Synthese (2006) 151: 547–559
DOI 10.1007/s11229-006-9025-0

ORIGINAL PAPER

Emergence: Core ideas and issues[1]

Jaegwon Kim

Published online: 9 August 2006
© Jaegwon Kim 2006

Abstract This paper explores the fundamental ideas that have motivated the idea of emergence and the movement of emergentism. The concept of reduction, which lies at the heart of the emergence idea is explicated, and it is shown how the thesis that emergent properties are irreducible gives a unified account of emergence. The paper goes on to discuss two fundamental unresolved issues for emergentism. The first is that of giving a "positive" characterization of emergence; the second is to give a coherent explanation of how "downward" causation, a central component of emergentism, is able to avoid the problem of overdetermination.

Keywords Downward causation · Emergence · Emergentism · Reduction · Supervenience

Since around 1990, the idea of emergence has been making a big comeback, from decades of general neglect and disdain on the part of mainstream analytic philosophers. Indications are that the emergence boom is going to continue, on an upward trajectory, for years ahead. What it is about emergence that makes it such an attractive idea to so many thinkers, with diverse and disparate backgrounds and agendas—philosophers, practicing scientists from a variety of scientific fields, and science writers—is itself an intriguing philosophical question, or perhaps a question in the sociology of science. The term "emergence" seems to have a special appeal for many people; it has an uplifting, expansive ring to it, unlike "reduction" which sounds constrictive and overbearing. We now see the term being freely bandied about,

[1] Kim (2004). This paper is largely based on my "Making Sense of Emergence", *Philosophical Studies 95* (1999), 3–36, and "On Being Realistic about Emergence". The latter was presented at the Emergence/Reduction Workshop at the Institut Jean Nicod in the fall of 2003; as this paper has been promised for another publication, the present paper was expressly composed in order to provide a critical target for the commentaries by Marras and Wong.

J. Kim (✉)
Department of Philosophy, Brown University, Box 1918, Providence,
RI 02912, USA
e-mail: Jaegwon_Kim@brown.edu

especially by some scientists and science writers, with little visible regard for whether its use is underpinned by a consistent, tolerably unified, and shared meaning (and if so what it is). This has created situations in which those discussing emergence, even face to face, more often than not talk past each other. Sometimes one gets the impression that the only thing that the participants share is the word "emergence". The intuitive associations this word evokes in us do not add up to a concept robust enough to do any useful work, or even to serve as helpful constraints on a theoretical account or construction of the concept. "Emergence" is very much a term of philosophical trade; it can pretty much mean whatever you want it to mean, the only condition being that you had better be reasonably clear about what you mean, and that your concept turns out to be something interesting and theoretically useful.

Of course we do not start with a totally blank page when we now ponder how best to understand emergence. For there are signposts that should guide our reflections, at least in their initial stages. Any account of emergence, I believe, should show significant continuity with the concept that the British emergentists of the early 20th century, such as Alexander, Morgan, and Broad, had in mind, and we should take care that our understanding of the concept should be as charitable as possible in regard to the central doctrines of the emergentism these philosophers espoused. This is not to say that these emergentists were in complete agreement with one another; however, there was a sufficient conceptual and doctrinal convergence among them to generate a shared philosophical perspective and a movement—what now goes by the name "British emergentism". Nor does it mean that we must stay with the early emergentists to the end; at some point, new possibilities and opportunities may well prompt us to proceed in different directions and explore new options.

As is well known, the intuitive idea of an emergent property stems from the thought that a purely physical system, composed exclusively of bits of matter, when it reaches a certain degree of complexity in its structural organization, can begin to exhibit genuinely novel properties not possessed by its simpler constituents. Questions and disputes arise when we try to make this idea more precise and turn it into something that is philosophically and scientifically useful. For the friends of emergence, to say that a given property is an emergent property of some system must be saying something significant and explanatory about the property and the system that has it. But exactly what does it mean to say that something is an emergent property? What does it mean to say that it is a "novel" property? What do emergent properties *do* after they have emerged?

In this paper, I want to set out and discuss two conditions that I believe must be considered necessary components of any concept of emergence that is true to its historical origins in the early 20th century. The conditions are supervenience and irreducibility. How reducibility is to be understood in this context will require some discussion. I will then take up the question whether supervenience and irreducibility together could be considered sufficient for emergence, and argue for a negative answer. Finally, I will turn to the question what needs to be done if we are to turn emergence into a robust and metaphysically coherent concept. As we will see, the idea of "downward" causation will loom large in my discussion. The picture I will present is not going to be reassuring to the friends of emergence, for my conclusion will essentially be something close to this: Emergentism cannot live without downward causation but it cannot live with it either. Downward causation is the raison d'être of emergence, but it may well turn out to be what in the end undermines it.

