Strong and Weak Emergence

David J. Chalmers
Philosophy Program
Research School of Social Sciences
Australian National University

1 Two concepts of emergence

The term 'emergence' often causes confusion in science and philosophy, as it is used to express at least two quite different concepts. We can label these concepts *strong emergence* and *weak emergence*. Both of these concepts are important, but it is vital to keep them separate.

We can say that a high-level phenomenon is *strongly emergent* with respect to a low-level domain when the high-level phenomenon arises from the low-level domain, but truths concerning that phenomenon are not *deducible* even in principle from truths in the low-level domain.[1] Strong emergence is the notion of emergence that is most common in philosophical discussions of emergence, and is the notion invoked by the British emergentists of the 1920s.

We can say that a high-level phenomenon is *weakly* emergent with respect to a low-level domain when the high-level phenomenon arises from the low-level domain, but truths concerning that phenomenon are *unexpected* given the principles governing the low-level domain. Weak emergence is the notion of emergence that is most common in recent scientific

---

In (P. Clayton and P. Davies, eds.) *The Re-emergence of Emergence* (Oxford University Press, 2006). Most of this chapter was written for discussion at a Granada workshop on emergence, sponsored by the Templeton Foundation. One section (the last) is modified from a posting to the Usenet newsgroup comp.ai.philosophy, written in February 1990. I thank the editors and the participants in the Granada workshop on emergence for their feedback.

[1] In philosophers' terms, we can say that strong emergence requires that high-level truths are not conceptually or metaphysically necessitated by low-level truths. Other notions in the main text can also be formulated in these modal terms, but I will mainly talk of deducibility to avoid technicality. The distinction between conceptual

discussions of emergence, and is the notion that is typically invoked by proponents of emergence in complex systems theory. (See Bedau 1997 for a nice discussion of the notion of weak emergence and its relation to strong emergence.)

These definitions of strong and weak emergence are first approximations which might later be refined. But they are enough to exhibit the key differences between the two notions. As just defined, cases of strong emergence will likely also be cases of weak emergence (although this depends on just how 'unexpected' is understood). But cases of weak emergence need not be cases of strong emergence. It often happens that a high-level phenomenon is unexpected given principles of a low-level domain, but is nevertheless deducible in principle from truths concerning that domain.

The emergence of high-level patterns in cellular automata—a paradigm of emergence in recent complex systems theory—provides a clear example. If one is given only the basic rules governing a cellular automaton, then the formation of complex high-level patterns (such as gliders) may well be unexpected, so these patterns are weakly emergent. But the formation of these patterns is straightforwardly deducible from the rules (and initial conditions), so these patterns are not strongly emergent. Of course, to deduce the facts about the patterns in this case may require a fair amount of calculation, which is why their formation was not obvious to start with. Nevertheless, upon examination these high-level facts are a straightforward consequence of low-level facts. So this is a clear case of weak emergence without strong emergence.

Strong emergence has much more radical consequences than weak emergence. If there are phenomena that are strongly emergent with respect to the domain of physics, then our conception of nature needs to be expanded to accommodate them. That is, if there are phenomena whose existence is not deducible from the facts about the exact distribution of particles and fields throughout space and time (along with the laws of physics), then this suggests that new fundamental laws of nature are needed to explain these phenomena.

The existence of phenomena that are merely weakly emergent with respect to the domain of physics does not have such radical consequences. The existence of unexpected phenomena

---

and metaphysical necessity will not be central here, but in principle one could formulate finer-grained notions of strong emergence that take this distinction into account.

in complex biological systems, for example, does not on its own threaten the completeness of the catalogue of fundamental laws found in physics. As long as the existence of these phenomena is deducible in principle from a physical specification of the world (as in the case of the cellular automaton), then no new fundamental laws or properties are needed: everything will still be a consequence of physics. So if we want to use emergence to draw conclusions about the structure of nature at the most fundamental level, it is not weak emergence but strong emergence that is relevant.

Of course, weak emergence may still have important consequences for our understanding of nature. Even if weakly emergent phenomena do not require the introduction of new fundamental laws, they may still require in many cases the introduction of further levels of explanation above the physical level in order to make these phenomena maximally comprehensible to us. Further, by showing how a simple starting point can have unexpected consequences, the existence of weakly emergent phenomena can be seen as showing that an ultimately physicalist picture of the world need not be overly reductionist, but rather can accommodate all sorts of unexpected richness at higher levels, as long as explanations are given at the appropriate level.

