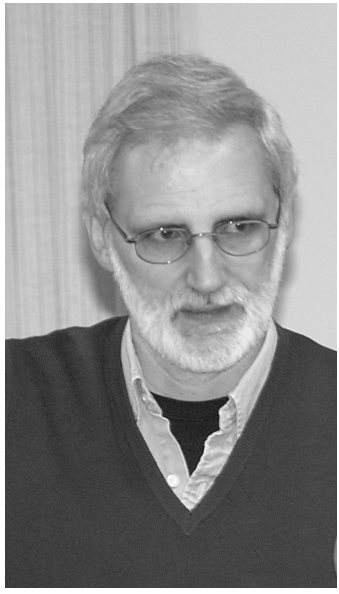

William Seager

The Emergence of Consciousness

William Seager

I. Varieties of Consciousness

When we speak of consciousness, or being conscious or modes of consciousness, there is a huge multitude of phenomena to which we could be referring. It is not completely obvious that there is any one thing that stands as the core referent of all our different ways of talking about consciousness (see Wilkes 1988). But it is important to distinguish the field of consciousness from mentality in general. Ever since Freud, and Helmholtz somewhat before him, the concept of unconscious mental states has seemed quite unproblematic, despite the previous widespread opinion that the idea was completely incoherent. Freud's innovation depends upon noticing that some mental states have a distinctive role in the etiology of behavior that is seemingly independent of consciousness. Roughly speaking, this role is to cause behavior so that it is susceptible to "rationalizing" explanation.

By this term I do not mean that the explained behavior is in itself rational but that it is rationally understandable in relation to some set of beliefs and desires attributed to the agent. Freud, for example, explains why a woman misspoke by means of a little story the upshot of which was that the woman was "under the influence of unconscious thoughts concerning pregnancy and the prevention of conception" (1938, 74). Or again, a woman dreams that her plans for a dinner party are wrecked when she realizes (in the dream) that there is no food to serve the guests. Freud "discovers" that the dream reflects the woman's jealousy of a friend who she knows her husband greatly admires but who – luckily – her husband regards as too thin. Obviously, she does not want her friend to gain weight and hence dreams that she will not be feeding her (1938, 226). None of this is very rational (nor very plausible either of course) but the link between the postulated unconscious beliefs and desires makes sense to us.

Something of the same underlying rationalizing structure can be seen in the inferential theory of perception of Hermann Helmholtz in which various aspects of visual perception, including various forms of illusory perception, are explained in terms of there being an underlying set of "premises" or assumptions about the usual conditions of perception from which, along with current sensory input, our perceptual systems "deduce" what we are seeing. An example might be the use of Emmert's Law to explain the moon illusion (which is the way the moon looks bigger when it is near the horizon than when it is overhead). Emmert's law says that the perceived size of an object varies with the perceived

distance. If we suppose that the moon looks further away when near the horizon (perhaps because it is clearly behind everything on the horizon) than when it is overhead (and cannot be compared to anything else), then Emmert's law predicts that the moon will be perceived as larger when on the horizon. But we are certainly not conscious of any such cognitive mechanism within us so presumably our perceptual system is unconsciously processing information in a way that is rational[1].

This kind of rational structure of belief, desire and behavior evidently does not require any consciousness of the mental states involved. It is now commonplace and legitimate for us to invoke such unconscious states, both in everyday life and throughout research in the "mental sciences". This naturally puts some pressure on the idea that consciousness has an essential role in the generation of behavior, which in turn has led some to raise the specter of "consciousness inessentialism": the view that any behavior which is mediated by consciousness could occur in the absence of consciousness (see Flanagan 1992, ch. 7). In general, the information processing which underlies the system of belief and desire-mediated generation of behavior appears capable of operation in the absence of consciousness. Already artifacts such as chess playing computers have at least simulations of beliefs (about the current position of the pieces and their relative value) and desires (as encoded in the global goal of checkmate and a host of local goals involved in the incremental improvement of position as the game proceeds). The "actions" of chess computers are highly susceptible to rationalizing explanation within the confines of the chess world. It is natural and useful to describe them in ways such as: the computer took the pawn in order to threaten my knight.

But even if we restrict ourselves to mental states that are conscious or involve consciousness, the range of phenomena remains immense. When we think about consciousness we may have in mind highly complex mental activities such as reflective self-consciousness or introspective consciousness, of which perhaps only human beings are capable. Or we may be thinking about something more purely phenomenal, perhaps something as apparently simple and unitary as a momentary stab of pain. Paradigm examples of consciousness are the perceptual states of seeing and hearing, but the nature of the consciousness involved is actually complex and far from clear. Are the conscious elements of perception made up only of "raw" sensations from which we construct objects of perception in a quasi-intellectual operation? Or is perceptual consciousness always of "completed" objects with their worldly properties?

The realm of consciousness is hardly exhausted by its reflective, introspective or perceptual forms. There is distinctively emotional consciousness, which seems to necessarily involve both bodily feelings and some kind of cognitive assessment. Emotional states require a kind of evaluation of a situation. Does consciousness thus include distinctive evaluative states, so that, for example,

consciousness of pain would involve both distinctive bodily sensations and a conscious "sense" of aversion? Linked closely with emotional states are familiar, but nonetheless rather peculiar, states of consciousness that are essentially other-directed, notably empathy and sympathy. We visibly wince when others are hurt and almost seem to feel pain ourselves as we undergo this unique kind of experience.

