrony-based binding, he deduced that the temporal quantum of consciousness should be approximately 12-15 msec, and subsequent experimental results (Joliot, Ribary and Llinás 1994) were consistent with this hypothesis: when two 1 msec auditory clicks were played in succession, subjects heard them both as long as the clicks were separated by at least 13 msec, but heard only one click (of normal duration) when the inter-click interval was less than 12 msec. This distinction was also visible in the “resetting” once or twice of a trans-cortical 40 Hz oscillation wave that normally resets in response to each consciously perceived click.¹⁴ From this and other evidence relating to the frequency of oscillations originating in the intralaminar nucleus of the thalamus, Llinás and his colleagues concluded that conscious experience is indeed quantized into discrete chunks of approximately 12.5 msec duration. More recently, VanRullen and Koch have suggested that a number of puzzling quantization and periodicity effects in visual psychophysics can be explained in terms of the oscillatory dynamics of the neural mechanisms underlying conscious perception, and have argued for the broader methodological claim that studying “the temporal evolution of perception” could be a good way to gain insight into “the computations leading to awareness”

¹⁴ It should be noted that, as with many other 40 Hz oscillation based models of consciousness, broader empirical validation of this phenomenon has been elusive. At this point, all one can say with confidence is that (approximately) 40 Hz oscillations are present at least some of the time in some areas of cortex in awake mammals; the source, function and significance for consciousness of these oscillations all remain controversial.

(VanRullen and Koch 2003). For present purposes, the important thing to note in these studies is the role of analogical modeling: the isomorphisms drawn by these researchers between perceptual and neural phenomena are what make it possible for the neural models to both predict and explain the results of psychophysical studies.

None of this is meant to suggest that consciousness studies should become a mere subdiscipline of neuroscience. The various sciences of mind and brain, while intimately related, approach their targets from different perspectives, using different tools and asking different questions. Integration is an important goal, but it cannot be rushed. At present there is much about consciousness that cannot even be described, let alone investigated, without recourse to the vocabularies of the social sciences and humanities. It would, however, be a mistake to assume that this implies any hard limits on the explanatory range or scope of NCC theories. Divisions and gaps in the structure of our engagement with the world (that is, in our scientific theories and in the domains of inquiry that shape them) should not be mistaken for unbridgeable fissures in the structure of reality itself (Farber 2000).

Only once this methodological point is recognized can the true importance of the analogical approach be understood. A longstanding problem for materialist philosophers of mind has been the lack of a plausible, concrete example of what a successful materialist theory of consciousness should *look like*; given the celebrated gap between the subjective and the objective, many have been tempted to declare either that some sort of revolutionary insight is required, or that we will never be able to find relations deeper than mere psychophysical correlation. The above history suggests an alternative prediction: that the gap will be bridged by the gradual development of new conceptual resources, and more specifically of theories which enable us to describe the phenomena of consciousness in a metaphysically neutral way – that is, theories which enable us to say useful things about consciousness without first specifying whether we are speaking of its phenomenal or physical aspect. As with the computational approach to thought and the informational approach to genetics, it is reasonable to believe that this sort of development will be driven by theories which establish analogical relations between the structure of consciousness and the structure of the physical system in which it inheres: in other words, by NCC theories.

4. Some specific consequences for the evaluation of NCC theories

Different scientific disciplines evaluate and test theories in very different ways, and one of the most vexing challenges in a new field can be that of figuring out just which methods and criteria are most appropriate. The bewildering and sometimes comical array of objections faced by NCC theorists when they present their work at conferences attests to the fact that the neuroscience of consciousness – and to some extent, cognitive neuroscience generally – is still in this “growth pains” stage. One of the risks of this situation is that, in relying on the standards of their home disciplines, researchers may fail to recognize or appreciate some of the special demands that accompany the goal of explaining consciousness in neural terms; a second and opposite risk is that certain objections – especially those which can be expressed in a nonspecialist vocabulary and which depend only on widely appreciated features of “the scientific method” – may consume far more time and attention than they deserve. A clearer understanding of the methodological structure of NCC theories can help to focus the debate, reducing both of these risks. In this vein, I will conclude by drawing on the analogical conception presented above

to offer some specific suggestions – two positive and one negative – regarding criteria for the evaluation of NCC theories.

On the positive front, I want to suggest two new criteria that should help to distinguish the merits of competing theories. In addition to all the usual scientific desiderata – fit with data, consilience with other theories, parsimony, tractability, etc. – a few less-standard features assume special importance in light of the analogical structure of the NCC approach. The first is what might be called **intertheoretic bridge potential**: that is, the ability of a model to reduce the “gappiness” of our understanding of consciousness by supporting the development of neutral or intermediate modes of description similar to those seen in genetics and classical cognitive science. This potential will be greatest in NCC theories which identify consciousness with patterns, dynamics and/or functional architectures, since such structures can be identified via different modalities and within different theoretical domains; conversely, it will be weaker in theories which identify consciousness with particular substances or anatomical structures, since these are definitionally locked to a particular discipline and mode of access.¹⁵ All else equal, theories with greater bridge potential are to be preferred because they’re more likely to lead to interesting conceptual advances, as well as to a reduction in the sense that consciousness is somehow metaphysically “special” or distinct from everything else that science investigates.