In a way, the philosophical morals of strong emergence and weak emergence are diametrically opposed. Strong emergence, if it exists, can be used to reject the physicalist picture of the world as fundamentally incomplete. By contrast, weak emergence can be used to support the physicalist picture of the world, by showing how all sorts of phenomena that might seem novel and irreducible at first sight can nevertheless be grounded in underlying simple laws.

In what follows, I will say a little more about both strong and weak emergence.

2 Strong emergence

We have seen that strong emergence, if it exists, has radical consequences. The question that immediately arises, then, is: are there strongly emergent phenomena?

My own view is that the answer to this question is yes. I think there is exactly one clear case of a strongly emergent phenomenon, and that is the phenomenon of consciousness. We can say that a system is conscious when there is something it is like *to be* that system; that is,

when there is something it feels like from the system's own perspective. It is a key fact about nature that it contains conscious systems; I am one such. And there is reason to believe that the facts about consciousness are not deducible from any number of physical facts.

I have argued this position at length elsewhere (Chalmers 1996; 2002) and will not repeat the case here. But I will mention two well-known avenues of support. First, it seems that a colourblind scientist given complete physical knowledge about brains could nevertheless not deduce what it is like to have a conscious experience of red. Secondly, it seems logically coherent in principle that there could be a world physically identical to this one, but lacking consciousness entirely, or containing conscious experiences different from our own. If these claims are correct, it appears to follow that facts about consciousness are not deducible from physical facts alone.

If this is so, then what follows? I think that even if consciousness is not deducible from physical facts, states of consciousness are still systematically *correlated* with physical states. In particular, it remains plausible that in the actual world, the state of a person's brain determines his or her state of consciousness, in the sense that duplicating the brain state will cause the conscious state to be duplicated too. That is, consciousness still *supervenes* on the physical domain. But importantly, this supervenience holds only with the strength of laws of nature (in the philosophical jargon, it is natural or nomological supervenience). In our world, it seems to be a matter of law that duplicating physical states will duplicate consciousness; but in other worlds with different laws, a system physically identical to me might have no consciousness at all. This suggests that the lawful connection between physical processes and consciousness is not itself derivable from the laws of physics but is instead a further basic law or laws of its own. The laws that express the connection between physical processes and consciousness are what we might call fundamental psychophysical laws.

I think this account provides a good general model for strong emergence. We can think of strongly emergent phenomena as being systematically determined by low-level facts without being deducible from those facts. In philosophical language, they are naturally but not logically supervenient on low-level facts. In any case like this, fundamental physical laws need to be supplemented with further fundamental laws to ground the connection between low-level properties and high-level properties. Something like this seems to be what the

British emergentist C. D. Broad had in mind, when he invoked the need for 'trans-ordinal laws' connecting different levels of nature.

Are there other cases of strong emergence, besides consciousness? I think that there are no other clear cases, and that there are fairly good reasons to think that there are no other cases. Elsewhere (Chalmers 1996; Chalmers and Jackson 2001) I have argued that given a complete catalogue of physical facts about the world, supplemented by a complete catalogue of facts about consciousness, a Laplacean super-being could, in principle, deduce all the high-level facts about the world, including the high-level facts about chemistry, biology, economics, and so on. If this is right, then phenomena in these domains may be weakly emergent from the physical, but they are not strongly emergent (or if they are strongly emergent, this strong emergence will derive wholly from a dependence on the strongly emergent phenomena of consciousness). In short, with the exception of consciousness, it appears that all other phenomena are weakly emergent or are derived from the strongly emergent phenomenon of consciousness.

One might wonder about cases in which high-level *laws*, say in chemistry, are not obviously derivable from the low-level laws of physics. How can I know now that this is not the case? Here, one can reply by saying that even if the high-level laws are not deducible from the low-level laws, it remains plausible that they are deducible (or nearly so) from the low-level *facts*. For example, if one knows the complete distribution of atoms in space and time, it is plausible that one can deduce from there the complete distribution of chemical molecules, whether or not the laws governing molecules are immediately deducible from the laws governing atoms. So any emergence here is weaker than the sort of emergence that I maintain is present in the case of consciousness.

