The Dualist

Stanford's Undergraduate Philosophy Journal

Volume XX Spring 2015

This issue is dedicated to our advisors:

Jonathan Ettel, Nathan Hauthaler
and Tamar Schapiro

THE DUALIST

Volume XX ■ Spring 2015

Department of Philosophy

Stanford University

The Dualist
Volume XX
Spring 2015

TABLE OF CONTENTS

Trapped in a Chinese Room:
What Computationalists Can Learn from Searle's Argument

Cody Rosevear
University of Toronto

John Searle's Chinese Room argument questions the capacity of formal symbol manipulation to serve as a suitable explanation of mind as such, with all of its intentional phenomena, rather than simply as a theory of the engineering of intelligent behaviour. In particular, Searle is concerned whether or not a machine suitably programmed to exhibit the linguistic behaviour of a native Chinese speaker can be justifiably said to possess an intentional understanding (in the sense that its mental states actually harbour propositional content) of Chinese, rather than being a mere simulation of such understanding, and whether any program capable of engendering the proper behaviour should be construed as constituting an explanation of the phenomena it simulates (Searle, 417).

Searle answers all these questions in the negative, in virtue of the following: that he would have no means of generating any intentional understanding of Chinese if he were sequestered in a room with English instructions telling him which Chinese symbols to output in response to the Chinese symbols fed to him as input. From this phenomenological fact, he infers that no formal system, qua formal system, is sufficient for intentional understanding. The argument may be summarized thusly:

> I have inputs and outputs that are indistinguishable from those of the native Chinese speaker, and I can have any formal program you like, but I still understand nothing. For

> the same reasons, Schank's computer understands nothing of any stories, whether in Chinese, English, or whatever, since in the Chinese case the computer is me, and in cases where the computer is not me, the computer has nothing more than I have in the case where I understand nothing. (Searle, 418)

It is evident from the above that on Searle's construal, all computational accounts of cognition, in virtue of being computational, are unable to instantiate the kind of intentional understanding possessed by human minds. Naturally, given the proclivity for conceptual ambiguity in pre-theoretical notions of mental phenomena, one is concerned to know precisely what it is about understanding that Searle believes to be so decisive a factor in allowing for judgment about the unsuitability of computational accounts of cognition, simply because of their (alleged) necessary lack of possession of said phenomena.

Searle makes it clear that what he is concerned with is not understanding insofar as it is construed as a behavioural capacity only, defined in terms of giving appropriate output in response to stimuli, linguistic or otherwise; but rather, the capacity to understand as defined from the first person point-of-view (Searle, 419). That is, Searle is not concerned with whether a machine can be programmed to replicate human performance, but with whether or not it actually has a comparable intentional mental state that harbors actual propositional content.

Indeed, the following makes clear that what Searle is concerned to demonstrate with regards to computational machines is that they cannot, in principle, harbour any mental states that are defined in terms of the possession of propositional content, as opposed to merely instantiating the proper syntactic relations which engender within the machine those behavioural norms appropriate to the cognitive domain being simulated:

> Second, the program is purely formal, but the intentional states are not in that way formal. They are defined in terms of their content, not their form. The belief that it is raining, for example, is not defined as a certain formal shape, but as a certain mental content with conditions of satisfaction, a direction of fit (see Searle 1979), and the like. Indeed the belief as such hasn't even got a formal shape in this syntactic sense, since one and the same belief can be given an indefinite number of different syntactic expressions in different linguistic systems. (Searle, 423)

Insofar, then, as the question is whether systems of formal symbol manipulation can be said to possess intentional understanding, Searle would have us infer from the dissociation between his theoretical capacity to produce the appropriate behavioural output by following the rules of a program, and the lack of any intentional understanding engendered within him by said symbol

manipulation, that computational accounts of cognition are thus necessarily incomplete. The grounds, however, are dubious, even if one rejects purely behaviouristic formulations of the reality of mental states.

In order to observe this, it is important to note the structure of the analogy being made between the man in the room and the implementational processes in the computer. While there can be little doubt that the man, simply in virtue of consciously manipulating formal symbols, will not come to understand Chinese, and that it is indeed true that any AI system ultimately consists of varieties of symbol manipulation such that it would possess nothing that the man in principle could not, there nevertheless remains a fatal difference between the scenarios. Namely, that in the case of the man's manipulation of symbols, the implementational process (his brain) is already a fully functioning agent with intentional understanding, while the computer's is not.

This is significant not with respect to determining whether the Chinese room is capable of intentional understanding in any meaningful sense, but in assessing the validity of Searle's inference from his intentional phenomenology while in the room, to the conclusion that anything that simply manipulates formal symbols is incapable of possessing intentional understanding.

To make his case, Searle points out that even though, by hypothesis, he is implementing a program capable of producing the right behavioural output, he in fact understands nothing in Chinese, and that since a computer implementing the program would have nothing more than he has at his disposal, (namely, the capacity to follow the program's instructions) then the formal symbol manipulation account of intentional understanding must be false.

This is in fact the same reasoning which motivates Searle's rebuttal to the well-known systems reply, which states that while the actual program of symbols on its own is incapable of understanding, the entire room, construed as a system composed of a CPU (the man), an input output system (the door), a set of memory banks (the scratchpads), as well as the actual program (the English to Chinese translation instructions) nevertheless does understand Chinese, even in Searle's first person sense. Against this, Searle argues that if he can implement the entire program which produces the Chinese linguistic output by memorizing all the rules and symbols, while nevertheless failing to understand the content of the symbols, then no other system can either, because, again, it will possess nothing more than he does.

In both cases, Searle utilizes his intentional phenomenology while he is implementing the program at the level of conscious awareness to debunk the claim that what intentional understanding requires is simply the appropriate sequences of formal symbol manipulation. The problem, however, is that the inference, relying as it does on a failure to distinguish the personal from the sub-personal levels of the mind, is fallacious.

Let the term intentional understanding refer to Searle's conception of the notion of understanding at issue; namely, whether or not the agent in question harbours mental states with propositional content in addition to the requisite behavioural capacity. Furthermore, let the terms 'personal level' refer to any activities/cognitions carried out by an agent as a whole (such as solving

arithmetic problems), and 'sub-personal' to refer to those processes carried out only by parts of the agent (such as any of the unconscious processes occurring below awareness).

Now, consider: since the man is already a fully functioning intentional system, he must possess sub-personal processes which account for his prior intentional understanding and behavioural linguistic capacity with regards to his native language, whether those processes are formal symbol manipulations or not. This, however, highlights an important point: given that what occurs at the personal level (the level of the man's intentional understanding) is a result of what occurs at the sub-personal level (the man's brain processes, whether formal or not), it is fallacious to infer that the lack of intentional understanding in the man while he is implementing a program entails that the formal symbol manipulation being carried out at the personal level by that man cannot instantiate intentional understanding in a computer when those processes are run entirely at the sub-personal level of the computer[1].

To illustrate more concretely the problematic nature of Searle's inference, consider the following: when an individual adds two numbers together, provided they are not too large, the personal level carries out the operation (almost) directly; there is very little (if at all) in the way of intermediate operations (such as carrying digits when the numbers are too large) which take place at the personal level. In contrast, a program designed to add will rely on a number of primitive hardware operations which, when executed, will ultimately result in the same output as a human performing the addition operation directly.

Now, consider the case where an individual is made to add some numbers together by carrying out a sequence of manipulations that correspond to the primitive operations of a computer on collections of numbers in unfamiliar notation. If the individual were to carry out these operations manually they would produce the appropriate output. It does not follow, however, that since those primitive operations being manually performed at the personal level are completely absent when the individual does mental arithmetic, that the processes underlying that personal level execution of arithmetic are not in fact like those primitive operations in the computer (again, whether they actually are or not is irrelevant, it is simply the form of Searle's inference here that matters).