1 Emergence and supervenience

In order to help fix the concept of emergence, let us begin with the question whether emergent properties should be thought of as supervenient on their basal conditions—conditions at the lower level from which they emerge. For our purposes, it will be useful to take as our starting point a helpful survey article by Van Gulick (2001). According to Van Gulick, three grades of emergence can be distinguished: (1) "specific value emergence"; (2) "modest kind emergence"; and (3) "radical kind emergence". The first, specific value emergence, is a pretty tame affair exemplified in a situation like the following: a whole, say a bronze statue, has a mass of 1 kg whereas none of its parts, though they all have a mass, has that particular value of mass, namely 1 kg. Whence the name "specific value emergence". It is clear that this is not a case of what emergentists have in mind when they speak of emergent properties; Van Gulick probably wanted to recognize cases of this kind because they do fit the letter of the emergentist slogan "A whole can (and often does) have properties that none of its constituents have".

Van Gulick explains the second, "modest kind emergence", as follows: "The whole has features that are different in kind from those of its parts... For example, a piece of cloth might be purple in hue even though none of the molecules that make up its surface could be said to be purple. Or a mouse might be alive even if none of its parts (or at least none of its subcellular parts) were alive" (Van Gulick 2001, p 17). The examples offered are the sort that has traditionally been claimed to be emergent. What Van Gulick does not say is just in what sense the purple color of a cloth "emerges" from the properties of its constituent molecules. But before we take up this question, let's look at Van Gulick's third and strongest kind of emergence.

Van Gulick defines "radical kind emergence" as follows: "1. (the emergent property is) different in kind from those had by its parts, and 2. (it is) of a kind whose nature and existence is *not necessitated by* the features of its parts, their mode of combination and the law-like regularities governing the features of its parts" (Van Gulick 2001, p 17, emphasis added). The second condition, which is what distinguishes this kind of emergence from its weaker siblings, asserts that an emergent property of a whole is not determined by the properties and relations characterizing its parts, or, to put it another way, an emergent property of this third kind does not *supervene* on the microstructure of the object that has it. Thus, two wholes may have identical microstructure (i.e., composed of identical basic physical constituents configured in an identical structure) and yet differ in respect of their emergent properties. For example, two molecule-for-molecule identical systems may be such that one of them is a live mouse and the second is not—that is, if being a live mouse is emergent in the sense of "radical kind" emergence. Van Gulick is uncertain whether there are actual cases of radical kind emergence, saying that accepting it would violate "atomistic physicalism". However, the real problem with this purported form of emergence is not its violation of the doctrine of microdeterminism (or mereological supervenience). A more immediate concern is whether it is a form of emergence at all. We should remember that classic emergentists accepted microstructural supervenience of emergent properties,[2] and it is easy to see why.

[2] For example, Broad, arguably the most astute of the British emergentists, writes: "No doubt the properties of silver chloride are completely determined by those of silver and of chlorine; in the sense that whenever you have a whole composed of these two elements in certain proportions and relations you have something with the characteristic properties of silver chloride." For Broad, the characteristic properties of silver chloride are emergent properties (Broad 1925, p 64).

Suppose that on a given occasion a mental phenomenon, say pain, emerges from a certain configuration of neural events. It is highly unlikely that an emergentist will deny that if the very same configuration of physiological events were to recur, the same mental phenomenon, pain, would emerge again. If the connection between pain and its neural substrate were irregular, haphazard, or coincidental, what reason could there be for saying that pain "emerges from" that neural condition rather than another? What would be the point of saying that pain is an emergent property rather than saying that it is a property randomly distributed over neural states? The very meaning of "neural substrate" implies the presence of a regular determinative, or necessitating, relationship. If supervenience, or upward necessitation, is taken away, that takes away something essential to the meaning of "emergence" in Van Gulick's radical kind emergence. I am not insisting here that the dependence of an emergent property on its "basal" condition be deterministic; I believe there may well be a viable concept of *statistical* or *stochastic* emergence, which assigns a stable objective chance of the emergence of a property given that an appropriate basal condition is present. But stochastic emergence in this sense must be based on statistical laws with nomological force, and these laws can warrant talk of *stochastic supervenience* and *stochastic necessitation*.

It is clear then that we must consider supervenience as a component of emergence—that is, we need to accept the following proposition:

> *Supervenience*: If property M emerges from properties $N_1, \ldots, N_n$, then M supervenes on $N_1, \ldots, N_n$. That is to say, systems that are alike in respect of basal conditions, $N_1, \ldots, N_n$ must be alike in respect of their emergent properties.

