PCID 2.3.1, October 2003

Groundwork for an Emergentist Account of the Mental

Timothy O'Connor

I. Conscious Experience and Action

Every moment of our waking lives, we confront an aspect of the natural world that seems set apart from all the rest. It stares us right in the face, so to speak: conscious experience (of varied forms) and thought, in ourselves and presumably in a great many other forms of biological life. To focus on just the most discussed of conscious phenomena, consider the *qualitative* content of our experiences—'what it is like' to have a certain thought or sensation. (The philosopher Ned Block proposes the term 'phenomenal consciousness,' to distinguish this feature from others that go under the label of 'consciousness' in ordinary as well as in specialized discourse.) There are two striking aspects of the qualities of conscious experience. The first is the apparent simplicity or nonstructurality of the most basic elements of experience, such as a homogenous patch of blue in one's visual experience. Thus, while we can recognize such constituent features within a single unified field of awareness, it doesn't seem to make sense to suppose that any of these basic constituent features themselves have *underlying* structure—that they are built up out of properties of parts (e.g., assemblies of neurons) that are themselves not directly apprehendable. Our awareness of our own conscious state is not mediated by causal signals, so that the awareness might give a merely partial—or even, in principle, completely erroneous—representation of the conscious state. (Compare the compatibility of unobservable, underlying structure in this page, despite your direct perceptual awareness of it, a compatibility that depends on the fact that your awareness partly

consists in information being transmitted to you and leading to an inner representation which captures just some features of the page, and those in a simplified fashion.) The second striking feature of consciousness is its apparently *sui generis* nature vis-a-vis the physical, owing to the feature of *subjectivity*. It seems that one can come in direct contact with a conscious property only by having it. It is precisely this in-principle asymmetry of access that leads one to describe consciousness as inherently subjective in nature—a real feature of the world, while not 'objective' in the sense of being intersubjectively accessible.

A no less salient, though less pervasive and conceptually more problematic, feature of our mental lives is our *willing* to do this or that, directing our own actions by selecting from among a plurality of options. The spontaneous, self-determined character of many of our actions seems different in kind from the causal activity of mere (impersonal) mechanisms. We might express this by saying that our freely-chosen actions are both undetermined and triadic in structure—they involve our causing an intention *for a reason*, and this triadic structure is not conceptually or ontologically decomposable into a process consisting of dyadic causal interactions among states of the agent.

As striking as conscious experience, thought, and deliberate action are, their irreducibility to physical processes within their subjects is hotly debated.[1] I shall ignore these debates entirely, as my purpose in this essay is constructive. Assuming that these mental qualities and processes are indeed irreducible to impersonal, non-purposive physical phenomena, I want to propose the very general form a non-reductive explanatory account of their underpinnings and dynamics should take. A suggestive label for my proposal is *ontological emergence*.

II. Emergence: What Is It?

As soon as one proposes a view under the banner of 'emergence', one has to explain what one does *not* mean, as the term has been used to cover a multitude of sympathies. All 'emergentist' views intend to position themselves between two polar extremes, which we might term hyper-dualism and hyper-reductionism.

The hyper-dualist position in the philosophy of mind is the substance dualism famously championed in the 17th century by Descartes. On Descartes' view, our mind or soul (the only essential part of ourselves) has no spatial location. Yet it directly interacts with but one physical object, the brain of that body with which it is, 'as it were, intermingled,' so as to 'form one unit.'[2] The radical disparity posited between a nonspatial mind, whose intentional and conscious properties are had by no physical object, and a spatial body, all of whose properties are had by no mind, has prompted some to conclude that, pace Descartes, causal interaction between the two is impossible. There is a real conceptual challenge here, which I've discussed elsewhere.[3] Here I'll simply note one familiar empirical problem for substance dualism that seems to me to be weighty. It must suppose that a composite physical system gives rise, all in one go, to a whole, self-contained, organized system of properties bound up with a distinct individual. For we cannot say, as we should want to do, that as the underlying physical structure *develops,* the emergent self does likewise, as there doesn't seem to be conceptual space for changing mereological complexity within a nonphysical simple. No, the substance dualist view will have to say, instead, that at an early stage of physical development, a self, or primitive mental substance, emerges with all the capacities of an adult human self, most

of which lie dormant owing to immaturity in the physical system from which it emerges. And this is surely implausible.[4]