In the same way, it does not follow from the lack of intentional understanding generated by explicit rule following, that the intentional understanding agents already have of their native language is not the result of formal symbol manipulation at the sub-personal level.

---

[1] Technically, Searle can't even infer that there is not a second mind of which he is unaware that is running on his own as he manipulates the Chinese symbols, since in that case, even though Searle is unaware of it at the level of his phenomenology, Searle's manual symbol manipulations done at the personal level would, simultaneously, constitute the sub-personal processes of any virtual mind present (whether or not this is actually possible is irrelevant; I am simply concerned with the validity of Searle's inference).

The point here is that there is an important distinction to be made between a cognitive system that is carrying out a program manually at its personal level, and the program which 'runs' said cognitive system . The former no more entails that the agent carrying out the program should come to intentionally understand anymore than manually carrying out the program that allows a robot to see would entails seeing what the robot sees. Searle's argument is problematic precisely because it runs roughshod over this distinction, equating manually carrying out a program with being run by it. Given, however, that computationalism is in no way committed to such an equivalency, Searle's inference from the intentional phenomenology of the man in the room as he manipulates symbols to the conclusion that no intentional understanding is possible in virtue of such manipulation is divested of its force.

This is not to say, however, that Searle's critique is entirely without merit, for although as a refutation of computationalism the Chinese room does not live up to its author's intentions, it nevertheless exposes a considerable explanatory debt that has yet to be discharged by the computationalist thesis. Namely, that if formal symbol manipulation is to explain how intentional understanding, with all of its propositional content, arises, then it should be able to allow for inferences regarding the presence of said phenomena on the basis of the principles of the program itself, independent of (or at least not entirely dependent on) behavioural considerations. If not, one's analysis of intentionality cannot help but reduce to behaviourism, since the grounds for attribution ultimately do not derive their weight from the agent operating on the basis of particular processes internal to the system (in this case, the program), but instead from the appropriateness of the behavioural output of said processes.

Thus, on pains of capitulating to a defunct theoretical model of mind, intentional understanding cannot be construed as being constituted by 'behaving as if one understands'; in this, Searle is correct. However, this does not mean that symbol formal manipulation, properly characterized, cannot ever provide a suitable explanation of intentional understanding (perhaps, for example, an account which included causal linkages with the world via embodiment will provide the necessary resources). Searle's critique, though it fails as a proof of the impossibility of machine intentionality, nevertheless highlights that one cannot conflate the problem of explicating intentionality with the problem of engineering intelligent behaviour.

The Chinese room is thus a re-articulation of the problem of original intentionality: the question as to how it is possible for physical systems to refer or be about states of affairs in the world, intrinsically, without doing so as the result of attributions of intentionality by already intentional agents (and what the conditions of the determination of content for such intentionally rich mental states are). That Searle's argument is merely a re-articulation of this problem can be observed from the fact that the Chinese room emphasizes that syntax and semantics are dissociable concepts, and that the presence of a syntax governed normatively by semantic constraints, even if it outputs the appropriate behaviour for any given cognitive domains, does not, in and of itself, constitute an explanation of how some formal systems actually come to harbour, without

derivative interpretation by an outside observer, the propositional content which serves as the basis for the constraint on the formal manipulation of syntax in the first place. This is just to say, however, that we don't have an adequate account of intentionality.

A common response to the problem is to simply stipulate that any syntactic system which satisfies the right input/output patterns are, ipso facto, vehicles of propositional content, and thus possess intentionality. This move, however, faces a variety of explanatory debts that must be discharged if the stipulation is to be justified.

Firstly, it can reasonably be inquired why any syntactic system in particular, even though it may be behaving incorrectly according to some normative standard, cannot also be said to nevertheless harbour propositional content. After all, to stipulate that only those systems whose state-transitions are governed by the appropriate normativity possess propositional content is, again, to capitulate to unacceptably behaviourist criteria for attributions of intentionality, since it would, in effect, be denying that 'unintelligent' systems can possess intentionality. Prima facie, however, there is no reason to think that this is true: just because one's thoughts are not appropriately governed by the right normative rules in any given context does not demonstrate that one's thoughts have no content at all; merely that the state transitions between them are faulty (for some given normative standard). One is thus owed an explanation as to why a system's patterns of inputs and outputs should be regarded as the sole criterion for attributions of intentionality, *independent* of a prior belief in computationalism.

In a similar vein, even if one were to articulate a convincing argument as to the necessity of a system adhering to normative standards in order to posses intentional content, there would nevertheless remain the question of what degree of intelligence is required before an arbitrary system can be said to possess intentional understanding. That is, if 'correct' and 'intelligent' behaviour is taken to be the definitive standard with respect to whether a system can plausibly be said to posses intentional understanding, one is still owed an explanation as to what the *principled* difference is between simpler systems which lack extensive 'intelligent' behaviour and would not normally be considered candidates for attributions of intentionality, but nevertheless satisfy their domain specific semantic constraints, (such as a simple numerical calculator); and those which *are* considered candidates, given that the difference is merely one of a degree of complexity, rather than kind. One could, of course, challenge the notion that there is a difference in kind between syntax and semantics, but this is, in effect, the very question at issue.

Therefore, to demonstrate that the determination of semantic content within the minds of cognitive agents is a result of syntax would be tantamount to providing an account of *how* and hence *why* one should adhere to the criterion of intentionality implicit within computationalism, independent of a prior commitment to it. Whether such an account of the determination of semantic content is possible remains an open question, whose solution would go a long

way towards vindicating the tenability of computationalism as a theory of mind, rather than merely intelligent behaviour.

Searle's own articulations make clear that he does not think that any variation on the computational theme will have the capacity to fully explicate intentional understanding:

> But the main point of the present argument is that no purely formal model will ever be sufficient by itself for intentionality because the formal properties are not by themselves constitutive of intentionality, and they have by themselves no causal powers except the power, when instantiated, to produce the next stage of the formalism when the machine is running. And any other causal properties that particular realizations of the formal model have, are irrelevant to the formal model because we can always put the same formal model in a different realization where those causal properties are obviously absent. (Searle, 422)

It might reasonably be asked, given the condemnation of formal systems above, what the appropriate methodological paradigm for the study of the mind should be if it is not to be in any way computational. Whether or not any shifts in methodological practice have come about as a result of Searle's argument, it is clear that insofar as neuroscience is construed as an alternative methodological paradigm for studying the mind, it would indeed seem to at least address Searle's conception of the problem of intentionality, at least as articulated above.

For insofar as neuroscience can be said to study the mind, it does so via the brain, which, being a biological organ, is studied much like all other biological systems in the life sciences: it is investigated form a purely physical perspective, such that talk of rules, representations, and formal properties are eschewed in favour of direct experimentation and observation of the object of interest in order to determine the neural basis of both behavioural and cognitive phenomena. This direct, experimental analysis of the causal nature of the brain as a physical system is certainly much more in line with Searle's emphasis on avoiding the alleged explanatory poverty of formal systems (due to their multiple realizability and hence physical causal variation), in virtue of focusing on the concrete, particular physical mechanisms underlying known cases of intentional understanding (humans). This is in direct contrast to the way of framing the problem that Searle decries; namely, abstracting from the physical particulars on the grounds that they are mere implementational details, of no theoretical consequence.

Although whether or not neuroscience will prove to have better intellectual resources when it comes to explicating intentional understanding (or that recourse to neuroscience is even necessary) is an open question, one might wonder why the issue of intentionality was overlooked during AI's original

formulation as a model of the mind in the first place, and why it is so often conflated with the possession of intelligence.

It might be said that the holy triumvirate of the cognitive sciences are intelligence, intentionality, and consciousness, and that though each of these are inextricably intertwined, they are nevertheless suitably distinct enough in character to justify focusing on one at the expense of the others.