Philosophers argue about whether all thinking is accompanied by or perhaps even constituted out of sensory materials (images have been the traditional favorite candidate material), and some champion the idea of a pure thought-consciousness independent of sensory components. In any event, there is no doubt that thought is something that often happens consciously and is in some way different from perception, sensation or other forms of consciousness.

Yet another sort of conscious experience is closely associated with the idea of conscious thought but is not identical to it: epistemological consciousness, or the sense of certainty or doubt we have when consciously entertaining a proposition (such as '2 + 3 = 5' or 'the word "eat" consists of three letters'). Descartes famously appealed to such states of consciousness in the "method of doubt" (see his Meditations 1641/1985).

Still another significant if subtle form of consciousness has sometimes been given the name "fringe" consciousness (see Mangan 2001, following James 1890, ch. 9), which refers to the "background" of awareness that sets the context for experience. An example is our sense of orientation or "rightness" in a familiar environment (consider the change in your state of consciousness when you recognize someone's face who at first appeared a stranger). Sometimes the oddity of dreams consists in what ought to be a familiar situation experienced without the normal sense of orientation. Moods present another form of fringe consciousness, with clear links to the more "overtly" conscious emotional states but also clearly distinct from these.

Despite the large diversity in this sample of the forms of consciousness, there is a fundamental common feature that they all share. Consciousness is distinctive for its subjectivity or its "first-person" character. There is "something it is like" to be in a conscious state, and only the conscious subject has direct access to this way of being (see Nagel 1974). There is nothing it is like to be a rock, no subjective aspect to an ashtray. But conscious beings are essentially different in this respect. Every case given above has its own peculiar but accessible subjective form. This is the hard core of what we mean when we think about the nature of consciousness. And as we'll see, it is this common feature of conscious mental states that makes the emergence of consciousness both mysterious and problematic.

II. The Place of Consciousness in Nature

There is no "problem" of consciousness until we come to see it as surprising or unexpected relative to some favored explanatory framework. The most powerful explanatory framework ever devised is that of empirical science. Why is consciousness and its emergence a special problem for science? To see this, consider what I think can be fairly characterized as the orthodox view of the structure and development of the universe, beginning with the rather grand scheme provided by the science of cosmology. I'll call this general picture the mainstream view and it goes something like the following.

The world is a purely physical system created some 13 billion years ago in the prodigious event Fred Hoyle derisively labeled the big bang. Very shortly after the big bang the world was in a primitive, ultra-hot and chaotic state in which normal matter could not exist, but as the system cooled the familiar elements of hydrogen and helium, as well as some traces of a few heavier elements, began to form (this opening act takes up some 300,000 years). Then very interesting things started to happen, as stars and galaxies quickly evolved, burned through their hydrogen fuel and went nova, in the process creating and spewing forth most of the elements of the periodic table into the increasingly rich galactic environments.

There was not the slightest trace of life, mind or consciousness throughout any of this. That was to come later. The mainstream view continues with the creation of planetary systems. At first these were poor in heavier elements, but after just a few generations of star creation and destruction there were many earth-like planets scattered through the vast – perhaps infinite – expanse of galaxies, and indeed some 7 or 8 billion years after the big bang the earth itself formed along with our solar system.

We do not yet understand it very well, but whether in a warm little pond, around a deeply submerged hydrothermal vent, amongst the complex interstices of some claylike matrix, as a pre-packaged gift from another world, or in some other way of which we have no inkling, conditions on the early earth somehow enabled the rather special – though entirely in accord with physical law— chemistry necessary for the beginnings of life.

But even with the presence of life or proto-life, consciousness still did not grace the earth. The long, slow processes of evolution by natural selection took hold and ultimately led at some time, somewhere to the first living beings that could feel – could feel pain and pleasure, want and fear, could experience sensations of light, sound or odors. The mainstream view sees this radical development as being conditioned by the evolution of neurological behavior control systems in co-evolutionary development with more capable sensory systems. Consciousness thus emerged as a product of increasing biological complexity, from non-conscious precursors composed of non-conscious components.

On this view it seems obvious that there would be a vanishingly small difference between the last, as it were, non-conscious organism and the first conscious one. So one problem with the emergence of consciousness is to identify the physical change that provides the basis for this momentous transition. A fanciful way to put the issue is to imagine that we were alien exo-biologists observing the earth around the time of the emergence of consciousness. How would we know that certain organisms were, while other organisms were not, conscious? What is it about the conscious organisms that explains why they are conscious? Furthermore, the appearance of conscious beings looks to be a development that sharply distinguishes them from their precursors, but the material processes of evolution are not marked by such radical discontinuities. To be sure, we do find striking differences amongst extant organisms. The unique human use of language is perhaps the best example of such a difference, but of course the apes exhibit a host of related, potentially precursor abilities, as do human beings who lack full language use. Thus we have possible models of at least some aspects of our pre-linguistic ancestors that suggest the evolutionary path that led to language.