2 Emergence and reduction

Morgan, a leader of British emergentism, said that "resultant" properties—that is, properties of a whole that are not emergent—are "additive and subtractive only, and predictable" (Morgan 1923, p 3) from information concerning the basal conditions. The implication of course is that emergent properties are neither additive nor subtractive, and not predictable, on the basis of the lower-level properties from which they arise. It has also been said that emergent properties are not explainable in terms of their basal properties—that is, for example, we cannot explain why someone is experiencing pain (rather than itch or tickle) on the basis of the neural processes from which pain emerges.

Let us consider some relations a property can bear to a set of other properties. The first, which we have just discussed, is supervenience:

1. Supervenience/determination: Property M supervenes on, or is determined by, properties $N_1, \ldots, N_n$ in the sense that whenever anything has $N_1, \ldots, N_2$, it necessarily has M.

I argued earlier that if M emerges from $N_1, \ldots, N_2$, these properties must meet the supervenience condition. According to classical emergentism—that is, British emergentism—this determination thesis must be sharply distinguished from each of the following claims (we will see below what Broad says on this issue):

2. Predictability: The occurrence of M—that is, whether M will be instantiated on a given occasion—can be predicted from the occurrence of $N_1, \ldots, N_2$; full information concerning whether $N_1, \ldots, N_n$ are instantiated in a system suffices for the prediction of whether the system instantiates M.
3. Explainability: Why a system instantiates M can be explained, understood, and made intelligible in terms of its instantiating $N_1, \ldots, N_n$.

Emergentists will deny that predictability and explainability hold for an emergent and its basal conditions. When Morgan denied that emergent properties are "additive" or "subtractive", what he had in mind presumably is that their occurrences cannot be mathematically calculated, or logically deduced, on the basis of the properties from which they emerge (so the whole is not a mere "sum" of its parts). Emergentists will say that the fact that pain emerges from a certain neural state (say, C-fiber excitation) is a brute fact that cannot be explained, and that full and ideally complete knowledge of the neurophysiology of the brain does not suffice for prediction of conscious states. Thus, determination is one thing; explainability and predictability are quite another.

This naturally leads to the following question: What is required for explanation and prediction that goes beyond mere supervenience or determination? Emergentists were quite aware that there is a sense in which the occurrence of an emergent phenomenon can be predicted. Consider an inference like this:

> Jones's C-fibers will be stimulated at *t*.
> Anyone whose C-fibers are stimulated will experience pain.
> Therefore, Jones will experience pain at *t*.

This may be called an "inductive" prediction of pain—based on our inductive knowledge of the pain/C-fiber stimulation correlation. It should be clear why a prediction of pain of this kind will not impress an emergentist who asks "Can an emergent phenomenon be predicted on the basis of knowledge of its basal conditions?" What is wrong with the above prediction of pain is that the evidence base, in the second premise of the inference, makes use of knowledge of facts going beyond those at the basal level; it assumes knowledge of the "emergence law" linking pain with C-fiber stimulation. Similar comments apply to the following "inductive" explanation of pain:

> Jones's C-fibers were stimulated at *t*.
> Anyone whose C-fibers are stimulated will experience pain.
> That is why Jones experienced pain at *t*.

What the emergentist had in mind, when he denied the explainability of emergents, is what we may call "reductive" explanation, or what some emergentists called "mechanistic" explanation—an explanation of why a whole has a certain property exclusively on the basis of its constituent microstructure.

Broad contrasted emergentism with "mechanism", a view that we would now call "reductionism". Consider the following passage in which the contrast is described:

> "(The two approaches) differ according to the view that we take about the laws which connect the properties of the components with the characteristic behaviour of the complex wholes which they make up. (1) On the first form of the theory the characteristic behaviour of the whole *could* not, even in theory, be deduced from the most complete knowledge of the behaviour of its components, taken separately or in other combinations, and of their proportions and arrangements in this whole. This alternative ... is what I understand

> by the 'Theory of Emergence'. ... (2) On the second form of the theory the characteristic behaviour of the whole is not only *determined* by the nature and arrangement of its components; in addition to this it is held that the behaviour of the whole could, in theory at least, be *deduced* from a sufficient knowledge of how the components behave in isolation or in other wholes of a simpler kind. I will call this kind of theory 'Mechanistic'" (Broad 1925, p 50; emphasis in the original).