Still, this suggests the possibility of an intermediate but still radical sort of emergence, in which high-level facts and laws are not deducible from low-level *laws* (combined with initial conditions). If this intermediate sort of emergence exists, then if our Laplacean super-being is armed only with low-level laws and initial conditions (as opposed to all the low-level facts throughout space and time), it will be unable to deduce the facts about some high-level phenomena. This will presumably go along with a failure to be able to deduce even all the low-level facts from low-level laws plus initial conditions (since if the low-level facts were

derivable, the demon could deduce the high-level facts from there). So this sort of emergence entails a sort of incompleteness of physical laws even in characterizing the systematic evolution of low-level processes.

The best way of thinking of this sort of possibility is as involving a sort of *downward causation*. Downward causation means that higher-level phenomena are not only irreducible but also exert a causal efficacy of some sort. Such causation requires the formulation of basic principles which state that when certain high-level configurations occur, certain consequences will follow. (These are what McLaughlin (1993) calls configurational laws.) These consequences will themselves either be cast in low-level terms, or will be cast in high-level terms that put strong constraints on low-level facts. Either way, it follows that low-level laws will be incomplete as a guide to both the low-level and the high-level evolution of processes in the world.[2]

To be clear, one should distinguish *strong* downward causation from *weak* downward causation. With strong downward causation, the causal impact of a high-level phenomenon on low-level processes is not deducible even in principle from initial conditions and low-level laws. With weak downward causation, the causal impact of the high-level phenomenon is deducible in principle, but is nevertheless unexpected. As with strong and weak emergence, both strong and weak downward causation are interesting in their own right. But strong downward causation would have more radical consequences for our understanding of nature, so I will focus on it here.

I do not think there is anything incoherent about the idea of strong downward causation. I do not know whether there are any examples of it in the actual world, however. While it is certainly true that we can't *currently* deduce all high-level facts and laws from low-level laws plus initial conditions, I do not know of any compelling evidence for high-level facts and laws (outside the case of consciousness) that are not deducible in principle. But I think it is possible that we will encounter some. (See Kim (1992; 1999) for some doubts.)

---

[2] In such a case, one might respond by trying to introduce new, highly complex, low-level laws to govern evolution in these special configurations, in the effort to make low-level laws complete once again. But the point of this intermediate sort of emergence will still remain. It will just have to be rephrased, perhaps as the claim that non-configurational low-level laws are an incomplete guide to the evolution of processes. See Meehl and Sellars (1956) for related ideas here.

Perhaps the most interesting potential case of downward causation is in the domain of quantum physics, at least on certain 'collapse' interpretations of quantum mechanics. On these interpretations, there are two principles governing the evolution of the quantum wave function: the linear Schrödinger equation, which governs the standard case, and a nonlinear measurement postulate, which governs special cases of 'measurement'. In cases of measurement, the wave function is held to undergo a sort of 'quantum jump' quite unlike the usual case. A key issue is that no one knows just what is the criterion for a measurement taking place. Yet it is clear that for the collapse interpretation to work, measurements must involve certain highly specific causal events, most likely at a high-level. If so, then we can see the measurement postulate as itself a sort of configurational law, involving downward causation.

Both consciousness and the quantum measurement case can be seen as strong varieties of emergence in that they involve in-principle non-deducibility and novel fundamental laws. But they are quite different in character. If I am right about consciousness, then it is a case of a strongly emergent quality, while if the relevant interpretations of quantum mechanics are correct, then it is more like a case of strong downward causation.

In principle, one can have one sort of radical emergence without the other. If one has strongly emergent qualities without strong downward causation, one has an epiphenomenalist picture on which there is a new fundamental quality that plays no causal role with respect to the lower level. If one has strong downward causation without strongly emergent qualities, one has a picture of the world on which the only fundamental properties are physical, but on which their evolution is governed in part by high-level configurational laws.

One might also in principle have both strongly emergent qualities and strong downward causation together. If so, one has a situation in which a new fundamental quality is involved in new fundamental causal laws. This last option can be illustrated by combining the cases of consciousness and quantum mechanics discussed above. In the familiar interpretations of quantum mechanics according to which it is consciousness itself that is responsible for wave-function collapse, the emergent quality of consciousness is not epiphenomenal but plays a crucial causal role.

My own view is that, relative to the physical domain, there is just one sort of strongly emergent quality, namely, consciousness. I do not know whether there is any strong downward causation, but it seems to me that if there *is* any strong downward causation, quantum mechanics is the most likely locus for it. If both strongly emergent qualities and strong downward causation exist, it is natural to look at the possibility of a close connection between them, perhaps along the lines mentioned in the last paragraph. The question remains wide open, however, as to whether or not strong downward causation exists.