It is, therefore, not necessarily surprising that given the intellectual milieu during which AI was originally formulated, when the consequences of theoretical computer science were being harnessed towards the construction of machines capable of performing cognitive feats hitherto relegated solely to the province of humankind, that those interested in studying the mind would opt to attempt to explicate that member of the triumvirate to which the newly developed insights would be most profitably be applied: intelligence.

Indeed, given the limitations of the technology at the disposal of neuroscience at the time, and the presence of the aforementioned intellectual tools, such a decision would have certainly have been justifiable on pragmatic grounds insofar as it is construed as a temporary methodological strategy. If, however, AI ever reaches human level performance, in order to be a truly viable theory of mind, rather than simply a methodology for the engineering of intelligent behaviour, proponents of the computational thesis will have to tackle the problem of intentionality head on.

Works Cited

Searle, John R. “Minds, Brains, and Programs.” *Behavioural and Brain Sciences* 3.3 (1980): 417-424. Print.

The Layered Model of Biological Emergence

Katie Morrow
Seattle Pacific University

Introduction

There are at least four different (though interrelated) questions about reductionism in the philosophy of science literature. (1) The first is methodological: is it predictively or explanatorily better for scientists to model system behavior in terms of interactions of parts, or in terms of system-level features? For example, there have been long-running disputes in numerous applications whether system-level models can be as epistemically secure as methodologically "reductive" models which describe system behavior in terms of the smallest feasible components. (2) Another issue is what relationship the theories and concepts of the special sciences have to fundamental physics (i.e., to base-level theory). Can theory from the domain of biology in some sense be derived from or closely mapped to theory utilizing only physical terms? (3) A third question is whether properties or states of higher-level entities (objects, processes, or events) can be reduced to, or perhaps identified with, those of their smaller constituent entities. This question is of clear interest in the metaphysics of mind, but the issue seems to generalize to any complex, organized physical systems, such as biological organisms. (4) A final issue is whether higher-level objects (or their states or properties) have real causal powers, or whether their apparent causal efficacy is merely epiphenomenal on fundamental physics.

It is clear that these questions about reduction have important consequences for the methodology of science, the epistemology of science, and the metaphysics of physical systems. I am interested in many of these

consequences, but I will not attempt to directly address them in this paper. My interest here is instead in the way in which many of these questions have been framed in the philosophical literature, particularly the literature pertaining to biological systems. It seems to me that a certain framework has been commonly implicitly or explicitly utilized in articulating these issues; that the framework has received little explicit evaluation; and that such an unevaluated framework has the potential to misrepresent the issues.

The framework I mean is this: all of the above questions about reduction have routinely been described as questions about *interlevel* relationships. I have followed this convention in stating the four questions in the first paragraph. To further illustrate, reductionists about the various questions typically make claims such as: *higher-level* models are predictively worse than *lower-level* models; *higher-level* theories, or properties, reduce to *lower-level* items; the causal powers of *higher-level* objects or properties can be reduced to or identified with those of items on a *lower level*, perhaps the *base level* of physics. So, the notion of levels is used both to frame these questions and to state the views that might be taken on them. It is the widespread assumption that physical entities can be usefully thought to occupy a hierarchical series of levels which I will dispute in this paper.

More specifically, I will argue that the layered model does not adequately capture the structure of biological systems. My project is closely related to that of Angela Potochnik and Brian McGill (2012), who have argued against the empirical utility and accuracy of a strict hierarchy of levels framework, especially as applied to the context of scientific explanation in ecology. Their paper provides a starting point for my project, although I will treat a slightly different set of issues. In particular, I am interested in consequences of the layered model for the metaphysics of complex systems, with a focus here on the case study of emergent properties—a special class of high-level properties (or states) which have been thought not to be causally reducible to lower-level properties.

There are three parts to this paper. First I will characterize the received layered model. I will briefly discuss some issues in the historical development of the model, and then I will examine the case study of emergent properties to illustrate what implications the model has in contemporary metaphysical applications. Then I will briefly argue that the layered model is not utilized by biologists, despite the common assumption that the model has been borrowed from science. Finally, I will spend the rest of the paper on two cases of emergence from biology that I think begin to show the inadequacy of the model to real physical applications.

The Layered Model

The layered model portrays the physical universe as stratified such that entities can be ranked as higher or lower than each other, where lower-level items are those which are more basic, universal, or fundamental, and higher-level items are those described by the special sciences. Furthermore, physical objects are thought to be arranged such that smaller (lower-level) ones compose

higher-level kinds of objects. Higher-level objects are thus dependent both for their existence and for their features on lower-level composition. Moreover, this hierarchical arrangement is supposed to be relevant to the discussion of causal and other metaphysical relationships within and among physical objects.

Little recently has been written to explicitly evaluate the layered model. An exception to this is a helpful paper by Jaegwon Kim entitled "The Layered Model: Metaphysical Considerations" (2002). We should take note of Kim's warning that "it isn't easy to come up with a neat and satisfying general model of levels that will serve useful philosophical purposes" (2002: 3). There have not been any detailed proposals since some highly idealized models, now clearly too neat to be true, were developed during the early to middle part of the last century. These included the British Emergentists, e.g. C. Lloyd Morgan (1923), as well as the well-known paper by Oppenheim and Putnam (1958). These historical views invoked discrete, universal layers, such that every object (or property) in the physical universe could be ranked as higher or lower with respect to every other object (property).

The most obvious issue for such a model of universal levels is that only select classes of entities have commensurable special-science descriptions. Both a rabbit-sized rock and a rabbit are made of molecules, but does the rock occupy a higher level than its molecules? If so, does it occupy the same level as the rabbit's cells, or the rabbit? Or how does a computer, which is highly complex but organizationally dissimilar from organic life, compare to a rabbit? These questions do not seem to have good answers.

A further problem is that similar entities we would intuitively put on the same level may have quite different constitutive levels. Take, for example, the organism. Some organisms are single cells, like any prokaryote. Others are multicellular individuals like the rabbit. For some plant species, the organism is a distinct individual, like a single tree. However, other species are able to reproduce clonally via their roots, such that, for example, an entire stand of aspen can constitute a genetically identical and physically continuous entity. One unusual organism, the plasmodial slime mold, is macroscopic and multinucleate but not composed of distinct cells—that is, it lacks internal cellular membranes.

So, not only is it sometimes unclear what counts as an organism, but different things that all clearly count as organisms—humans, bacteria, and slime molds—have radically different decompositions. How could we decide whether, say, a bacterium is on the same level as my whole body or as one of my cells? I think, again, the answer is simply that many physical items, on account of different internal structure, are not directly comparable with respect to levels.

Having reflected on some problems similar to the ones given above as well as some in addition, Kim concludes that preserving a meaningful view of levels requires a localized and top-down approach: "we first pick a nomic kind of interest to us and go [i.e., look downward] from there, rather than start with a comprehensive levels ontology and then try to locate each object, or kind…" (2002: 20). That is, contrary to historical proposals, it looks hopeless to

universally stratify the physical world and then see where everything fits into an overall hierarchy. Nevertheless, if we select any single entity, we can still see how it relies on the lower-level entities that make up its structure. In this way we can retain hierarchical stratification on a case-by-case basis.

Another issue for a historical levels model has to do with the messiness of compositional stratification within objects. It has been pointed out by very many authors—including Kim (2002) and Potochnik and McGill (2012); Beckner (1974) also gives a helpful treatment of hierarchies—that most macroscopic objects are not, in Beckner's terminology, perfect hierarchies. A perfect hierarchy is one in which objects at level L are composed exhaustively of objects at level L-1. Contrary to Oppenheim and Putnam (1958), this fails to hold for biological objects. For example, a human body has no full decomposition into organ systems, but rather is composed of organ systems plus various fluids. A cell is similarly best seen as being composed of a mixture of objects we would intuitively place on different levels, from free ions to large organelles.