Both emergentism and reductionism—that is, Broad's "mechanism" – agree in holding that "the behavior of the whole" is determined by—that is, it supervenes on—the properties and structural relationships characterizing its components. Where they differ, according to Broad, concerns the in-principle deducibility—presumably, logical/conceptual deducibility—of the properties of the whole from facts at its basal level. Reductionism asserts, and emergentism denies, the deducibility of all properties of a whole from the properties at the basal level—that is, the properties and relationships for its constituents. Deducibility is crucial in this context, for when the British emergentists discussed predictability and explainability in connection with emergents, it is plain that they construed prediction and explanation as a matter of deduction or derivation. The question "Is emergent property *M* explainable in terms of basal facts *F*?" comes down to "Is the fact that *M* is instantiated deducible from *F*?" Similarly, and only more obviously, the question "Is the occurrence of *M* predictable from *F*?" is simply the question "Is the occurrence of *M*—that is, the fact that *M* will be instantiated—deducible from *F*?"

This throws further light on why the two displayed inferences above, concerning Jones's pain, do not count as an explanation and prediction of the kind demanded by reductionism. The reason, as noted, is that they both invoke, in their second premise, facts about the emergent property in question (i.e., pain). This is why the Nagelian model of "bridge-law" reduction (Nagel 1961), which dominated recent discussions of reduction, is irrelevant to the question of reductive prediction and explanation in the present context. Nagelian reduction, too, is derivational reduction; however, as is well known, the Nagel model allows as auxiliary premises of the derivation the use of "bridge laws" connecting properties to be reduced with base-level properties—what Broad called "trans-ordinal" laws, like the second premise of the two inferences above. This is expressly prohibited in reductive derivations as envisaged by Broad: for the emergentist the real explanatory question is why these particular bridge laws—for example, one connecting pain with C-fiber excitation—hold. Why does pain arise from, or correlate with, C-fiber excitation, but not another kind of neural state? Why does pain, not itch or tickle, correlate with C-fiber excitation? Nagelian reduction allows the use of psychoneural correlations as unexplained additional premises of reductive derivations, and this, by the emergentist's light, begs the question on hand.

However, there is another model of reduction—reduction through functionalization, or, briefly, functional reduction—which better suits our purposes.[3] Bridge laws are the heart of Nagel reduction: in fact, it can be shown that the availability of bridge laws is necessary and sufficient for Nagel reduction (Kim 1998, p 91). In contrast, functional reduction proceeds via the "functionalization" of the properties to be reduced in terms of properties at the base level. Suppose pain can be given a "functional" definition like this:

[3] For more details see Kim (1988, 1999). For some criticisms see Marras (2002).

To be in pain $=_{\text{def.}}$ to be in a state that is typically caused by tissue damage and trauma and that typically causes aversive behavior.

This definition connects pain conceptually with physical/behavioral properties. Reduction of pain is accomplished for a population of interest to us (say, humans, mammals) when we are able to identify a "realizer" of pain so conceived, namely a physical state that fits the functional definition, for that population. So suppose neuroscientific research has identified N_1 (say, the activation of a group of neurons in certain cortical areas) as the state that is typically caused by tissue damage and which in turn triggers aversive behavior. N_1 would be what is standardly called the neural "correlate" or "substrate" of pain (for humans, mammals). When we have such a neural state for the population of interest to us, we have a neural reduction of pain for this population.

A neural reduction of pain, it should be noted, does not require a logical derivation of pain from a neural state—in particular, from its neural realizer; and it does not require logical or conceptual connections between pain and neural states. Trying to derive a pain statement from statements about neural states would be hopeless. What we should keep in mind is the fact that the mind–body problem involves three players, not two; they are pain (and other mental states), the brain, and behavior. Reduction requires conceptual connections, but these connections connect pain with behavior, not directly with the brain. Brain phenomena enter the picture as the realizers of the functionally conceived mental phenomena. It is important to notice that the fact that N_1 is a realizer of pain (for a given group of organisms), or that the brain is the realizer of mentality, is an empirical and contingent (though lawful) fact. What is not contingent is the relation between pain and pain behavior. I am not saying that pain can in fact be reduced this way; what I am saying is that if pain is to be reduced to a brain process, the following is what must be accomplished: pain must first be given a functional definition or interpretation and then we must identify its neural realizers. The first step involves conceptual work: Is the concept of pain functionally definable or interpretable and if so how should a functional definition of pain be formulated? The second step, that of discovering the realizers of pain, is up to empirical scientific research. It is in effect the research project of finding the neural correlates of conscious experiences. From a philosophical point of view, the crucial question, therefore, is whether pain can be given a functional characterization, in terms of physical input and behavioral output; the rest is up to science. Philosophical functionalism, still the orthodoxy on the mind–body problem, holds that pain, along with other mental phenomena, can be functionalized; if philosophical functionalism is correct, all mental phenomena will be functionally reducible and hence nonemergent. I am with those who do not believe pain and other sensory states ("qualia") can be given functional characterizations (e.g., Chalmers 1996). However, this does not change the fact that functionalizability is crucial to reduction and reducibility, and hence to understanding emergence (as we shall shortly see). Conscious experience, or anything else for that matter, is reducible if and only if it is functionally reducible, and it is functionally reducible only if it is functionally definable or interpretable.