But the slightest, most fleeting, spark of feeling is a full-fledged instance of consciousness that entirely differentiates its possessor from the realm of the non-conscious. Note here a striking dissimilarity to other biological features. Some creatures have wings and others do not, and we would expect that in the evolution from wingless to winged there would be a hazy region where it just would not be clear whether or not a certain creature's appendages would count as wings. Similarly, as we consider the evolutionary advance from non-conscious to conscious creatures, there would be a range for which we would be unclear about whether or not creatures within that range were conscious. But in this latter case, there is a fact whether or not the creatures in that range are feeling anything, however"dimly" or "weakly," whereas we do not think there must be a fact about whether a certain appendage is or is not a wing (a dim or faint feeling is 100% a kind of consciousness but a few feathers on a forelimb is not a kind of wing). It is up to us whether to count a certain sort of appendage as a wing or not – it makes no difference, so to speak, to the organism what we call it. But it is not up to us to decide whether or not organism X does or does not enjoy some smidgen of consciousness – it either does or it does not.

The abstract and general, perhaps metaphysical, question lurking in the background here links to the issue noted above of conscious inessentialism. Given that creatures capable of fairly complex behavior were evolving without consciousness, why is consciousness necessary for the continued evolution of more complex behavior? Just as wings are an excellent solution to the problem of evolving flight, brains (or more generally nervous systems) are wonderful at implementing richly capable sensory systems and co-ordinated behavior control systems. But why should these brains be conscious? Although perhaps of doubtful

coherence, it's useful to try to imagine our alien biologists as non-conscious beings. Perhaps they are advanced machines well programmed in deduction, induction and abduction. Now, why would they ever posit consciousness in addition to, or as a feature of, the increasingly complex sensory and behavioral control systems they find in their examination of Earth? Thomas Huxley once said: "how it is that anything so remarkable as a state of consciousness comes about as a result of irritating nervous tissue, is just as unaccountable as the appearance of Djin when Aladdin rubbed his lamp" (1866, 8, 210). We might, rather fancifully, describe this core philosophical question about consciousness as that of how the genie of consciousness gets into the lamp of the brain, or why, to use Thomas Nagel's (1974) famous phrase, there ever is in the world "something it is like" to be some entity?

The mainstream view implicitly appeals to a view of the world that is hierarchical or structured in layers of complexity but in which the complex is a determinate function of the simple. We can try to articulate this idea in terms of three interlocking principles that I call completeness, closure and resolution. Completeness is the doctrine that everything in the world is physical and as such abides by closure and resolution. Closure entails that there are no "outside forces" – everything that happens, happens in accordance with fundamental physical laws so as to comply with resolution. Resolution requires that every process or object be resolvable into elementary features that are, by completeness, physical and whose abidance with laws governing these features leads to closure.

Take anything you like: a galaxy, a person, a flounder, an atom, an economy – it seems that anything can be resolved into the fundamental physical constituents, processes and events which determine its activity. Indeed, our best theory of the creation of the universe maintains that at very early times after the big bang the universe was quite literally resolved into its elementary constituents; at that time the universe consisted of an extremely hot, highly active "sea" of quarks, leptons and elementary force exchange bosons. It is important, however, to distinguish the mainstream view – and completeness, closure and resolution – from simple mechanism or part-whole reductionism. We are now told that the fundamental physical features of the world involve non-mechanistic processes and that the properties of quantum mechanical systems are not straightforward functions of the local properties of their parts. But such bizarre properties as quantum entanglement were predicted by quantum mechanics based upon purely physical considerations and are themselves elements of the fundamental physics of the world.

Completeness, closure and resolution and their inter-relations are concisely expressed in the thought that the universe is "running" entirely and solely upon the interactions of these elementary constituents no less today than when it was 10^{-37} seconds old. It's worth pausing to think about this. Of course the

universe is very different today than it was the instant after creation, but the mainstream view is committed to this claim. There just aren't any other forces and their aren't any other constituents.

The crucial difference between that early time and now is the absence of high level structure. From an explanatory point of view, nothing would be gained by trying to impose any high level theory onto the initial chaos (this is not quite true, a certain very abstract high level theory already applied: thermodynamics). That is why, in fact, we can have a pretty fair characterization of the universe very shortly after its creation despite it being the case that the "number" of interactions per femtosecond, so to speak, was no less then than it is now and probably very much greater. It seems that the only change from then to the present is the creation of structure which admits of high level explanatory description, but it is very doubtful that, at bottom, the world is any less "running on physics" now than it was then.

This seems to me the heart of the issue of emergence. Do high level structures contribute to the "go" of the universe (to use Morgan's (1923) evocative term), or are they merely explanatory aids imposed onto the world by us? Of course, there is no doubt that high level structure exists; it is, for example, a plain fact that there are trees in the world. And, seeing they predate us by millions of years, if there were no people about there still would be trees. High level structure is obviously an important feature of our universe. It is perhaps an amazing fact that the basic features of our world permit the emergence of high level structure. Perhaps if we conceive vast multitudes of universes each of which obey our most basic physical laws, differing only in ways permitted by those laws, only an infinitesimal fraction would allow for high level structure. That would be an interesting and significant fact. But it would not by itself suggest that high level structure adds to the causal powers of the world. What it seems to add is the possibility of new explanatory approaches to the world which exploit the existence of high level structure. In fact, it seems that the possibility of imposing such explanatory order on the world is definitive of high level structure.