However, it is not clear that an idealized perfect hierarchy is a necessary feature of a levels model. It only seems to be necessary (assuming physicalism) that every object have a complete decomposition into objects *some* level below, unless the item in question is already basal. Furthermore, it remains intuitively true that most biological objects are roughly hierarchical, such that much of an animal's behavior can be explained in terms of the coordinated behavior of organ systems, and organ systems are largely conglomerates of cells, and cellular behavior is driven by organelles, and so on. Based on these observations, we might think that a sufficiently qualified model of levels will retain empirical usefulness.

The above two fixes—a move towards localization, and the allowance that objects are not perfect hierarchies—do bring the model better in line with empirical science, and should be uncontroversial. But so far the only positive claim we have made is that physical objects are composed of smaller parts; this is hardly a metaphysically interesting layered model. There are, however, some stronger further assumptions which are made in contemporary applications of the model. These assumptions are what potentially make the model interesting and useful when applied to specific metaphysical problems, such as the levels of causation; I will argue, however, that these assumptions also make the model empirically inadequate. I will illustrate these features of the model by considering the case of emergent properties.

Emergent Properties

Emergence is a good case study because it is a classic position on certain interlevel relationships. The view in philosophy is a form of strong causal antireductionism. It is sometimes thought to be a mere philosophical working out of a concept borrowed from the biological sciences. In fact, however, I will argue that the way emergence is typically defined in philosophy makes it inconsistent with how biologists understand emergent features of biological systems. Metaphysical definitions of emergent properties have

construed them as intrinsic and static features, whereas emergence in biological applications pertains to dynamic, extrinsic features of systems. I will later attempt to show how these distinct conceptions have their origins in the way philosophers have utilized the layered model.

Emergent properties are typically defined in philosophy along the following lines. For my purposes here, I will give only a rough and partial characterization. Emergent properties are: *mereologically supervenient* features of systems which are *qualitatively novel* with respect to lower-level features and are *not causally reducible* to the features of the system's lower-level parts alone or in conjunction (see O'Connor 1994; McLaughlin 1997; Kim 1999, 2006; Francescotti 2007). I will define each italicized part of the definition sketch in what follows.

First, the phrase "qualitatively novel" is intentionally vague. It is merely intended to capture the intuitive idea behind any emergence view, which is the fact that large and complex enough arrangements of matter can begin to exhibit interestingly new kinds of behavior.

Second, the causal irreducibility clause is typically thought by philosophers to require that emergent phenomena cannot be given a mechanism or functional explanation on any lower level—that is, the emergent feature does novel causal work not done by the system's micro-parts alone or in conjunction. It is popularly held that this sort of novel causal power must consist in reflexive downward causation, that is, causal influence of high-level features of the system on the system's parts.

Biologists, by contrast, tend to assume that all the processes they study can be given mechanisms, that is, processes can be described in terms of interactions among smaller-scale objects which compose relevant systems. Nevertheless, they also think select features count as emergent. This seems puzzling at first glance, but I will try to show below that downward causation only makes sense within the layered model. Biologists have no need for downward causation because they are utilizing an importantly different model of system structure.

Finally, emergent properties are thought to mereologically supervene, which is to say, they supervene on properties of the system's ("lower level" or microphysical) parts, taken in conjunction. Roughly, supervenience holds between two classes of properties when any object intrinsically alike in terms of one class (here, features of the object's lower-level composition) must also be alike in terms of the other class of properties (here, higher-level or emergent features). That is, anything with the same physical composition will have the same emergent properties.

This statement seems innocuous, insofar as the supervenience of all features of the world on the microphysical is an assumption of physicalism. However, the explicit appeal to mereology imports the following assumption: that emergent features of systems supervene on just the features of their own parts, that is, without considering any features of the environment. This means emergent properties, on the standard metaphysical definition, can only be intrinsic features of decontextualized objects. I will argue later that this is a bad

restriction since biologists understand emergent features to be partially dependent on the environment.

In addition, the supervenience clause is typically (at least implicitly) taken to put the emergent feature and its lower-level subvening base in temporal lockstep. That is, systems compositionally identical *at a time* must be the same with respect to their emergent properties (if they have any of those). For example, here is a statement by Kim from a discussion about emergence; notice the temporal claims:

> 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… (2006: 557; emphasis in original).

Kim is here in the process of making a specific argument about mental causation (M* stands for a mental property, though it could be read as any other emergent property), but mind is not my immediate concern here. Instead, notice his assumption that emergent features are fixed within temporal slices, absent information about the preceding or following states of the system.

To summarize, what this emergence case illustrates is that when it comes to particular objects or systems, some philosophers have tended to model system behavior in terms of causal relationships among temporal sections through the (local) compositional hierarchy of levels. It is usually reported, or clearly assumed, that mereological supervenience holds at those temporal instants, such that any system compositionally alike at an instant of time will also be alike in terms of salient systemic features.

Although I have only discussed one example for the sake of space, I do not think I have cherry-picked an unusual case. Emergence is just one of many related discussions about reduction and the "levels" of causation (including downward causation), and the mentioned assumptions—that the causal structure of physical systems is adequately captured within a framework of composition and temporal slices—looks to me to be made throughout this literature.

Having summarized what I take to be some important features of the received layered model in philosophy, I will now transition to my argument that this model inadequately meshes with the structure of emergence as seen by biologists. I will first make a note on how biologists use the term "level," which I think is unrelated to how philosophers use the term. Then I will discuss how physiologists understand emergent properties, and why these real cases of emergence fit poorly into the layered model and thus clash with the metaphysical conception of emergence discussed above.

"Levels" in Biology

In colloquial or scientific usage, "level" can refer to all sorts of things which are only vaguely related to each other and to philosophical usage. The term often pertains to sequences of inclusiveness, abstractedness, or theoretical broadness, in application to descriptions or explanations of system behavior. Alternatively, it can apply to series of larger-scale objects or processes, with or without invoking composition.

Here are examples of the use of "levels" in science. (A) Geneticists often speak of explanation on the level of genes versus phenotype. The phenotype consists of gene products, so moving between these two modes of description is not really moving up-down in the sense of composition, but rather going back and forth between two causally interconnected networks. (B) We might discuss the description of an entity's behavior at the level of physics (i.e., in terms of basic physical magnitudes) versus at the level of biology (e.g., in terms of behaviorally salient facts). Here we are considering one and the same entity under the conceptual schemes of different sciences. (C) We might compare a model that predicts system behavior at the level of a whole population versus at the level of individual interactions. In this case our different models vary in granularity. This case approaches the layered-model sense of "level," except that differences in the temporal scale of interactions may be built into the two models. More examples of scale-related issues, and why they problematize the layered model, will be considered in what follows. (D) Ecologists often talk about the biological hierarchy moving from spatially smaller to larger things: atoms all the way up to the biosphere. It bears noting, though, that it is an oversimplification to consider each level that of a larger kind of object. Populations, for instance, are arguably not objects. So, these levels are of explanatorily salient kinds, not of metaphysically comparable sorts of physical entities. Alternatively, (E) ecologists might use "level" almost synonymously with "scale" in application not to different sorts of entities but to relevant processes within a single entity. For example, an ecosystem process that occurs slowly and over a greater area (e.g. succession) might be called higher-level with respect to a faster, more localized process (e.g. nutrient flow).

I think the above should make clear how inconsistent levels talk tends to be. This is alright for scientists, who use the notion of levels only as a metaphor for the way we can describe or resolve entities and processes on different scales, from different perspectives, or with differing abstractedness. It should go without saying that this sort of un-clarity is problematic when using a layered model to frame philosophical discussions. I want to stress that, contrary to what is apparently sometimes assumed, the philosophical levels model is not contained within or clearly derivable from scientific usage of the term. Scientists do not use "levels" with any sort of consistency, let alone with any particular ontological structure of the world in mind.