As noted earlier, it would be futile to try to derive pain from its neural correlate. However, it is quickly shown that if pain can be functionalized, we can derive a statement to the effect that pain occurs from facts at the behavioral/neural level. And this will vindicate functional reduction as characterized here as the appropriate concept of reduction for the purposes of defining and clarifying emergence. So suppose that

pain has been functionally reduced with neural state N_1 identified as its realizer for a given population. A derivation of the occurrence of pain can proceed as follows:

System *s* is in neural state N_1 at *t*.
N_1 is such that tissue damage in *s* and systems like *s* causes them to go into N_1, and N_1 causes these systems to emit aversive behavior.
By definition, a system is in pain iff it is in some state *P* such that *P* is caused by tissue damage and *P* in turn causes aversive behavior.
Therefore, *s* is in pain at *t*.

The derivation is valid. Note that the third line, as a conceptual definition, does not count as an additional premise. It introduces no facts about pains or about how pains correlate with instances of N_1; if it is about anything, it is about the meaning of the term "pain" or the concept pain; it does not state a fact about pains. In general, definitions come free in proofs; they do not count as premises. Consequently, the derivation gives an affirmative answer to the emergentist question "Can the occurrence of pain be derived from information about lower-level facts alone?" This is the crucial difference between this derivation and the earlier derivation using Nagelian bridge laws as auxiliary premises. Both derivations invoke laws; the difference is that the law used in the Nagelian derivation is a "trans-ordinal" or "bridge" law connecting the emergent level with the base level, whereas the law used in the functional derivation (i.e., the second line) concerns the base level alone.

It is clear that the above derivation can serve as a reductive explanation of the occurrence of pain and also as a predictive inference to the occurrence of pain. Thus, a functional reduction of a mental property, *M*, guarantees the following:

(1) Instantiations of *M* can be predicted on the basis of information concerning neural and behavioral processes alone (including laws concerning these processes).
(2) Similarly, why an organism instantiates *M* at a time can be explained on the basis of information concerning facts at the lower level, namely neural and behavioral facts.

This shows that functional reduction gives a unified account of the emergentist idea that an emergent property is irreducible to the basal phenomena and neither explainable nor predictable in terms of them. Moreover, a functional reduction of pain has the following causal and ontological implications:

(3) Each occurrence of pain has the causal powers of its neural realizer; thus if pain occurs by being realized by N_1, this occurrence of pain has the causal powers of N_1. In fact, the pain can be identified with this instance of N_1. In general, if *M* occurs by being realized by N_1 on a given occasion, the *M*-instance has the causal powers of the *N*-instance.

If two individual events have identical causal powers, there is little reason to think of them as distinct. In fact, Davidson has proposed a causal criterion of event identity: "Events with the same causes and the same effects are one and the same event" (Davidson 1969). Even if we do not accept Davidson's suggestion as a "criterion" of event identity, it may nonetheless be true and can serve as a good metaphysical guide. If this is right, functional reduction also accomplishes ontological reduction:

(4) If M is instantiated in virtue of the instantiation of its realizer N_1 on a given occasion, the M-instance is identical with the N_1-instance.[4]

We have, therefore, identified a second condition of emergence:

Irreducibility of emergents: Property M is emergent from a set of properties, $N_1, \ldots, N_n$, only if M is not functionally reducible with the set of the Ns as its realizer.

Thus, supervenience and functional irreducibility are two necessary conditions of emergence. Are they together sufficient? I will return to this question in the section to follow and argue for a negative answer. I believe, however, that the two conditions together capture the concept as it was intended by the classical emergentists like Alexander, Morgan, and Broad. This means that the notion of emergence in British emergentism was under-characterized. When we consider recent proposals concerning emergent properties in complex systems, in terms of such ideas as chaotic, nonlinear dynamics, the necessity of simulation (rather than computation), and so forth, we should, first of all, examine them to see whether they fit the classic conception of emergence encapsulated in our two conditions. Of course, to judge that one or another of these new proposals does not fit the classic conception does not by itself show that it is not an interesting and potentially fruitful concept. But the conjunction of supervenience and functional irreducibility can serve as a useful benchmark; any deviation from it is a deviation from the classic conception, and new proposals can be analyzed and compared with one another in terms of how far, and in what ways, they deviate from the classical conception.