III. Forms of Emergence

The most trivial and unobjectionable definition of emergence is simple novelty. If some entity has a property that its constituents lack then that property can be called "emergent". The liquidity of water is an emergent according to this definition. Somewhat embarrassingly however, while having mass is not an emergent feature, having a mass of 10 kg seems to be, since presumably none of the constituents of a 10 kg brick themselves weigh 10 kg. But no matter how we avoid these minor peccadilloes, it seems clear that consciousness is going to count as an emergent according to the novelty conception of emergence.

But is that enough?

In the first place, a lot of novel emergents are predictable given information about the constituents of the system in question. The bizarre property of superfluidity exhibited by liquid helium (and other so-called Bose-Einstein condensates) is highly novel but predicted by quantum mechanics. One of the most bewitching examples of this kind of emergence derives from what are called cellular automata. You are probably familiar with at least one CA, John Conway's "game of life". But since Daniel Dennett says that every philosopher should study it closely, let me review it briefly here. A CA is a grid or lattice of "elements" which can have various abstract properties that change according to strict rules of succession that depend upon the state of a given element and its neighbors in the lattice. The game of life has its own particular rules, which assign to each element or "cell" of the lattice one of two states, labeled "alive" and "dead" according to rules that govern how the lattice changes from generation to generation. The rules of life are:

> (1) Any cell which is alive and has either 2 or 3 living neighbors remains alive.
> (2) Any cell which is alive and has 4 or more living neighbors dies.
> (3) Any cell which is alive and has 1 or fewer living neighbors dies.
> (4) Any cell which is dead but which has exactly 3 living neighbors becomes alive.

Some features of life are obvious. An initial lattice consisting of only dead cells will remain totally dead forever. A small amount of thought reveals that three live cells in a row or column will perpetually switch between the row and column orientation (the two end cells die, the middle cell remains and the dead cells adjacent to the middle cell become alive since they have three living neighbors). Some simple patterns will remain forever unchanging, such as a 2x2 box (each cell in the box has exactly three living neighbors). Some simple patterns reproduce themselves over a few generations while in the process moving across the lattice (the simplest is the glider)[2].

But in general it is very hard to tell what is going to happen given some initial arrangements of living and dead cells. It turns out, in fact, that it is impossible for there to be any way to predict in advance how an arbitrary initial condition will evolve. All we can do is let the setup evolve and observe the outcome. This is because of a remarkable mathematical fact about the game of life: it is "Turing complete", which means that any computable function can be implemented by some life configuration or, to put the same point another way, any Turing machine can be simulated in some life configuration. Turing's most famous theorem was that it is in general impossible to compute whether a given Turing machine will halt or not (for a certain input) and this impossibility

translates directly to the game of life given the latter's Turing completeness (see Bedau 1997).

Does that make the structures of the game of life unpredictable in a way that makes them emergent in a stronger sense than mere novelty? I don't think so. It seems uncontroversial that simulatability is a mode of predictability. Our weather predictions nowadays are directly based upon computer simulations and while it may not be mathematically provable that there is no way to analytically solve the weather equations for a given initial condition, it is quite beyond the realm of any practical feasibility. Weather is, in effect, predictable only by simulation (and, of course, not very accurately predictable even by our most powerful supercomputers).

One of the classical emergentists – C. D. Broad— was acutely aware of the issue of practical computability long before anyone dreamed of the possibility of computer based simulations of natural processes. Thus he deployed the fiction of a mathematician of immense powers. As Broad put it (perhaps poking some fun at the famous physicist's weakness in mathematics), the point is to "bring out the logical distinction between mechanism and emergence. Let us replace Sir Ernest Rutherford by a mathematical archangel, and pass on" (1925, 70). Our computers do not yet aspire to angelic status, but we can imagine computational systems that are not burdened with the limitations of time and space. Let them have as much memory as needed and as much time. The issue ought only to be what can be simulated based upon the putatively fundamental description of the system at issue.

The game of life is of course simulatable in this sense. And it gives us no kind of emergence beyond that of trivial or predictable novelty. Nothing ever happens in the life world that would not show up in a computer simulation of that world. Perhaps this robs the example of all significance, because the abstract structure of cellular automata does not seem to be any kind of a model of the physics of the real world. It's worth pausing to note that this claim is not universally accepted.