It might be worth mentioning here why I think we should care about consistency with scientific concepts when considering very abstract metaphysical issues. I think that where scientists and philosophers are both interested in the same subjects—like the structure of complex biological systems—we might aim for more than the mere absence of factual inconsistency

between philosophical and scientific views. We might further ask how well the philosophy coheres with the scientific approach. I think it constitutes reason to search for a better metaphysical framework if that framework fails to sit together well with the assumptions scientists make to conduct their empirical work—even if the metaphysical thesis is abstract enough to guarantee it will not directly contradict any factual claim made by a scientist. To be upfront, my reason for taking this perspective is some optimism that both science and philosophy have the potential to converge on the truth about the world. Defending this optimism is well beyond the scope of this essay; anyone who prefers a more antirealist view about scientific theory will not likely find my argument here very compelling.

Having set aside the above issue, I will turn now to cases of emergence in physiology.

Emergence in Physiology

The Heartbeat

Denis Noble is a systems physiologist who, unlike most biologists, has written explicitly about causation and levels. He is a convenient source for my comparative purposes here, since he is writing from the perspective of a working scientist and not a philosopher. I do not agree with everything he has written in his (2008) or elsewhere, but he makes some observations that will be helpful here.

Noble observes that molecules in a biological system often behave very differently from those in a non-biological system, but only (we tend to assume) because they are parts of systems that are organized differently to begin with, not because of any intrinsic difference—that is, the basic chemistry of the molecules remains the same.

Intricate feedback networks, which result from the initial physical organization of many biological systems, may be particularly important drivers of such novel behaviors of system parts. In order to accurately model the functioning of such a system, we need to identify "the level at which such networks are integrated" (3012). What Noble means by this, I think, is that it is necessary to figure out what components of the system are minimally necessary to observe the behavior in question, which is produced by feedback among those components. (Note that this is another example of "level" being used in a descriptive but somewhat vague sense.) When we know which parts participate in the behavior, it is then possible to model how those parts must be arranged and interact with each other to exhibit the function observed in nature. No such function will appear if we look at parts in isolation or consider too few parts at a time. Note too that the relevant organization for a given behavior may occupy various scales, i.e., the behavior may be exhibited at the scale of a cell or may be invisible until an entire organ is considered. This is a contingent, empirical matter, not a matter that can be settled by considering intrinsic features of isolated microphysical parts.

To illustrate the above claims, Noble discusses how electrochemical oscillation within the heart is produced. There is no single component of the

heart that oscillates in isolation. (He notes that in mathematically modeling the beat of the heart, no individual component of the equations includes an oscillator like a sine wave.) However, when a number of ions and protein channels are put together within a network of cells, rhythmic oscillation emerges within the system as a result of feedback between the electric potential of the cells and the channels that allow the movement of ions.

There are several points I want to make about biological emergence. I will address what biological emergence has to say about downward causation here, and make further points in the next section.

The heart's oscillation is considered emergent because it is a feature that only occurs at the scale of the heart: it is "novel" with respect to the constituent molecular parts of the heart considered in isolation. However, this claim does not entail any strong metaphysical antireductionism.

Metaphysically strong emergence, as I mentioned previously, is often thought to require reflexive downward causation, i.e., causal influence of features of the whole system on the system's micro-parts. Downward causation has been posited by philosophers as a way to save the causal relevance of macro-properties from reductive projects. Such moves are motivated by the observation that construing systems in terms of a static hierarchy of levels threatens to make the higher-level properties causally superfluous. However, rather than positing strong downward causal powers—which is very controversial from both an empirical and a metaphysical standpoint[1]—we might do better to reject the entire notion that macro-properties are merely higher-level integrations of a system's micro-state.

The salient macro-micro relationships among biological properties are not really interlevel relationships in the standard sense, but rather pertain to larger and smaller scales of description.[2] Consider the case of the heart's oscillation. What Noble calls the "level" at which oscillation is observed is actually the way in which participating cellular components (cell membrane proteins, ions) must be arranged to allow oscillation to occur. In philosophical parlance, this oscillation is an organ-level phenomenon. But the supposed levels-based ontological distinction between the heart qua organ and the conglomerate of "lower-level" molecules that compose the heart is irrelevant to (and even confuses) the scientific question. Again, the scientific question pertains to the *arrangement* and *spatial scale* on which the molecular system

---

[1] For critical discussions of downward causation, see, e.g., Kim (2000), Robinson (2005), Craver and Bechtel (2007). I am sympathetic to the view that the notion of reflexive downward causation may be incoherent; however, I will not go into that issue here for the sake of space.

[2] When I talk about "describing" or "observing" systems at certain scales, I do not mean to suggest that spatiotemporal scales or physical features of systems at those scales are merely a matter of language or human observation. Rather, certain physical properties, states, and processes only occur over sufficiently large scales; so, if a scientist is interested in explaining one such feature, she must study the system at the appropriate scale.

will oscillate. Nothing *over and above* cellular components (in the sense of an extra metaphysical level) is needed for oscillation; the process can be fully causally described in terms of proteins and ions interacting. The key for biologists is that these interactions only take place when we observe the component parts over an appropriate scale. A lone ion cannot participate in oscillation; it can only do so within the larger environmental context of the organ of which it is a component part.

Here is why, more generally, distinguishing among scales in this manner cannot correspond to any philosophical notion of levels. First, there is no guaranteed supervenience relationship between states of the same system described at different spatial scales, e.g. the state of an individual ion versus the whole heart's phase in the oscillation process. Additionally, the process of oscillation occurs over a greater timescale than behaviors of individual molecules, which disrupts the neat lockstep causation posited by an idealized layered model.

Since downward causation has been developed precisely as an *interlevel* relationship, it fails to apply naturally to biological systems (or at least to the one we have looked at so far), whose dynamics are better understood as dependent on the *scale* of description. Further, the fact that the heart's oscillation can only occur on a certain minimal scale might save the process from appearing merely redundant with behaviors of individual molecules. The motivation for positing downward causation dissolves when we appreciate that novel behaviors track the scale at which we describe a system, not the purported ontological layer it occupies.

The Circadian Rhythm

Here is a second case of emergence from animal physiology. In our brains we have a region known as the SCN, which is a large organized group of neurons responsible for our circadian rhythm. The cells on their own will oscillate, but they tend to deviate from the target 24-hour cycle. When organized together in the SCN, however, they jointly produce an accurate, highly robust 24-hour rhythm. Furthermore, because of how the network of neurons is organized, some sort of oscillation can occur even if the protein components of the cells responsible for intrinsic rhythmicity are removed. This is from a review paper on circadian rhythms: "When the individual [neural] cells are no longer rhythmic [due to gene knockout], the coupling pathways within the SCN network can propagate stochastic rhythms…. Thus, … rhythmicity can arise as an emergent property of the network in the absence of the component pacemaker or oscillator cells" (Mohawk et al. 2012: 449). This is a purely scientific paper with no discernable metaphysical agenda. What they report is that a complex behavior of a system can arise due to how its parts are arranged, where those parts in isolation would not necessarily exhibit that sort of behavior, and where the mechanism of the behavior involves the system's organizational features. This is exactly what we saw in the first example as well.

One lesson of this particular case, which builds on the point I made above, is that it is a matter of empirical fact that an emergent feature occurs at

the scale of the organ. Removing the cellular oscillators might have stopped rhythmicity entirely. This is why the authors report experimental results showing that rhythmicity of the SCN can in fact be emergent. Emergence in this sense thus has little to do with a *level* of organization and everything to do with the *particular* organization of a given system. Emergent behavior pops up inconsistently; gross features of an entity's composition (e.g., the fact that something is a brain region made up of neurons) does not determine whether any of its causally salient behaviors are emergent in the empirical sense. This observation again precludes us from suggesting that ontological levels might be salvaged by replacing layers with a robust concept of scale. Biological processes certainly are scale-dependent, but the specific process or emergent feature is what determines the relevant scale of description, not vice versa. The same object might well contain emergent features on different scales of description.