3 Are supervenience and irreducibility sufficient for emergence?

There are reasons for thinking that emergence would be under-characterized and under-explained if supervenience and irreducibility together were taken to constitute its full definition—that is, as a necessary and sufficient condition of emergence. At least, this would be the case if one expects emergence to be a robust and natural relation holding between a higher-level property and its lower-level base properties.

First, consider supervenience, or its core notion of dependence. Brief reflection will show that supervenience is not a homogeneous natural relation. This becomes clear when one considers some sample cases of supervenience and asks why supervenience holds in each case. Take, for example, the supervenience of normative properties on factual, nonnormative properties. Why does supervenience hold for these two sets of properties? What grounds normative supervenience? Different metaethical theories give competing and mutually exclusionary answers. Ethical naturalists will say that the supervenience holds because ethical properties (e.g., being good) are fully definable in terms of nonormative, naturalistic properties (e.g., being desired, being pleasurable). Nonnaturalists, or intuitionists, will say that there are synthetic necessary connections between ethical properties and certain nonethical, natural properties that we intuit through our moral sense. Noncognitivists and moral antirealists will see normative supervenience as arising from certain consistency requirements on the use of ethical language or our moral and evaluative practices (e.g., the requirement of universalizability or generalizability). Each of these is a possible answer to the question what

[4] For objections to this line of argument for (4) see Gillett (forthcoming).

grounds normative supervenience, but they all advert to different grounding relations. Under each explanation normative supervenience holds; however, the only thing common to the three explanations is the fact that ethical/normative properties covary in a certain way with naturalistic properties. But property covariations are "phenomenological" relations—surface phenomena arising from, and requiring explanations in terms of, deeper underlying relations.

We see an analogous situation with mind–body supervenience. Many diverse mind–body theories accept supervenience; for example, type physicalism, functionalism, epiphenomenalism, emergentism, and the double-aspect theory. (Token physicalism and even substance dualism are at least consistent with mind–body supervenience.) Each will give a different explanation of why the mental supervenes on the physical; for example, functionalism will advert to the fact that physically indiscernible systems have identical physical causal powers; epiphenomenalism will invoke the "same cause, same effect" principle; emergentism, rather like nonnaturalism in regard to normative supervenience, will say that mind–body supervenience is grounded in brute and fundamental physical–mental law-like connections (primitive "laws of emergence"), which Alexander counseled us to accept with "natural piety". And so on. Again, this shows that supervenience, or dependence, is not a homogeneous relation; it is not a genuine, "natural" relation, but rather something that arises from natural relations holding at a deeper level, like causation and reduction.

This means that the bare statement that a family of properties supervenes on another does not tell us much. For it to be philosophically informative and enlightening we must know the deeper relation that grounds and explains why supervenience holds between these two sets of properties. Now, according to classical emergentism, *that is precisely the kind of information we cannot have*: emergence is brute, and that means that the supervenience relation holding in cases of emergence, too, is brute and unexplainable. If we could explain why pain, not itch or tickle, supervenes on C-fiber stimulation (or, more correctly, why pain has C-fiber stimulation as its supervenience base), the emergence of pain from a neural state would no longer be a brute unexplainable fact. *So the demands of emergentism make the supervenience relation involved in emergence necessarily unexplainable; we cannot know what kind of dependence grounds and explains the supervenience relation involved in emergence*. This impossibility of knowledge has nothing to do with our epistemic limits; rather, for classical metaphysical emergentists like Broad, there is nothing here, no fact of the matter, to be known.[5] The upshot is that the supervenience condition on emergence simply amounts to the assertion that there is an in-principle *unexplainable* covariation between the putatively emergent properties and their base properties. This cannot be considered a substantive positive characterization of the emergence relation.

Let us now turn to irreducibility. The first thing we notice is that it is a *negative* characterization. If we know that X is reducible to Y, we know something interesting and important about the relationship between X and Y. And if we also know that U is reducible to W, we know something common that the pairs $\langle X, Y\rangle$ and $\langle U, W\rangle$ share. I believe we can take reducibility as a genuine relation characterizing two domains of properties, or two theories. But this does not mean that irreducibility, namely the *absence* of reducibility, is also a genuine and informative relation. As has often been

[5] I will shortly be urging that the first item on the emergentists' "to-do-list" ought to be a positive characterization of the emergence relation. However, the point I am making in this paragraph makes the prospects of producing such a characterization dubious, to say the least. Here is perhaps where the neo-emergentists would have to depart from classical emergentism.

observed, being red is a property but that does not mean that being nonred is also a genuine property. There are too many diverse things that are nonred: green things, yellow things, transparent things, numbers, atoms and molecules, thoughts and ideas, propositions, and countless other sorts of things. The same applies to relations and their negations. Number theory is irreducible to hydrodynamics and vice versa. Chemical properties are irreducible to biological properties; geological properties are irreducible to economic properties and vice versa. If emergent properties are irreducible to their base properties, does this instance of irreducibility have anything in common with those other cases of irreducibility? The answer, I believe, has to be "none".