It has sometime been conjectured that the ultimate foundation of the actual world is a cellular automaton (or cellular automaton-like) system (see Wolfram 2002 for some recent speculation; the first person to investigate this idea may have been Konrad Zuse 1969; the most actual work on the idea seems to be that of Edward Fredkin). The general program of modeling the world at its most fundamental level as a cellular automaton is called "digital physics" or, more provocatively, "finite nature" by Fredkin. The scale at which this hypothetical cellular automaton , "computes" the universe (let's call it the ultimate cellular automaton or UCA) would be very much smaller than even subatomic dimensions, probably approaching the Planck length (about 10^{-35} meters). In fact, there is no particular reason to suppose that the UCA works in what we call space. The neighborhoods around the cells of a CA are

purely abstract. We might thus hope that space, and time as well, are no less emergent features of the world than its more ponderable denizens. This would be an amazing vindication of Leibniz's old idea that the radically non-spatial monads' systems of perception generate the relational structure we call space. Time is perhaps a more problematic emergent for Leibniz since the monads themselves seem to evolve against a temporal background, but perhaps we can go so far as to consider the "instantaneous" (or infinitesimally temporally extended) states of all possible monads as the fundamental entities of the universe. Then time itself will appear as a special set of relations amongst these states (such an idea has been explored in the context of more standard physics by Julian Barbour 2001). Then instead of saying that the UCA operates at a length scale of about the Planck length, we should say that it is at that length that the universe would give clear evidence that it is the product of a CA. Unfortunately, experiments that could probe such a length scale are hard to imagine, but hopefully as digital physics is developed some experimentally accessible implications will emerge. Otherwise the doctrine can only remain pure metaphysics, albeit mathematical metaphysics.

No one knows how or if digital physics will pan out. The core problem is to link some CA architecture to the physics we already know. As Fredkin puts it: "the RUCA [reversible universal cellular automaton] runs a computation. As a consequence of that process and of appropriate initial conditions, various stable structures will exist in the lattice. For each such stable structure, we expect that its behavior will mimic the behavior of some particle such as a muon or a photon. What we demand of a correct model is that the behavior of those particles obeys the laws of physics and that we can identify the particles of the RUCA with the particles of physics" (Fredkin 2003, p. 192). The particles (and fields and everything else for that matter) of standard physics would all be emergent entities. The task of establishing how standard physics can be linked to some RUCA is immensely difficult and not yet very far advanced.

It might be thought that the continuous mathematics of the calculus, which serves as the basic mode of description for all our basic theories, and which has been applied with huge success throughout all domains of science, precludes any serious interpretation of the world as based upon the digital and discrete workings of a cellular automaton. After all, according to standard formulations of physics, many systems are capable of instantiating an infinite and continuous range of states (e.g. the position of a particle can take on any real number value). The situation is not so clear-cut however. It is possible to demonstrate rigorously that some CA systems generate behavior asymptotically describable in continuous terms. For example, the Navier-Stokes equations governing hydrodynamical flow can be retrieved from so-called lattice gas models that are types of CA (see Frisch et. al. 1986). Of course, I emphasize that no one knows whether all of continuous physics can thus be retrieved

or approximated from a CA model, but no one knows otherwise either (see Fredkin 2003, 2004). It is important to bear in mind that the "space" of the CA lattice is not the space of our physical universe (as deployed in scientific theory) and the time-tick of the CA is not the time we measure or experience. So the mere fact that the space and time of the CA are discrete does not rule them out even if we grant that the best scientific description of our space and time makes them continuous quantities.

The weird behavior of some quantum systems, in which two particles that have interacted maintain a kind of link across any distance so that interaction with one will instantaneously affect the state of the other, is called "entanglement" (the literature on what Einstein called "spooky action at a distance" is immense; see Mermin 1990; Hughes 1992, especially ch. 6). But, again, it is far from clear that this is a real problem for digital physics and for the same reason. The spatial distance separating the "parts" of an entangled system that makes entanglement seem "weird" need not be reflected in the underlying workings of our hypothetical universal CA. In fact, could it be that the phenomenon of quantum entanglement is trying to tell us that what we call spatial separation is not necessarily a "real" separation at the fundamental level of interaction?

The point of this digression is to emphasize that it remains an open possibility that the fundamental physics of our world could be CA-like. If so, then all the emergence there is in our world is just the kind of emergence found in the life world. And that is no more than trivial novelty.

But perhaps it is rather more likely that fundamental physics is not based upon cellular automata. Of course, that does not entail the existence of some form of emergence beyond trivial novelty. In fact, so long as the lynchpin principles of the mainstream view: completeness, closure and resolution are endorsed, it is hard to see how there can be any other kind of emergence. This claim may seem to be in tension with the undeniable existence of hierarchical structure in nature, as discussed above. Is there a real distinction between the expression of complex combinatory powers of low level entities which entirely stem from low level properties, and genuine emergence – an emergence which goes beyond trivial novelty? Or again, is there any real content to the purported distinction between the explanatory usefulness of high level structural accounts and the causal powers of high level structure? I think there is, but to see it requires focusing on the mainstream view as a metaphysical rather than a methodological or epistemological doctrine. We are not concerned with obtaining a practical understanding of everything in terms of full resolution. In fact, such understanding is quite impossible, for reasons of complexity that the mainstream view itself can spell out and fully expects to encounter. Innumerable immensely difficult questions arise at every stage of resolution, and there is no practical prospect whatsoever of knowing the full details of the physical resolution of even very simple physical complexes.

IV. The Superduper Computer Thought Experiment

The metaphysical picture is nonetheless clear. And since the world has no need to know the details but just runs along because the details are the way they are, the problems we have understanding complex systems in terms of fundamental physics are quite irrelevant to the metaphysics of the mainstream view. So, leaving aside issues of verifiability, let us take flight on the wings of imagination and engage in a purely philosophical thought experiment.