There is a further issue for the strong, levels-based emergence concept that arises with this example. The SCN, I take it, requires input from the visual system about when it is daytime in order to sync our circadian rhythm with the earth's daylight period. An SCN network removed from an organism and being sustained in a lab would therefore (as far as I understand) produce some sort of oscillation—perhaps even a very consistent one—but not a proper *circadian* rhythm, since it would no longer receive any visual input. The same goes if we take a whole mammal and put it in an environmental chamber with no light. More drastically, rhythmicity would stop entirely if we put a mammal in a freezer with a simulated 24h light cycle but with the air temperature kept at -30°C. So, it bears noting, the systemic property of circadian rhythmicity does not supervene on features of the parts of the SCN in conjunction, even after we stipulate the neurons are properly arranged. It is not a purely intrinsically fixed feature of a system of given microphysical organization. Rather, circadian rhythmicity occurs only if there is (1) a properly arranged network of neurons which is (2) imbedded within a reasonably typical brain and hooked up to various sensory inputs while (3) the brain exists within a reasonably well-functioning body which is living within a narrow range of *further* environmental conditions.

One might reasonably point out about the freezer case that a certain temperature of the SCN itself—not of the environment—is what is necessary for physiological function. Importantly, however, physiological processes always take place over some temporal interval. While the *capacity* to oscillate under certain conditions may be a temporally instantaneous feature of a system, the oscillations themselves are not. And when we consider the diachronic causal story, it is clear that the system's having all of the intrinsic (structural) features necessary for oscillation at some time t_1 does not guarantee it will have those same features at t_2. To put it crudely, one could stick a mammal in the deep freeze at any time. And if something is not oscillating from t_1 to t_2, then it is not oscillating at all, since oscillation requires a time interval. Since an animal can maintain its body temperature only under a limited range of environmental

temperatures, its external environment must remain within a favorable temperature range over a time span for a circadian rhythm to ever occur.

Because of the causal relevance of factors like external temperature, it is always necessary to consider the environment when fully explaining biological processes. This is true even if all of the features that are intuitively involved in generating the process are features intrinsic to the system, like internal temperature and the arrangement of molecular components. Since the layered model considers only system composition and instantaneous relationships, it assumes the internal structure of the system at t_1 must be causally sufficient for the state of the system at t_2. But in fact this is probably always false, since the environment can always causally influence the system when we take a diachronic perspective, that is, when we consider the behavior of the system over time intervals appropriate to a given feature. The instantaneous micro-configuration of an animal does not guarantee that it is ever in a state of circadian oscillation, even if its state at the given time is appropriate for or consistent with the production of rhythmicity. On the other hand, saying that it is exhibiting circadian rhythmicity at some time guarantees that it is in that state at some other time as well. The issue is that the former description neglects the appropriate temporal scale for the process in question.

Very many causally important features of biological systems are diachronic processes: cellular respiration, photosynthesis, circadian or monthly or annual rhythms, circulation and digestion and metabolic reactions, ecosystem succession, nutrient cycling, migration and colonization and gene flow, genetic drift and selection, cell division and reproduction. So, a problem for the layered model is as follows. If we take Kim's suggestion of looking downward from systemic properties (i.e., to the micro-structure of the system), we might retain a rough hierarchy of local compositional levels. However, in doing so we neglect causally salient parts of the environment. The structure of a complex system is nothing more than a *part* of an explanation of its behaviors, particularly those novel behaviors which we might want to call emergent.

If we do wish to include the environment in our consideration then we will lose anything like a consistent hierarchy of layers, since features of the environment (especially in, say, a real ecosystem) will almost certainly not be commensurable with the system under consideration in terms of levels of composition. This goes back to the problem with comparing rocks and rabbits discussed earlier in the paper. Rocks and rabbits—as well as more nebulous entities like stands of trees and the atmosphere—all participate in ecosystem processes. But there is no good way to decide whether these entities are on the same or different metaphysical levels. This being the case, the levels model can only muddy discussions of causal pathways among different ecosystem components.

Moreover, if we consider a system in a series of instantaneous temporal slices, we may well lose emergent processes entirely.

There is often a positive correlation between the temporal and spatial scales appropriate for observing a process, though this correspondence only holds very neatly for generalizations; for specific cases, appropriate spatial and

temporal scales might come apart. This has been stressed by Potochnik and McGill (2012), who note that it is a matter of your research question and of empirical fact what scales are important as well as how finely one must resolve differences in scale; it is not just a matter of how large component objects are. For example, they argue, the difference in size between a squirrel and a tree whose seeds it eats is important for their co-evolutionary dynamics, but it makes no difference to the rate of their mutual northward movements in response to climate change.

But, consider the fact that it is *often* the case that system processes occupy greater timescales than processes involving their microphysical constituents, a point I made above in discussing the heartbeat. This leads to the following problem, diagnosed by Sandra Mitchell in a paper in which she criticizes the way some philosophers have treated biological emergence:

> If we take a snapshot view of the higher and lower levels, then the dynamics of *how* the higher level is constituted and stabilized is lost. Contemporary sciences show us that there are processes, often involving negative and positive feedback or self-organization, that are responsible for generating higher-level stable properties, and these processes are not captured by a static mapping (2012: 177).[3]

That is, if we take a time slice of a system, we will capture its microphysical organization, but we will be blind to the stable system processes—like all of the biological processes I listed just above—which tend to occur over larger temporal intervals. Moreover, we will miss *how* these processes occur, which is the scientifically interesting and predictively useful issue. That is, interesting system processes may involve complex dynamics (like feedback) across longer time intervals, which are excluded from consideration by the layered model. Because of the complexity of biological systems, both the spatial and the temporal scales at which we will see stable, explanatorily useful processes are context-specific facts which must be determined empirically; one cannot make such discoveries by thinking about static part-whole relationships from an armchair.

---

[3] Since I have quoted her out of context, I should clarify what Mitchell is arguing in her paper. When she describes "static mapping," Mitchell has in mind specifically the metaphysical projects that attempt to functionally map and thus reduce high-level properties to low-level properties. In many ways, Mitchell's strategy is similar to mine: she accuses a metaphysical project (functional reduction) of using a bad framework that misses the scientific phenomena. However, as far as I understand, Mitchell disputes the claim that purportedly emergent properties can be functionally reduced by disputing the way some metaphysicians (Jaegwon Kim in particular) have understood interlevel reduction to work. So, whereas Mitchell attempts to show that high level emergent properties do not reduce in a proposed manner, I am attempting to argue that emergent (and other systemic) properties are not "high level" in the first place because any metaphysically interesting notion of levels lacks grounding in empirical science.