What we have in supervenience and irreducibility, therefore, are two essentially negative conditions, and they do not amount to a *positive* account of what emergence really is. They tell us what emergence is not; they do not tell us anything—at least, not much—about what it is. I believe one pressing item on the emergentist agenda is to provide an illuminating positive characterization of emergence. Some current work on emergence, in fact, can be seen as attempts in that direction.[6] However, it remains to be seen whether any of them will succeed. Success here includes at least two things: first, the proposed characterization of emergence must explain why emergents so characterized supervene on their base properties and why, in spite of the supervenience relation, the former are not reducible to the latter; second, it must successfully cope with the problem of downward causation. We turn to this problem in the next section.

4 The problem of downward causation

It is critically important to the emergentists that emergent properties have distinctive causal powers of their own, irreducible to the causal powers of their base properties. Properties that are lacking in causal powers—that is, whose possession by an object makes no difference to the causal potential of the object—would be of no interest to anyone; in fact, it was Alexander who equated the existence of an entity with its having causal powers, saying that an epiphenomenon "might as well, and undoubtedly would in time be abolished" (Alexander 1927, p 8). In fact, some philosophers identify properties with clusters of causal powers (Shoemaker 1980). And the causal powers the emergents bring with them must be new and distinctive (remember: emergent properties are supposed to be "novel"); if they were reducible to the causal powers of the base-level properties, they would be bringing nothing new and would have nothing new to contribute to the evolving causal structure of the world.

Suppose then that an instance of an emergent property, M, causes another emergent property M^* to instantiate. This, we might say, is an instance of "same-level" causation. Now, M^*, as an emergent, must have a basal (physical) property P^* from which it emerges; M^* cannot be instantiated unless some appropriate basal condition, say P^*, is present; moreover, the presence of P^* by itself guarantees that M^* will be instantiated at that time, *no matter what has preceded this occurrence of* M^*. That is, as long as P^* is there at the time, M^* will be there at the same time *whether or not* M^**'s purported cause,* M*, had been there at all*—unless, that is, M had something to do with P^*'s presence at that time. In fact, the only way to save the claim that M caused M^* appears to be to say that M caused M^* *by causing* P^*. It makes sense to think that in order to bring about an

[6] Newman (1997), Humphreys (1997), Bedau (1997), Silberstein (2001), Rueger (2000), O'Connor and Wong (forthcoming), Campbell and Bickhard (forthcoming), and many others.

emergent phenomenon (or a supervenient property), you must bring about an appropriate basal condition from which it will emerge. If pain emerges from neural state N, then to cause pain you must bring about N. (And to extinguish pain you must get rid of N—that's why we take ibuprofen to alleviate pain.) In any case, if these considerations are correct, they show that same-level causation, from M to M^*, entails "downward" causation from M to P^*. This is downward causation because P^* is a property at the base level; if M and M^* are mental properties, P^* would be a physical/neural property.

This means that to understand the possibility of causation between emergent phenomena we must understand the possibility of downward causation. Apart from this, downward causation is of paramount importance to the emergentists. For they want to claim that the emergence of consciousness and rational thought has made a fundamental difference to the world at the physical level. It is because of our emergent mental powers that we have built cities and bridges, sent spaceships to Jupiter and Saturn, destroyed rain forests, and burned holes in the ozone layer. But can we understand how downward causation is possible?