Imagine the day when physics is complete. A theory is in place that unifies all the forces of nature in one self-consistent and empirically verified set of absolutely basic principles. Of course, the mere possession of this theory of everything will not give us the ability to provide a complete explanation of everything: every event, process, occurrence and structure. Most things will be too remote from the basic theory to admit of explanation in its terms; even relatively small and simple systems will be far too complex to be intelligibly described in the final theory. But we are interested here in the ontological aspect of causation, not the explanatory side.

Seeing as our imagined theory is fully developed and mathematically complete it will enable us to set up detailed computer simulations of physical systems. The range of practicable simulations will in fact be subject to pretty much the same constraints facing the explanatory use of the theory; the modeling of even very simple systems will require impossibly large amounts of computational resources.

But consider a thought experiment that flatly ignores the inevitably insuperable problems of computational reality. Imagine a computer model of the final physical theory operating under "relaxed computational constraints": the simulation has no computational limits (we can deploy as much memory as we like and compute for as long as we like). Imagine that detailed specifications of the basic physical configuration of any system are available, so that if the configuration of any physical system is specified (to any given degree of accuracy) as input then the subsequent states of the system can be calculated (to a specified degree of accuracy). If the final theory is non-deterministic then we can permit multiple simulations, thus duplicating the statistics to be found in the real world.

Now, even though it is not physically realizable, I think the idea of such a computer program is perfectly well defined. So, let us imagine a superduper computer simulation of a part of the world. Consider first something "simple" – a bob on a spring on distant Sedna say. The simulation covers a restricted region of space and time (the programmer would set up "boundary conditions" representing the influence of the rest of the world), and must be defined solely in terms of the values of fundamental physical attributes over that region. The

programmer is not allowed to work with gross parameters such as the mass of the bob or the stiffness of the spring, or the gravitational force on Sedna, but must write her code in terms of the really basic physical entities involved. The mainstream view predicts that the output of this computer simulation, appropriately displayed, would reveal a bob bouncing up and down, suspended above Sedna's desolate and frozen surface. Now up the ante. Imagine a simulation of a more complex situation, for example a father and child washing their dog, in their backyard on a lovely sunny day. Do you think the simulation would mimic the actual events? The mainstream view asserts that such a simulation would "re-generate" both the action of the pendulum and the behavior of the father, child and dog (along with tub, water, soap, sunlight, etc.).

This thought experiment allows us to characterize emergence. An emergent is anything – objects, processes, behavior, etc – that appears in the simulation that is not coded explicitly into it. For example, computer models of weather systems do not have any explicit code for tornadoes yet tornadoes nevertheless manage to turn up in the simulated weather. In Dennett's (1981) famous example of the chess-playing program that tends to use its queen early, we have an emergent behavior. There is no "get-the-queen-out-early" code in the chess program.

Our old friend, emergence as trivial novelty appears when the emergents in the simulation exactly match what we find in the real world target of the simulation. Tornadoes are real features of the weather and the early use of the queen is a genuine aspect of chess play; and they appear in their simulations despite not being explicitly coded into it. This is a kind of predictability essentially similar to that found in the life world.

A stronger form of emergence, which we might call ontological or radical emergence, is revealed through events that either fail to appear in the simulation or diverge in behavior from their simulacra. To avoid the worry that our simulation fails simply because it depends upon an inaccurate account of the basic features it is attempting to simulate, let us just stipulate that the simulation is thus accurate (we are free to do this since we are assuming that the fundamental theory is the correct theory of basic physical reality). The definition of ontological emergence will involve the idea that the real world's behavior will depart from that of the simulated world even though the simulation contains a fully accurate portrayal of the fundamental features of the simulated system.

This computational metaphor avoids one obvious difficulty of real world searches for ontological emergence. In the real world, scientific investigation involves the use of complex instruments that might themselves possess ontologically emergent features. Inside the computational environment however the only "causal" powers at work are the powers of the fundamental level of the simulation. This is because we have manufactured computers so that – no

matter what emergent properties they might possess as complex physical systems in their own right – they compute only the functions they are programmed to compute. If we code in only the laws governing quarks, leptons, exchange bosons and the four forces (with initial condition) then these are the only processes that will drive the simulation. This will not of course stop the simulation from exhibiting high level structure and high level causal interaction, but these will be guaranteed to be – within the simulation – merely trivially novel features.

The concept of ontological emergence can be illustrated within the game of life example. Suppose we were given some cellular automata grids and were allowed to watch them evolve with the object being to figure out the underlying rules which governed the system. It is also permitted that we set up initial conditions as we wish upon the grid and set the system in action. We are allowed to assume that all our samples conform to the same set of rules but of course we are not told what these are. The analogy with the goal of fundamental science and experimentation is clear. Now suppose that our initial investigation strongly supports the idea that we are simply watching an implementation of the game of life itself, with its three rules of birth, death and persistence. We set up a computer simulation of the game of life and compare our simulation with our target, but we notice that our simulation sometimes diverges from the evolution of the target.