Forcing biological cases of emergence into the layered framework is thus both inconsistent with the claimed metaphysics of emergence and with the way biologists understand emergent properties. Emergence in biology can only be understood in terms of *diachronic processes on an appropriate spatial scale involving a specified structure within an appropriate environment.* Such features are not nicely captured by the layered model, which forces systems into temporal slices rather than accommodating process-specific changes in the proper spatiotemporal scale of description; and where micro-macro relationships are explained solely in terms of static system composition. In fact, given how badly it meshes with a scientific conception of biological systems, one might want to question the widespread assumption that the layered model is neutral regarding questions about reductionism.[4]

In closing this section, I will address the potential objection that biologists and philosophers simply mean different things when they use the (somewhat unclear) term "emergence." I think it would be reasonable to wonder, in response, what is the use of a conception of emergence that has no application to things which scientists take to exhibit emergence.[5] Regardless, my goal is not to engage in a semantic dispute regarding what should count as emergent but rather to call into question the layered framework of physical systems. Recall my argument above that biologists do not recognize anything resembling an ontological layers framework, even if they use the term "level" frequently. This being the case, and assuming that biologists are as good as anyone to assess the structure of complex physical systems, the onus is on philosophers who wish to utilize a strict levels framework. Insofar as this framework is in no way derivable from scientific concepts, it is up to philosophers to demonstrate that the framework (1) is consistent with the science and (2) accurately and usefully represents real physical systems. I have argued that the framework, applied to one important class of examples, is

[4] Among other things, John Dupré's *The Disorder of Things* (1993) can be read as an extended argument to the effect that trying to construe scientific explanations in terms of a hierarchy of levels is inherently reductive. Here I mean reductive in the negative sense (oversimplified, misleading) rather than in a positive sense (parsimonious, explanatory). Part of the problem, for him, is that the same entities must be abstracted in different ways in order to explain different processes. Although he does not put it exactly like this, in part he seems to be getting at the issue that many explanatory systemic properties are extrinsic properties, so various salient properties of one and the same organism will have very different supervenience bases. This makes it misleading to construe the organism in terms of a single hierarchy of levels, let alone with any straightforward determinative relationship running from its microphysical parts to its higher-level properties. On this perspective, to assume that there is such a relationship comes dangerously close to begging the question whether the organism can be reduced to its microphysical composition.

[5] Compare Sandra Mitchell: "our philosophical understanding of concepts (like emergence) should track not just logical consistency, but also empirical adequacy" (2012: 181).

inconsistent with scientific concepts and fails to accurately represent the structure of complex biological systems.

Conclusion

The layered model portrays physical objects as stratified into a hierarchy of layers. On this model, the lower-level objects that compose higher-level systems are the most important determinants of the system's properties and behavior. Furthermore, the system's micro-properties are supposed to fix its macro-properties on very small (instantaneous) temporal scales, which allows for clean mapping between micro- and macro-descriptions.

Above I have argued that empirical cases of emergence in physiology, which are supposed to be a class of "high-level" properties, are badly misrepresented when forced into a levels framework. I lack the space in this essay to start at a positive argument that *none* of biology can be understood in terms of levels. Still, since the levels model is supposed to be a general model applicable to the whole physical world, one would expect it to apply to organisms' oscillators just as well as to anything else. This is especially true since it has been almost universally assumed by philosophers that the question whether there are emergent properties just is the question whether purported examples of emergence are subject to causal reduction to lower-level items. If my argument in this paper is right, then the latter statement of the question is not coherent. Emergent properties are not higher level than any other class of physical property, assuming that we expect our understanding of "levels" to be both metaphysically interesting and consistent with empirical fact.

This is not to say that the layered model is internally implausible, lacks any good application, or can be conclusively demonstrated false via empirical investigation. Rather, taking a broad view of relevant issues, the model does not cohere well with how biologists conceptualize the systems they study. This suggests that the philosophical discussion of biological systems—including discussions about causal reduction, among the many other issues I mentioned above in the introduction—might be improved in sophistication and in empirical adequacy by a departure from the layered model.

Works Cited

Beckner, Morton. "Reduction, Hierarchies, and Organicism." *Studies in the Philosophy of Biology*. Eds. F. J. Ayala, and T. G. Dobzhansky. Berkeley: University of California Press, 1974. 163-176.

Craver, Carl F., and William Bechtel. "Top-Down Causation without Top-Down Causes." *Biology and Philosophy* 22 (2007): 547-563.

Dupré , John. *The Disorder of Things*. Cambridge: Harvard University Press, 1993.

Francescotti, Robert M. "Emergence." *Erkenntnis* 67.1 (2007): 47-63.

Kim, Jaegwon. "Making Sense of Emergence." *Philosophical Studies* 91.1 (1999): 3-36.

---. "Making Sense of Downward Causation." *Downward Causation*. Eds. P. B. Andersen, C. Emmeche, N. O. Finnemann, and P. V. Christiansen. Aarhus: University of Aarhus Press, 2000. 305-321.

---. "The Layered Model: Metaphysical Considerations." *Philosophical Explorations* 5.1 (2002): 2-20.

---. "Emergence: Core Ideas and Issues." *Synthese* 151.3 (2006): 547-559.

McLaughlin, Brian P. "Emergence and Supervenience." *Intellectica* 25 (1997): 25-43.

Mitchell, Sandra D. "Emergence: Logical, Functional, and Dynamical." *Synthese* 185 (2012): 171-186.

Mohawk, Jennifer A., Carla B. Green, and Joseph S. Takahashi. "Central and Peripheral Circadian Clocks in Mammals." *Annual Review of Neuroscience* 35 (2012): 445-462.

Morgan, C. Lloyd. *Emergent Evolution*. London: Williams and Norgate, 1923.

Noble, Denis. "Genes and Causation." *Philosophical Transactions of the Royal Society A* 366 (2008): 3001-3015.

O'Connor, Timothy. "Emergent Properties." *American Philosophical Quarterly* 31 (1994): 91-104.

Oppenheim, Paul, and Hilary Putnam. "Unity of Science as a Working Hypothesis." *Minnesota Studies in the Philosophy of Science* 2 (1958): 3-16.

Potochnik, Angela, and Brian McGill. "The Limitations of Hierarchical Organization." *Philosophy of Science* 79 (2012): 120-140.

Robinson, William S. "Zooming in on Downward Causation." *Biology and Philosophy* 20 (2005): 117-136.

An Interview with Christine Korsgaard

May 2015
Harvard University

Each year The Dualist *includes an interview with a contemporary philosopher chosen by the staff. This year we are very pleased to have* Christine Korsgaard *answer questions posed by the* The Dualist *and the Stanford Philosophy Department.*

The Dualist:
How did you discover philosophy, and what made you want to be a philosophy professor?

Christine Korsgaard:
I am a philosopher by nature. What I mean is that I have always been interested in working out what I think about the "big questions." In my family we sometimes had political arguments at dinner – my grandfather, who lived with us when I was a child, was more conservative than the rest of the family, and this got us into arguments. It made me wonder how there could be objective answers to ethical and political questions. What exactly makes a view about ethics true or false? Of course I also wondered about the things many young people wonder about – whether there is a god, what we should aim for in life, whether various actions were right or wrong. I even kept notebooks in which I would write down my thoughts about these issues and argue with myself about them. But I am a first-generation college student and this was all long before I knew there was such a thing as philosophy. When I was in high school I developed an ambition to educate myself, and bought a set of "great books." This included some Plato and some Nietzsche, and that's how I learned that the sort of thing I did was a regular subject with a name. I was home.

But I didn't decide then to be a philosophy professor. I don't think I ever exactly decided that. I was extremely shy when I was young, and could not really imagine myself as a teacher, standing up in front of a class and talking to a large number of people. It just seemed impossible. I didn't even go to college right away, because although I was "bookish," my shyness made school difficult to enjoy. After high school, I took a secretarial course so that I would have a job skill. After a brief stint working as a secretary, and trying to teach myself philosophy, I realized that it was too hard, and I needed teachers. So I went to college after all. After college, I wanted to keep studying philosophy, so I went

to graduate school. In those days, the job market was bad, and when you applied to graduate school, the schools that accepted you also sent you a letter that basically said, "we can't get you a job: don't come." But I'd always been able to find work as a secretary – I continued to do office work during the summers while in college – so I wasn't worried about whether I could find work. I went to graduate school, and as part of my training I did some teaching, and found out I could handle it. And of course, I still wanted to go on doing philosophy. So from there it was natural to become a professor.

The Dualist:
How do you think academic philosophy contributes to the wider culture?