The second new criterion is **detailed mapping**. It has been widely observed that isomorphism is cheap; in other words, it is often (or in some contexts, provably always) possible to rig up some sort of ad hoc mapping between any two domains. The best way to filter out ad hoc mappings is to test whether proposed models are “projectible,” that is, whether they can be used to make novel empirical predictions that are then confirmed by subsequent observations. A theory’s ability to support such predictions, however, will depend heavily on the level of detail in its proposed mapping, and in my opinion this is the most important area of weakness in the current crop of NCC theories. While there’s clearly something right and relevant in such concepts as “competing activity clusters” and “global accessibility,” as metaphors they remain awfully vague. Recent versions of these proposals are incorporating more specific claims about the anatomical underpinnings, but such details alone cannot validate these models *as models of consciousness* unless they can be mapped in an interesting way onto elements in the phenomenological domain.

The temporal quantization work discussed above is one example of how this can be done; relating neural models to the dynamics of bistable perception (Leopold and Logothetis 1999) is another. The detailed mapping criterion expresses the idea that a good NCC theory is one which gives rise to these sorts of fine-grained, testable isomorphisms. In addition to superior testability, high-resolution mappings are desirable for the simple reason that they have more to *say* about the workings of consciousness. As the field of consciousness studies matures and we look for NCC

¹⁵ This is not to say that there is anything wrong with seeking to identify the particular neurotransmitters, cell types and anatomical structures that underlie the distinctive properties of consciousness; indeed, such details will be a crucial part of any sophisticated NCC theory. The problem is that such identifications can sometimes be derived directly from gross correlational evidence (such as the loss of consciousness due to particular lesions, anesthetics or focal seizures) even in the absence of a detailed model of what the entities in question *do*. Theories that identify consciousness with particular entities in this way can thus purchase empirical support at the cost of conceptual or dynamical emptiness, a move which dramatically reduces their bridge potential. Another way of putting this point is that intertheoretic bridging requires not just a model but a *mechanism*, in the sense of Machamer, Darden and Craver (2000).

theories that are not just plausible but genuinely explanatory, this criterion will thus become increasingly important.

In closing, let me return to the first question I raised above, the one regarding correlation and causation. If the analysis offered herein is even approximately correct, then it follows that one of the most popular objections to NCC theories – the charge that they necessarily express a “mere correlation” rather than a genuine causal explanation – is misguided and should be rejected entirely. Founded as it is on an inaccurate conception of the empirical basis and theoretical structure of the NCC approach, it expresses a worry that is neither relevant nor helpful; furthermore, its imagined goal – a model which explains how neural activity “causes” consciousness – is not something that we *should* be seeking even if we could, since it would build into the foundation of our understanding the very metaphysical gulf that theorists of consciousness are trying to overcome. A satisfying theory of consciousness will be one which allows us to see the properties of consciousness – its chains of association, its waxing and waning, its selectivity and reflexivity, its special relations to language and imagination and sociality, its puzzling ineffability – right there *in* the structures and movements of the brain. Rudimentary theories of this sort are already taking shape; what they most pressingly need, right now, is not some special new bit of metaphysics, but a higher level of precision and detail in the explanatory mappings that they offer.

References

- Banks W. and Farber I. (2003), 'Consciousness', in *Handbook of Psychology, v.4: Experimental Psychology*, ed. A. Healey and R. Proctor (New Jersey: Wiley).
- Black M. (1962), *Models and Metaphors* (Ithaca: Cornell University Press).
- Black M. (1979), 'More about metaphor', in *Metaphor and Thought*, ed. A. Ortony (Cambridge: Cambridge University Press).
- Boyd R. (1979), 'Metaphor and theory change: What is 'metaphor' a metaphor for?', in *Metaphor and Thought*, ed. A. Ortony (Cambridge: Cambridge University Press).
- Chalmers D. (1995), 'Facing up to the problem of consciousness', *Journal of Consciousness Studies* 2 (3), pp. 200-219.
- Crick F.C. and Koch C. (2003), 'A framework for consciousness', *Nature Neuroscience* 6 (2), pp. 119-126.
- Dennett D. (2001), 'Are we explaining consciousness yet?', in *The Cognitive Neuroscience of Consciousness*, ed. S. Dehaene (Cambridge: MIT Press).
- Edelman G. and Tononi G. (2000), *A Universe of Consciousness: How Matter Becomes Imagination* (New York: Basic Books).
- Farber I. and Churchland P.S. (1995), 'Consciousness and the neurosciences: Philosophical and theoretical issues', in *The Cognitive Neurosciences*, ed. M. Gazzaniga (Cambridge: MIT Press).
- Farber I. (2000), *Domain Integration: A Theory of Progress in the Scientific Understanding of Life and Mind* (doctoral dissertation, University of California, San Diego).
- Farber I., Peterman W. and Churchland P.S. (2001), 'The view from here: Spatial representations', in *The Foundations of Cognitive Science*, ed. J. Branquinho (Oxford: Oxford University Press).
- Joliot M., Ribary U. and Llinás R. (1994) 'Human oscillatory brain activity near 40 Hz coexists with cognitive temporal binding', *Proceedings of the National Academy of Sciences USA* 91 (24), pp. 11748-11751.
- Kay L. (2000), *Who Wrote the Book of Life? : A History of the Genetic Code* (Stanford: Stanford University Press).
- Koch C. and Davis J., eds. (1994), *Large-scale Neuronal Theories of the Brain* (Cambridge: MIT Press).
- Koch C. (2004), *The Quest for Consciousness: A Neurobiological Approach* (Englewood: Roberts & Company).
- Lakoff G. and Johnson M. (1980), *Metaphors We Live By* (Chicago: University of Chicago Press).