Then we notice something odd. Whenever a glider appears and exists for more than 20 time steps it becomes invulnerable. Completely contrary to the standard laws of life and in total opposition to the fundamental principles of purely local action in the standard life world, the age of some configurations makes a difference in how the grid evolves. The natural response would be to look for some internal clock that kept track of the age of the life world – a "hidden variable" within each cell. But, in general, there is nothing incoherent in the idea that there is simply an emergent feature depending upon time. This would be ontological or radical emergence.

V. Perils and Paradoxes of the Mainstream

Overall, the mainstream view is a beautiful and grand vision of the universe. It aims to unify everything in the world across all of space and all of time. It accepts and celebrates the overwhelming complexity of emergent systems while organizing them into a magnificent hierarchy of structure, development and function. But it leaves no place for consciousness.

To try to be more precise about what is already an intuitively disturbing situation, there are three especially significant, and inter-related, problems that arise when we think about the integration of consciousness into the mainstream view. The first problem stems from the way in which mainstream

emergents stand as mere epistemic resources in our quest to understand the world. The difficulty is that while the mainstream view is happy to embrace a "layered" view of nature in which there is a set of levels of increasing complexity in natural systems, from atoms to molecules to crystals, cells, organisms, etc. these levels are no more than explanatory shorthand. No one disputes that high level features are indispensable to our understanding of the world. The issue is whether high level features are, so to speak, part of nature's own way of structuring the dynamics of the world. So far as the mainstream view can discern, the drivers of the world are all, solely and entirely the fundamental features. While it would be wrong to say that high level features are mind dependent in the sense that without minds there would be no high level structure, it is absolutely correct to say that that high level features play no role in the world except as apprehended by minds. Until and unless some high level feature is taken up in conscious perception and thought by some cognitive agent, that high level feature stands as a merely potential epistemological resource for such agents.

If consciousness is an emergent within the purview of the mainstream view then it must similarly be no more than such a potential epistemic resource. This flies in the face of our introspective awareness of our own consciousness as something that is an occurrent, robustly ontologically solid, existent. In fact, we are more certain of the existence of consciousness than anything else, and a good deal more certain of it than we are of the mainstream view, which seems to demote consciousness to a kind of potential existence dependent upon the explanatory practices of other conscious beings.

The second problem is that the mainstream view makes conscious experience entirely epiphenomenal. This is because all high level emergents are epiphenomenal under the mainstream view. It is important to be clear about what this means. It would be evidently absurd to deny that hurricanes can wreak havoc, but that is not what is intended by my claim that the mainstream view entails that hurricanes are epiphenomenal. What I mean is that hurricanes don't cause anything in virtue of being hurricanes; the high winds inside a hurricane don't cause anything in virtue of being "high winds". Hurricaniness, if you will pardon the expression, does exactly nothing in the world's dynamics. But the concept of a hurricane is extremely useful to us in organizing experience as well as in explaining and predicting events. All the causal work is performed by the fundamental physics that underlie those complex systems that we, as epistemic agents, usefully categorize as hurricanes. This picture of the place of consciousness in the world is profoundly at odds with our current understanding of the role of consciousness. Each of the forms of consciousness that I cataloged at the beginning of this paper appears to us as possessing a distinctive and genuine causal efficacy. The way the pain feels – the conscious experience of it – surely matters to how we behave and in fact

constitutes our suffering.

The third problem is that the mainstream view's treatment of high level structure or emergent features generates a kind of paradox when applied to consciousness itself. The paradox is that if consciousness is just another high level feature of the world it stands as a mere epistemic potentiality until it is taken up and categorized as consciousness by some other cognitive agent who finds this categorization explanatorily or predictively useful. But my current consciousness is a fact about the world that is not at all merely potential and would persist as a fact even if all other consciousnesses were, this instant, obliterated. Perhaps the problem can also be expressed as a kind of vicious regress. The explanatory standpoints that take high level features beyond their mere epistemic potentiality are properties of conscious beings, but if conscious beings are themselves no more than high level features, then they require standpoints for them to go beyond mere epistemic potentiality. Each standpoint, as a high level feature, requires a conscious appreciation of it to transcend mere potentiality, but if consciousness itself is a high level feature then each consciousness requires a further consciousness to appreciate it. This vicious regress, or paradox, does not arise for any other of the high level features of the world that fit into their roles as epistemic resources smoothly and without complaint.

Thus another way to put the problem is to note that the mainstream view, contrary to appearances, subtly presupposes consciousness in its account of emergence. The layered view of reality that it endorses only makes sense if there are conscious beings around to appreciate all the high level structures, and their inter-relations, which can be discerned from various viewpoints. Without the viewpoints, the high level structure is completely otiose, playing no role in the intrinsic dynamics of the world.

The paradox is unavoidable, so long as the mainstream view forms the basis of our understanding of the universe. But the paradox is intolerable because it leaves a gaping hole in the metaphysical picture that stems from the mainstream view, undercutting the claim of completeness which is one of the main principles and supposed virtues of the mainstream view itself.

I can only give the merest sketch here of a few possible responses to this paradox. Two of them are rather hoary metaphysical standbys; one seems to me a rather different approach. None of them are altogether attractive.