3 Weak emergence

Weak emergence does not yield the same sort of radical metaphysical expansion in our conception of the world as strong emergence, but it is no less interesting. I think that understanding weak emergence is vital for understanding all sorts of phenomena in nature, and in particular for understanding biological, cognitive, and social phenomena, as is demonstrated in many of the other chapters in this volume.

I gave a quick definition of weak emergence earlier. But it is more satisfactory to understand the notion by example, and then attempt to analyze it. The concept of emergence is often tacitly invoked by theorists in cognitive science and in the theory of complex systems, in such a way that it is clear that a notion of other than the notion of strong emergence is intended. We can take it that something like weak emergence is at play here, and we can then use the examples to make sense of just what weak emergence comes to.

It will help to focus on a few core examples of weak emergence:

(A) The game of Life: high-level patterns and structure emerge from simple low-level rules.

(B) Connectionist networks: high-level 'cognitive' behaviour emerges from simple interactions between simple threshold logic units.

(C) The operating system (Hofstadter 1977): the fact that overloading occurs just around when there are thirty-five users on the system seems to be an emergent property of the system.

(D) Evolution: intelligence and many other interesting properties emerge over the course of evolution by genetic recombination, mutation, and natural selection.

Note that in all these cases, the 'emergent' properties are in fact deducible (perhaps with great difficulty) from the low-level properties, perhaps in conjunction with knowledge of initial conditions, so strong emergence is not at play here.

One sometimes hears it suggested that emergence is the existence of properties of a system that are not possessed by any of its parts. However, this phenomenon is too ubiquitous for our purposes. Under this definition, file cabinets and decks of cards, and even XOR gates, have many 'emergent' properties. So this surely not what theorists generally mean by 'emergence'.

One might suggest that weak emergence involves 'deducibility without reducibility'. Of course the notion of reducibility is itself controversial and somewhat unclear. Biological and psychological laws and properties are frequently said not to be reducible to physical laws and properties, simply on the grounds that they might be found associated with all kinds of different physical laws and properties as substrates. However, some standard examples of weak emergence, such as the emergence of thermodynamics from statistical mechanics, involve phenomena that are reducible in this sense. And other phenomena that are not reducible in this sense, such as the functioning of a telephone, are not obviously emergent. So reducibility in this sense does not seem to be the key to weak emergence.

We might instead understand weak emergence in terms of the *ease of understanding* one level in terms of another. Emergent properties are usually properties that are more easily understood in their own right than in terms of properties at a lower level. This suggests an important observation: *weak emergence appears to be an observer-relative property*. Properties are classed as 'emergent' based at least in part on (1) how interesting the high-level property at hand is to a given observer, and (2) how difficult it is for an observer to deduce the high-level property from low-level properties. The properties of an XOR gate are an obvious consequence of the properties of its parts; emergent properties aren't. To capture this, we might suggest that weakly emergent properties are *interesting, non-obvious consequences* of low-level properties.

This still cannot be the full story, though. Every high-level physical property is a consequence of low-level properties, usually in a non-obvious fashion. It feels unsatisfactory, for instance, to say that computations performed by a COBOL program are an emergent property relative to the low-level circuit operations—at least this example feels much less naturally classed as 'emergent' than a connectionist network. So something is missing. The trouble seems to lie with the complex, jury-rigged *organization* of the COBOL system. The low-level processes may be simple enough, but all the complexity of the high-level behaviour is due to the complex *structure* that is given to the low-level mechanisms (by programming). By contrast, in the case of connectionism or the game of Life there is simplicity both in low-level mechanisms and in their organization. Consequently, in those cases the high-level processes have more of the character of 'something for nothing'.

To capture this, one might suggest that weak emergence is the phenomenon wherein complex, interesting high-level function is produced as a result of combining simple low-level mechanisms in simple ways. I think this is much closer to a good definition of emergence. Note that COBOL programs, and many biological systems, are excluded by the requirement that not only the mechanisms but also their principles of combination be simple. (Of course simplicity, complexity, and interestingness are observer-relative concepts, at least for now, although some have tried to explicate them in terms of Chaitin–Kolmogorov–Solomonoff complexity.) Note also that most phenomena that satisfy this definition should also satisfy

the previous definition, as complex and interesting consequences of simple processes will typically be non-obvious.