If not substance dualism, then what? The irreducibility of conscious experience and self-determining action we affirmed at the outset already commits one to a kind of dualism, a duality of physical and conscious properties. The challenge for an ostensibly mediating emergentism is to locate property dualism within a more integrationist picture than that of Descartes. Let us be clear, however, that the dualism we must accommodate is ontological. It is not simply that the categories and theories we employ in describing mental phenomena cannot be rephrased without loss in the terms of our fundamental physical theories. Rather, mental qualities and processes themselves, quite apart from our ways of identifying them, possess a character distinct in kind from physical qualities and processes, of whatever degree of complexity we please. Consequently, the needed account of emergence is quite different from other, epistemologically-rooted notions of emergence to be found in some contemporary theories of mind in philosophy and cognitive science. There is, indeed, a pronounced tendency among recent philosophical commentators to conflate or blur ontological and epistemological issues when applying 'emergentist' ideas to nonlinear phenomena, artificial life, and human mentality. As a symptom of this, discussions of senses of 'unpredictability' dominate these other expositions. While seeing how an ontologically emergent property would be unpredictable from a certain, limited empirical standpoint *is* a useful way of getting a fix on the concept, this should be but a consequence of its core metaphysical tenets.

On the view I advocate, there are three such tenets. First, as already indicated, an emergent property is a basic feature of the world, as much as unit negative charge may

turn out to be, although in the case of the emergent property, it is had by a complex system, rather than a partless individual such as an electron. An emergent property is ontologically basic, nonstructural, or simple in that any instance of such a property in a complex system does not even partly consist in, nor is it 'constituted' or 'realized' by, the having of distinct properties by the system's parts and their standing in various relations. (Emergence is thus to be distinguished from versions of 'nonreductive physicalism' which hold that mental features are in some obscure sense 'realized' in while supervening on physical features of the world.)

A second feature of emergence concerns its causal influence. Quite apart from the special dynamics of freedom of choice, conscious mental states affect our behavior in all sorts of ways. Given the structural simplicity of an emergent property, this implies that an emergent property will *fundamentally* alter the behavior of that system. In contrast to the operation of an ordinary structural macro-property, such as the mass of your body, whose causal influence occurs via the activity of the micro-properties which constitute it, a structurally simple property will bears its influence in a direct, 'downward' fashion on the object's microstructure.This has been dubbed 'downward causation' to contrast it with the 'bottoms-up' determination characteristic of non-emergent systems.

An emergent property, then, is a non-structural or basic property which is exemplified by objects or systems of some requisite level and kind of organizational complexity and which exerts downward causation. A final question concerns the relationship of the emergent property to the lower-level properties of the object. What explains the emergent's presence in complex systems of a particular sort? Must we say, to use Samuel Alexander's memorable phrase, that it is "to be accepted with the natural

piety of the investigator," admitting no explanation?[5] This would be to abandon a very basic assumption of scientific inquiry—that all of nature constitutes a single, causally-connected system—and with it any possibility of a scientific understanding of emergent phenomena. But we needn't suppose this. Instead (and here is our third emergentist tenet), the occurrence of an emergent property will be a product of certain joint causal potentialities of the underlying properties of the system's parts. We generally think of microphysical properties such as charge and spin as manifesting their dispositions 'locally', regardless of the wider context in which they are embedded. However, we may consistently conceive that some such microphysical qualities are *also* disposed to generate large-scale effects (our emergent properties, basic features of complex systems) in tandem with other such properties when these are jointly embedded in a properly organized context. From a certain empirical standpoint, the complexity threshold required for manifesting this propensity will be arbitrary, as this disposition towards a joint effect would not be empirically discernible in systems below the threshold, among which are just those fairly simple systems that particle physicists study in formulating models of microscopic interactions. Finally, note that over time, emergent features will undoubtedly influence the future distribution not only of microphysical qualities of the system's elements, but also of other emergent qualities of the system as a whole. (At any rate, this will be a plausible assumption for an emergentist account of conscious states.)