In the face of the paradox of consciousness, we could modify the mainstream view in a way that retains some of its core principles. The ancient doctrine of panpsychism asserts that mind is a fundamental and ubiquitous feature of the world. This of course eliminates the problem of emergence. Since consciousness has been around always and everywhere it does not emerge out of radically non-mental precursors. But panpsychism faces many difficulties. One is simply the intuitive implausibility of the idea that everything – electrons, chairs, planets,

etc— partakes of mentality. A more principled objection to panpsychism is that even if we grant that the ultimate constituents of things are in some sense conscious, or proto-conscious, or however we might like to express it, we face the problem of moving from the individual minds of the constituents to the conscious experience of the composites formed from them. Evidently, we face here a kind of emergence whose explanation may be no less intractable than the emergence we are supposed to be rejecting.

If emergence is unavoidable, then perhaps we should embrace it "whole hog". Thus, another response to the paradox is radical emergence. This entails that complete rejection of the mainstream view. In particular, it entails that the models of fundamental science are empirically inadequate – that fundamental physics cannot provide a model in which all phenomena are correctly predicted. This is a bold claim and also a decidedly odd one insofar as radical emergence accepts both that fundamental physics can describe the elementary features of the world accurately and that all emergents appear as a matter of law entirely dependent upon these elementary features.

The final approach to the paradox I will consider depends upon a distinction between the empirical aspirations of science and the metaphysical hopes placed in science by philosophers and, perhaps, implicit in the mainstream view itself. The alternative view, which I call "surface metaphysics", questions the distinctively 20th century idea that science is the leader on the quest towards ultimate reality. In fact, it suggests to the contrary that "ultimate reality" is staring us all in the face, and that conscious beings are a part of it that does not need to be "explained" by fundamental physics in the sense demanded by the mainstream view. This does not mean that science has nothing to say about consciousness. Far from it. Science can investigate the empirical conditions that underlie consciousness and of course already has amassed a huge amount of knowledge about the relation between brain and consciousness. But perhaps the idea that science can go behind the world we experience to find its metaphysical "foundation" is a mistake – an error that reflection on consciousness itself reveals.

University of Toronto at Scarborough

References

Barbour, J. (2001). The End of Time, Oxford: Oxford University Press.

Bedau, M. (1997). "Weak Emergence", in J. Tomberlin (Ed.),"Philosophical perspectives: mind, causation, and world, Vol. 11. New York: Blackwell.

Broad, C. D. (1925). The Mind and Its Place in Nature, London: Routledge and Kegan Paul.

Dennett, D. (1971). "Intentional Systems", in The Journal of Philosophy 68, pp. 87-106. Reprinted in Dennett's Brainstorms, Cambridge, MA: MIT Press.

Descartes, R. (1641/1985). Meditations on First Philosophy, in J. Cottingham, R. Stoothoff and D. Murdoch (eds.) The Philosophical Writings of Descartes, Cambridge: Cambridge University Press.

Flanagan, O. (1992). Consciousness Reconsidered, Cambridge, MA: MIT Press.

Fredkin, E. (2003). "An Introduction to Digital Physics", in"International Journal of Theoretical Physics 42, No. 2 (2003) 189-247. (Fredkin's work can be found online at http://digitalphysics.org/)

Freud, S. (1938). The Basic Writings of Sigmund Freud, A. Brill (Ed. Trans.), New York: Random House.

Frisch, U., B. Hasslacher and Y. Pomeau (1986). "Lattice-Gas Automata for the Navier-Stokes Equation" Phys. Review Letters 56, pp. 1505-1508.

Hatfield, G. (2002). "Perception as Unconscious Inference" in Perception and the Physical World: Psychological and Philosophical Issues in Perception, Dieter Heyer and Rainer Mausfeld (eds.), New York: Wiley.

Hughes, R. (1992). The Structure and Interpretation of Quantum Mechanics, Cambridge, MA: Harvard University Press.

Huxley, T. (1866). Lessons in Elementary Physiology, London: Macmillan.

James, W. (1890/1950). The Principles of Psychology, v. 1, New York: Henry Holt and Co. Reprinted in 1950, New York: Dover.

Mangan, B. (2001). "Sensation's Ghost: The Non-Sensory 'Fringe' of Consciousness", Psyche 18, no. 7.

Mermin, D. (1990). Boojums All the Way through : Communicating Science in a Prosaic Age, Cambridge: Cambridge University Press.

Morgan, C. (1923). Emergent Evolution, London: Williams and Norgate.

Nagel, T. (1974). "What is it Like to be a Bat?" in Philosophical Review, 83, 435-50.

Wilkes, K. (1988). "____, Yishi, Duh, Um, and Consciousness", In A. Marcel and E. Bisiach (eds.) Consciousness in Contemporary Science, Oxford: Oxford University Press.

Wolfram, S. (2002). A New Kind of Science, Champaign, IL: Wolfram Media.

Zuse, K. (1969). Rechnender Raum, Braunschweig: Friedrich Vieweg & Sohn. Translated as Calculating Space, MIT Technical Translation AZT-70-164-GEMIT, MIT (Proj. MAC), Cambridge, Mass. 1970.

Footnotes

[1] For a careful discussion of Helmholtz's notion of'unconscious inference' see Hatfield (2002).

[2] Any number of implementations of the game of life can be found on the internet. Here's one: http://www.bitstorm.org/gameoflife/.