In our schematic example above, we concluded that M causes M^* by causing P^*. So M causes P^*. Now, M, as an emergent, must itself have an emergence base property, say P. Now we face a critical question: if an emergent, M, emerges from basal condition P, why cannot P displace M as a cause of any putative effect of M? Why cannot P do all the work in explaining why any alleged effect of M occurred? If causation is understood as nomological (law-based) sufficiency, P, as M's emergence base, is nomologically sufficient for it, and M, as P^*'s cause, is nomologically sufficient for P^*. It follows that P is nomologically sufficient for P^* and hence qualifies as its cause. The same conclusion follows if causation is understood in terms of counterfactuals—roughly, as a condition without which the effect would not have occurred. Moreover, it is not possible to view the situation as involving a causal chain from P to P^* with M as an intermediate causal link. The reason is that the emergence relation from P to M cannot properly be viewed as causal.[7] This appears to make the emergent property M otiose and dispensable as a cause of P^*; it seems that we can explain the occurrence of P^* simply in terms of P, without invoking M at all. If M is to be retained as a cause of P^*, a positive argument has to be provided. If M is somehow retained as a cause, we are faced with the highly implausible consequence that every case of downward causation involves causal overdetermination (since P remains a cause of P^* as well).[8] Moreover, this goes against the spirit of emergentism in any case: emergents are supposed to make distinctive and novel causal contributions. However, if there is systematic causal overdetermination in all cases of downward causation, emergents cannot fulfill their causal promise; anything they causally contribute can be, and is, contributed by a physical cause. This result, unless it is successfully rebutted, threatens to bankrupt one of the central claims of emergentism. If downward causation goes, so goes emergentism.

The above line of consideration is often referred to as the "exclusion" or "supervenience" argument (for other ways of formulating the argument see Kim 2003, 2005).

[7] Morgan explicitly denies that emergence is a form of causation (Morgan 1923, p 28). Moreover, there is little to recommend in the claim that a neural state causes pain and then pain in turn causes, say, my hand withdrawal. How can there be a causal chain from pain to the hand motion that is separate and independent from the physical causal chain from the neural state to the motion of the hand?

[8] For an argument that systematic overdetermination is an acceptable option for mental causation see Mills (1996). However, Mills does not specifically argue from the perspective of emergentism.

5 Concluding remarks

The idea of emergence is an attractive, and initially appealing, one in many ways, and it is not difficult to understand its popularity. But it is not easy to make the idea precise and give it substantive content. Two important unresolved items remain on the emergentists' agenda. The first is to give emergence a robust positive characterization that goes beyond supervenience and irreducibility. The second is to come face to face with the problem of downward causation. Somehow the emergentist must devise an intelligible and consistent account of how emergent properties can have distinctive causal powers of their own—in particular, powers to influence events and processes at the basal level.

References

Alexander, S. (1927). *Space, time, and deity*, vol. 2. London: Macmillan.
Bedau, M. (1997). Weak emergence. *Philosophical Perspectives, 11*, 375–399.
Broad, C. D. (1925). *The mind and its place in nature.* London: Routledge & Kegan Paul.
Campbell, R. J., & Bickhard, M. H. (forthcoming). Physicalism, emergence and downward causation.
Chalmers, D. J. (1996). *The conscious mind*. New York and Oxford: Oxford University Press.
Davidson, D. (1969). The individuation of events. In N. Rescher et al. (Ed.), *In Essays in Honor of Carl G. Hempel* Dordrecht, Holland: Reidel Publishing Co. (Reprinted in Davidson (1980). *Essays on Actions and Events*. Oxford: Oxford University Press).
Gillett, C. (forthcoming). Understanding the new reductionism: the metaphysics of realization and reduction by functionalization.
Humphreys, P. (1997). How properties emerge. *Philosophy of Science, 64*, 1–17.
Kim, J. (1998). *Mind in a physical world*. Cambridge, Mass.: MIT Press.
Kim, J. (1999). Making sense of emergence. *Philosophical Studies, 95*, 3–36.
Kim, J. (2003). Blocking causal drain and other maintenance chores with mental causation. *Philosophy and Phenomenological Research, 67*, 151–176.
Kim, J. (2005). *Physicalism, or something near enough*. Princeton: Princeton University Press.
Marras, A. (2002). Kim on reduction. *Erkenntnis, 57*, 231–257.
Mills, E. (1996). Interactionism and overdetermination. *American Philosophical Quarterly, 33*, 105–117.
Morgan, C. L. (1923). *Emergent evolution*. London: Williams and Norgate.
Nagel, E. (1961). *The structure of science*. New York: Harcourt, Brace and World.
Newman, D. V. (1997). Chaos, emergence, and the mind-body problem. *Australasian Journal of Philosophy, 79*, 180–196.
O'Connor, T., & Wong, H. Y. (forthcoming). The metaphysics of emergence. *Noûs.*
Rueger, A. (2000). Physical emergence, diachronic and synchronic. *Synthese, 124*, 297–322.
Shoemaker, S. (1980). Causality and properties. In P. V. Inwagen (Ed.), *Time and cause*. Dordrecht, Holland: Reidel Publishing Co.
Silberstein, M. (2001). Converging on emergence. *Journal of Consciousness Studies*, 8, 61–98.
Van Gulick, R. (2001). Reduction, emergence and other recent options on the mind-body problem: a philosophic overview. *Journal of Consciousness Studies, 8*, 1–34.