Our schematic picture of emergence, then, is that organized systems of particles causally generate and sustain ontologically basic features of the system, which in turn exert a basic influence on the future evolution of the system in both microphysical and emergent-level terms. Note that on this picture, everything that occurs, including

emergent features and their consequences, rests on the total range of causal dispositions of the fundamental physical properties. For the occurrence of any emergent properties are among those dispositions, and so the effects of the emergent features are indirectly a consequence of the physical properties, too. Notwithstanding this fact, emergent features *are* fundamental and *do* make a basic causal difference to how the world unfolds. The causal difference that emergence makes is that what happens transcends the immediate, or local, interactions of the microphysics. What is more, in a world whose basic physics is indeterministic, the distribution of emergent qualities will fail to supervene on the global (and diachronic) distribution of fundamental physical qualities.

III. Emergence: Is It Plausible?

Many contemporary thinkers will dismiss my explication of emergence as a metaphysical castle in the air. Human beings are a part of nature, so any plausible conjecture about our mental lives must conform to 'the emerging scientific picture of the world.' I confess that I am puzzled by the claim that emergence *is* implausible on the current state of scientific knowledge.

Let us make a simple distinction that will prove useful. One might maintain that all *macro*-level phenomena (consciousness included) have arisen through and continually causally depend on *micro*physical causal processes. Call this the 'Causal Unity of Nature Thesis.' There is good reason to affirm this thesis. A second and much stronger thesis is that every (token) macro-level phenomenon is *constituted* by a nested structure of microphysical processes. Observable macro-level phenomena are, in all cases, nothing over and above a whole bunch of microphysical goings-on. (In our earlier terms, all

macro-level properties are *structural* properties.) We may call this the 'Micro-Macro Constitution Thesis.'

And now I ask my opponent, what compelling reason is there to affirm not merely the Causal Unity Thesis, but also the Constitution Thesis? Broadly speaking, there are two ways the latter could be established. I believe that some (mostly philosophers) have supposed that it *follows* from the Causal Unity Thesis. (This is suggested by the tendency not to carefully distinguish the two.) The model of emergence sketched above makes plain that there is no such entailment. The second way of arguing for the Constitution Thesis is to argue that it has been directly established or made highly probable by work in relevant sciences (principally, various branches of biology and psychology).

Is the thesis that human mental properties and their activity emerge from and causally affect the nervous systems of human beings wholly implausible on current scientific knowledge? To argue that this is so requires evidence drawn from the various studies of *complex* physical systems, particularly those falling under the purview of the biological sciences. (It manifestly does *not* suffice to point to the fact that fundamental physics makes no reference to the direct influence of macroscopic properties.[6] The evidence for fundamental theories is gathered by analyzing relatively simple, decomposed physical systems. Moreover, as Nancy Cartwright has emphasized,[7] the technical devices of physics used in these contexts are built to ensure that no interference from factors outside the domain of the theory occurs. If some scientists say that they believe these results to hold quite generally, regardless of the macro-level complexity in which a microphysical system is imbedded, and despite the fact that there isn't even at present a worked-out procedure for applying quantum mechanics to many particle

systems on anything remotely like the order of biological systems, then they are merely expressing their faith in a reductionist metaphysics.) While there are no widely-accepted working theories that are expressly *committed* to the existence of emergent properties (apart, possibly, from theories governing small quantum-mechanical systems) contemporary scientific knowledge is sufficiently incomplete as not to rule out an emergentist picture of some factors within some highly organized phenomena. Indeed, it is precisely the incompleteness of scientific theory that leaves open the question of the status of higher-level properties. How could one suppose the matter to be settled (in favor of reductionism) apart from the end game where completed theories are compared?