- Leopold D.A. and Logothetis N.K. (1999), 'Multistable phenomena: Changing views in perception', *Trends in Cognitive Science* **3**, pp. 254-264.
- Llinás R. and Ribary U. (1994) 'Perception as an oneiric-like state modulated by the senses', in *Large-scale Neuronal Theories of the Brain*, ed. C. Koch and J. Davis (Cambridge: MIT Press).
- Logothetis N.K. (1998) 'Single units and conscious vision', *Philosophical Transactions of the Royal Society of London B* **353**, pp. 1801-1818.
- Machamer P., Darden L. and Craver C. (2000), 'Thinking about mechanisms', *Philosophy of Science* **67**, pp. 1-25.
- Searle J. (1992), *The Rediscovery of the Mind* (Cambridge: MIT Press).
- Searle J. (1997), *The Mystery of Consciousness* (New York: New York Review of Books).
- Singer W. (1994) 'Putative functions of temporal correlations in neocortical processing', in *Large-scale Neuronal Theories of the Brain*, ed. C. Koch and J. Davis (Cambridge: MIT Press).
- VanRullen R. and Koch C. (2003), 'Is perception discrete or continuous?', *Trends in Cognitive Science* **7**, pp. 207-213.
- Velmans M. (2002), 'How could conscious experiences affect brains?', *Journal of Consciousness Studies* **9** (11), pp. 3-29.

from Turing

idea	↔	physical symbol
thinking	↔	syntactic manipulation of symbols
thinking person	↔	Turing Machine
algorithm	↔	procedure implementable in a TM
complexity of a problem	↔	worst-case time/space required to solve a problem using the best TM program

from cognitive science

brain	↔	computer hardware
mind	↔	computer software or operations
thought	↔	information processing
memory	↔	information storage and retrieval
volition	↔	executive control
consciousness	↔	feedback or monitoring

figure 1: Mind-computer analogies. Various of these elements went to make up different models, though no one model contained them all. For a contemporary list of computer metaphors that were common in the heyday of classical cognitive science, see Boyd (1979).

Brains

Components are erratic and fault-tolerant
Architecture is modified by activity
Massively parallel processing
No central timekeeper for computation
Processing mechanisms are content-specific
Evolved primarily for sensorimotor integration

Digital computers

Components are consistent and fault-sensitive
Architecture is fixed
Serial or weakly parallel processing
Central timekeeper is essential
Processing mechanisms are general
Designed primarily for symbol processing

figure 2: A few functionally significant brain-computer disanalogies.

organism	↔	container
“factor” (gene + trait)	↔	contents
inheritance	↔	random pairwise mixing of contents

figure 3: Mendel’s analogy for heredity (1860s, rediscovered in 1900).

Wilhelm Roux (1890s)

organism	↔	machine
embryonic development	↔	mechanical movement
hereditary material	↔	mechanical part

Thomas Hunt Morgan (1910s)

“factor” (hereditary material)	↔	mechanical part that guides assembly
inheritance	↔	division and duplication of above parts

Wilhelm Johannsen (1910s)

“gene” (heritable trait)	↔	characteristic of complete machine
--------------------------	---	------------------------------------

Watson and Crick (1950s)

DNA	↔	plan or blueprint
development	↔	decoding and following of plans

figure 4: Further development of analogies for heredity.

quale	↔	local feature representation
image	↔	integrated cluster of feature representations
“core” or “general” consciousness	↔	approximately 40Hz oscillation sweeping across cortex (“global wave”), driven by intralaminar nuclei of the thalamus
specific contents of consciousness	↔	integrated cluster firing in synchrony with the global wave
train of consciousness	↔	sequence of representations produced by competition among activity clusters
consciousness vs. unconsciousness	↔	presence or absence of the global wave
waking vs. dreaming consciousness	↔	ability or inability of sensory input to reset the global wave
temporal quantum of consciousness	↔	duration of one sweep of the global wave
memory, motor choice	↔	associations driven by the greater associative power of representations oscillating in synchrony with each other and with the global wave

figure 5: An example of a detailed analogical mapping between neural dynamics and consciousness. It should be stressed that this is just one possible model, based primarily on the work of Rodolfo Llinás (see text for references) and incorporating ideas shared by other major theories (e.g. those of Singer, Crick and Koch, Edelman and Tononi, and Damasio).