This conclusion captures the feeling that weak emergence is a 'something for nothing' phenomenon. And most of our examples fit. The game of Life and connectionist networks are clear cases: interesting high-level behaviour emerges as a consequence of simple dynamic rules for low-level cell dynamics. In evolution, the genetic mechanisms are very simple, but the results are very complex. (Note that there is a small difference, in that in the latter case the emergence is diachronic, i.e. over time, whereas in the first two cases the emergence is synchronic, i.e. not over time but over levels present at a given time.)

A residual problem is that it is not clear how (C), the operating system example, fits this paradigm of way of understanding emergence. But an appeal to principles of design should get us the rest of the way. We *design* the game of Life according to certain simple principles, but complex, interesting properties leap out and *surprise* us. Similarly for the connectionist network—we only design it at a low level, and we hope that complex high-level properties will emerge. For more traditional computer programs, by contrast, what one gets out is much closer to what one puts in. The operating system example also fits in well. The design principles of the system in this case are quite complex—unlike the other cases that fit (5) above—but still the figure 'thirty-five' is not a part of that design at all.

So we might suggest an alternative: A weakly emergent property of a system is an interesting property that is unexpected, given the underlying principles governing the system. Here the notion of 'underlying principles' is deliberately vague, so that it can be understood in multiple ways. One way to understand it is in terms of the principles according to which a principle is designed. Doing so well help capture cases discussed above. But we can also apply the definition to cases where the underlying principles are not, strictly speaking, designed at all. Corresponding to different ways of specifying the underlying principles of a system, we will have different sets of emergent properties.

In the case of evolution, for example, we might see the underlying principles as operating at the level of the gene. In this case the complex, interesting, high-level properties, such as intelligence, are unexpected relative to the underlying principles, and hence qualify as emergent. Alternatively, we might see the underlying principles as operating at the level of the

organism. On this construal, the most salient adaptive phenomena like intelligence are no longer unexpected in the same way, so they are less clearly emergent. However, there will then be other kinds of emergent phenomena, such as unexpected by-products of the evolutionary process (e.g. Gould and Lewontin's 'spandrels'). This construal also allows a potential account of one sense in which consciousness seems emergent. Raw consciousness may not have been selected for, but it somehow emerges as an unexpected by-product of selection for adaptive processes such as intelligence.

Overall, our initial understanding of weak emergence, in terms of phenomena that arise from a low-level domain but that are unexpected given the principles of that domain, seems to fit the cases quite well. But of course there is little point in deciding just which of these notions is the definitive analysis of 'weak emergence' as the notion is used in the sciences, just as there is little point in deciding just which of the notions in this chapter is the definitive analysis of 'emergence' itself. Typical uses of the term 'emergence' may well express cluster concepts with many different elements.

Still, we can reasonably hope that most or all of the notions discussed in this chapter may play some role in understanding the many uses of the term 'emergence' in the sciences and in philosophy. More importantly, we can hope that they can play some role in understanding the phenomena to which the term has been applied.

Bibliography

Bedau, M. (1997), 'Weak Emergence', *Philosophical Perspectives*, 11: 375–99.

Broad, C. D. (1925), *The Mind and its Place in Nature* (New York: Routledge).

Chalmers, D. J. (1996), *The Conscious Mind: In Search of a Fundamental Theory* (Oxford: Oxford University Press).

—— (2002), 'Consciousness and its Place in Nature', <http://consc.net/papers/nature.html>.

Chalmers, D. J., and F. Jackson (2001), 'Conceptual Analysis and Reductive Explanation', *Philosophical Review* 110:315–61.

Hofstadter, D.R. (1977), *Gödel, Escher, Bach: An Eternal Golden Braid* (New York: Basic Books).

Kim, J. (1992), 'The Nonreductivist's Trouble With Mental Causation', in J. Heil and A. Mele (eds.), *Mental Causation* (Oxford: Oxford University Press).

—— (1999), 'Making Sense of Emergence', *Philosophical Studies* 95: 3–36.

McLaughlin, B. P. (1992), ‘The Rise and Fall of British Emergentism’, in A. Beckermann, H. Flohr, and J. Kim (eds.), *Emergence or Reduction?: Prospects for Nonreductive Physicalism* (Berlin: De Gruyter).