Consider also that claims *are* occasionally made that certain highly structured phenomena unconnected to animal consciousness are plausible candidates for instances of emergence. What I've just said, of course, implies that these claims are necessarily speculative. One example some cite is the 'dissipative structures' in nonequilibrium thermodynamics studied by Nobel laureate Ilya Prigogine.[8] Perhaps another is implicit in appeals to the role of 'hierarchical control' in the philosophy of biology.[9] Whatever the merits of these conjectures may be, they support my outsider's contention that the body of firmly-established fact in the biological disciplines can countenance property emergence. This is especially true with regard to the complex workings of the human neurophysiological system. Given that much remains to be understood about the detailed interactions of parallel and hierarchically-ordered subsystems of this system, how could one confidently assert that a completely general "bottom up" picture of this system is empirically established?

The strategy of micro-reduction has had enormous success in modern theoretical science, at least in application to restricted features of higher-level phenomena. But this does not strongly support the general negative thesis that there are no macro-determinative emergent properties in nature. (I have the impression in discussing this matter with other philosophers that many think that if there were emergent properties, they would be ubiquitous in nature, appearing at many or all importantly unified levels of natural organization. This strikes me as a groundless judgment.) More generally, the commitment to micro-reducibility, particularly among philosophers, is partly an overreaction to the now falsified crude conception of levels of nature propounded by early emergentists, one based in a much simpler picture of higher-level phenomena than we have at present. And in the end, it must be stressed that, whatever one thinks of purely *theoretical* inferences to emergence from the failure of existing lower-level theories to accommodate some phenomena, with the case of consciousness in all its facets, it certainly appears that we have powerful *evidence* for irreducible high-level features.

IV. Conceptual Issues for Emergentist Theories

While there is currently no sound basis, as far as I can see, for an anti-emergentist stance concerning conscious phenomena, there are plenty of general practical difficulties facing any attempt to develop concrete emergence-based theories (of, say, the specific dynamics of conscious experience, thought, and action in human beings). One will require a way to precisely identify, categorize, and measure the elementary qualities of such phenomena. One will then have to identify general dynamical principles characterizing the impact of such qualities on the microphysical processes which underpin them.

Taking the phenomena of *freedom of choice* seriously brings yet further difficulties. It appears to involve the idea of a fundamentally personal causal capacity to initiate behavior for the sake of reasons. If there is such a capacity in normally-functioning, mature human beings, one should be able to determine the precise underlying properties on which this distinctive capacity depends. Conversely, what structural transformations in the human nervous system would result in long-standing (or permanent) loss of this capacity? Another theoretical question results from the fact that, if there are distinctive events of willings, or the personal initiation of causal chains constituting actions, there is no neat and simple way of dividing those events from 'garden variety' mental and neurophysiological ones. It surely must be allowed that some human behavior, even consciously-governed behavior, is entirely brought about in quasi-deterministic fashion by proximate psychological factors. (Not all action is *free* action.) This is pretty obviously true, for example, of some behavior powerfully influenced by unconscious factors and of highly routinized actions. Precisely to what extent, then, is an ordinary human's behavior directly regulated by the agent himself, and to what extent is it controlled by unwilled physical-cum-psychological processes? Furthermore, even when I ostensibly act freely, I usually am not even trying to control directly the precise degree of muscle contraction, limb trajectory, and so forth. This makes it plausible to hold that our memory system stores action sequences that we simply activate through conscious choice.[10] It may be that these choices at times simply unfold, while we simply monitor the result and retain the capacity to redirect things as need be. Hence, a theorist would want to be able to identify factors whose presence or absence determines which scenario obtains on a given occasion.

Finally, one might also worry that free will requires the emergence of a degree of indeterminism far beyond what we have any reason to believe is operative (as a function of quantum indeterminacy) at the complex level of neural structures. My reply is that since an emergent property has, relative to the properties which underly it, a unique, nonstructural nature, we have no *a priori* reason to think it must result in processes exhibiting precisely the same degree of indeterminism as is present in its sustaining lower-level processes. Still, we are not supposing 'something's coming from nothing,' as many have thought: the presence of any emergent, on the view I have sketched, will be caused by the object's more fundamental features. What the view does allow is that a stable set of processes may give rise, at certain critical junctures, to a somewhat different order of affairs via 'top-down' controlling features. It is just this possibility that allows the right sort of emergentist view to overcome the *opposite* complaint from Cartesian sympathizers that agents with such emergent capacities are 'ontologically superficial'—not among the truly basic entities whose activities determine the way the world is. While it is true, on my picture, that the capacity for freely willing outcomes in select complex entities has always been among the potentialities of the world's primordial building blocks, the way those potentialities are exercised is not so prefigured. The agents themselves determine these outcomes. In consequence, any way of completely characterizing what happens in the world must make reference to these agents and their distinctive capacities. This is as ontologically 'deep' as any entity that is not a primordial fount of being could aspire to.

Indiana University

* Extensive portions of this article were previously published in Ch.6 of my *Persons and Causes: The Metaphysics of Free Will* (Oxford University Press, 2000) and "Causality, Mind, and Free Will," *Philosophical Perspectives 14: Action and Freedom*, ed. J. Tomberlin, 2000, 105-117.

---

[1] The literature on the irreducibility of consciousness is massive. An influential case in favor of irreducibility is Frank Jackson, "Epiphenomenal Qualia," Philosophical Quarterly, 32 (1982), pp.127-36, and one form of response is given in David Lewis "What Experience Teaches," in William Lycan, ed., Mind and Cognition (Blackwell, 1990), pp.499-519. More recent discussions include David Chalmers, The Conscious Mind (Oxford University Press, 1996), and Michael Tye, Ten Problems of Consciousness (MIT Press, 1995).

On the irreducibility of conscious willing, see my Persons and Causes. The psychologist Daniel Wegner gives a recent sustained empirical argument against this in *The Illusion of Conscious Will* (MIT, 2002). I have given talks critiquing Wegner's case, and hope to publish a response soon.

[2] See Descartes' *Meditations on First Philosophy*, Meditation VI.

[3] "Causality, Mind, and Free Will," *op cit.*

[4] The developmental problem for dualism is presented in Armstrong 1968. Hasker 1999 defends an emergentist form of substance dualism.

[5] Space, Time, and Deity Vol.I-II (New York: The Humanities Press, 1920), 46-7.

[6] Pace David Lewis in "What Experience Teaches," in William G. Lycan, ed., Mind and Cognition (Cambridge, MA: Blackwell, 1990), p.513.

[7] "Fundamentalism vs. The Patchwork of Laws," Proceedings of the Aristotelian Society N.S. 94 (1994), pp.279-92.

[8] See, e.g., G. Nicolis and I. Prigogine, Self-organization in Nonequilibrium Systems : From Dissipative Structures to Order Through Fluctuations (New York: Wiley, 1977).

[9] See Uko Zylstra "Living Things as Hierarchically Organized Structures," Synthese 91 (1992), pp.111-133 and Marjorie Grene, "Hierarchies in Biology," American Scientist 75

(1987), pp.504-510, for attempts to distinguish biological systems that exhibit hierarchical control from other sorts of hierarchically-arranged systems.

[10] An early philosophical discussion of the implications of this for theories of action is A. Farrer's The Freedom of the Will (London: Adam & Charles Black, 1958). Concrete proposals for how to account for such phenomena within recent cognitive science models may be found in S. Kosslyn and O. Koenig in Wet Mind (New York: Free Press, 1992) and Donald Norman and Tim Shallice, "Attention to Action: Willed and Automatic Control of Behavior," in Michael Gazzaniga, ed., Cognitive Neuroscience: A Reader (Oxford: Blackwell, 2000), pp. 376-390.