Christine Korsgaard:
Philosophy represents an ideal – the ideal of thinking a problem or a question all the way through. That means that you have an argument for your solution to the problem or your answer to the question; that you know how to defend the premises of that argument; that you have grasped the full implications of the solution or the answer, and that you can show that these implications are acceptable. It means that all of the connections of the issue to everything else have been thoroughly explored. In other words, it's the ideal of good thinking, nothing less. If it takes a thousand years to establish the answer, then that's what philosophy is going to do. It's an impossible ideal even for philosophers to meet, but philosophers at least have the time and the intellectual resources (in the great philosophical works of the past) to try. In everyday life, especially where moral and political issues are concerned, we have to settle for much less – after all, you can't take a thousand years thinking a question through if you need an answer in time to vote in the next election or decide what to do in an emergency. But of course it is important that people think as well as they can, since thinking is how human beings do what we do, and therefore it is important that we all have some idea what really good thinking would look like. It is especially important that people should see how hard it is, not so they will despair, but so that they won't be overconfident and forget to question their own views. For those reasons alone, I think that everyone should study philosophy.

Philosophy also represents humanity's main resource and occasion for rigorous thinking about moral and political issues, normative issues that cannot be settled by the empirical sciences.

Perhaps most importantly, philosophy is the discipline of self-knowledge. It is in works of philosophy that we find conceptions of what we are and what we are doing when we lead human lives and live together in human societies. If you have a conception of what you are doing, then you can do it better. For example, I think that people who live in democratic republics can be better citizens if they study social contract theory. They'll be better citizens because they will better understand what they are doing when they vote or hold office. I'd like to think that people who understand that they are in charge of their own self-constitution will constitute themselves better, too.

But perhaps what I should say is that these are some of the roles that philosophy *should* play. Two things prevent philosophy from having as much effect as it should. One is the fact that, at least in America, people don't study philosophy in secondary school. There are otherwise very well-educated people who have absolutely no conception of what goes on in philosophy, and as a result are extremely suspicious or even a little scared of it. People should start doing philosophy early, while they are intellectually fearless and playful and willing to explore.

The other problem is the appalling quality of most philosophical prose. People who aren't professional philosophers can't just read philosophy for pleasure and edification, the way they can read novels or history or art or film criticism, because of the horrible way that most philosophers write. Contemporary philosophical writing is needlessly technical and clubby, explicitly written for other professional philosophers and them alone. It is aimed more at fending off criticism than at conveying understanding or creating interest. Most philosophers pay little attention to style, and formalize things that could just as well be said in words because they think it makes them look smart. I think many philosophers prefer to believe that they are just writing for a small audience that shares an eccentric taste for intellectual puzzles, because they are afraid to take on the responsibility of saying something that people in general might find significant and meaningful.

When you put those two facts together - I mean that people don't study philosophy in secondary school and that it is not written with a wider audience in view (or really, with any audience in view) you get a very bad result – people who first approach philosophy later in life often find it impenetrable and unrewarding, and give up any hope that it will illuminate their lives. And that's a shame, because that's what philosophy should do.

The Dualist:
Kant thought that the Categorical Imperative was somehow implicit in everyone's common sense moral understanding. Do you think that? Why?

Christine Korsgaard:
If what Kant thinks is correct, the categorical imperative *has* to be implicit in everyone's common sense understanding, just as the principles of logic are implicit in everyone's thinking. The categorical imperative is a formal principle of practical reasoning, just as the principles of logic are formal principles of reasoning in general. For Kant, a formal principle is constitutive of the activity it governs: it's a kind of description of the activity. In other words, you aren't really *thinking* unless you are following the principles of logic – unless you are trying to avoid contradictions, make valid inferences, and so on. And you aren't really *acting* unless you try to ensure that your movements are governed by your own mind, and unless you try to be effective. That's what it is to be an agent – to have effects in the world that are determined by your own mind. Those two standards give us the categorical and hypothetical imperatives. The categorical imperative tells us to act in accordance with the principles we ourselves think

should be laws, and in that sense to be governed by our own minds. The hypothetical imperative tells to take effective means to our ends. So the effort to be autonomous, which is what the categorical imperative enjoins, is built right into the nature of action.

Apart from that, there's plenty of evidence that everyone's common sense moral understanding involves appealing to a universalizability principle. When we criticize an action by saying, "What if everyone did that?" we are suggesting is that the principle of that action cannot be universalized. When we say, "Do unto others as you would have them do unto you," or "put yourself in the other person's place," we are urging people to act in ways that are acceptable from every point of view, on considerations that everyone can regard as reasons. We are saying, "Treat others in accordance with the laws you would wish other people would follow in their actions towards you." All of these familiar sorts of moral argument can be seen as implicit appeals to the categorical imperative.

The Dualist:
Are you in any way influenced by Elizabeth Anscombe's philosophy of action? If so, how?

Christine Korsgaard:
This question surprised me. No, I'm not. Anscombe and I are both influenced by Aristotle, and I suppose there are some similarities in our views that derive from that. But there are also differences that come from my allegiance to Kantian ideas.

What got me interested in the philosophy of action was trying to answer the question what makes the principle of instrumental reason normative. Kant seems to have been the only philosopher who addresses this – everyone else seems to think this is a normative principle you get for free, without needing to explain it. According to Kant's argument the instrumental principle is a constitutive principle of action, and that seems right to me. After I worked out what that meant, I started thinking about whether we could explain the normativity of the categorical imperative that way too. Obviously, if you are going to make claims about certain principles being built right into the nature of action, you have to know what action is, so that's why I started working on the question.

Anscombe certainly doesn't think that the categorical imperative is a constitutive principle of action. Another difference is that Anscombe thinks you act from a knowledge of what you are doing, while I think that you act from a conception of what you are doing that isn't necessarily knowledge. She thinks that because she thinks human action partakes a little of divine creative activity as we conceive it – just thinking something makes it so. I think human action is more like human production: an attempt to impose form on matter – in this case, the form of law on our recalcitrant selves.

The Dualist:

Does moral luck play any role in the constitution of the self? Can the unification of the parts of the soul be made impossible or inordinately difficult by circumstances outside the control of an agent, or is the agent the only one who can do harm to her own soul, by her failure to unify it in action? More generally, how does the Kantian regard moral luck?

Christine Korsgaard

Luck certainly plays a role in self-constitution. Many people have very little choice in general about what they do. Rigid social forms and traditions leave little scope for choice. Society can make it hard to put two forms of practical identity together – an obvious example is the way society makes it hard for women both to be mothers and to have meaningful careers. Since I believe (as Kant did) that there is a general duty to obey the law, I believe that bad laws can make it difficult or maybe even impossible to do the right thing. Since in a democratic republic we all share the responsibility for our social institutions, we are all implicated when they are unjust. These are all circumstances that can make it harder to constitute yourself in a satisfactory way, so to find yourself in them is a kind of bad luck. Some of them also make it hard to maintain your moral integrity, and so count as a kind of bad moral luck.

Bernard Williams talked about a specific form of moral luck in which you can be justified by the result if you do something that is morally – well, let's say, dicey – and it turns out well. In a paper I wrote about Kant's attitude towards revolution, I argued that engaging in revolution could be like that. There's an obvious way in which revolution is wrong: it can't be right for individuals to engage in violence, overthrow the government, and disobey the laws just because they don't happen to agree with the laws or the regime. Yet if a regime is sufficiently unjust, as I said above, people are implicated in the injustice, may feel it is their responsibility to set it right, and may see no way to do that besides armed revolt. When laws and practices are sufficiently bad, justice is divided against itself – it both requires and forbids you to obey the law. In that kind of situation, I think you can be justified by the result. If you correct the injustice and get a substantially better system in place, then what you did was right. If you don't, you not only have disobeyed the law, but you may also have killed or injured people for no good reason.

I'm also tempted by the view that paternalistic acts can be like this. Suppose you put an addict in a treatment clinic against his will. Kant's views are strongly anti-paternalistic: he would say that this could only be justified if the person is so far gone in addiction that he is incapable of making the decision for himself. And it may just be fundamentally unclear whether that is so. We might think that if he is cured, you did the right thing, while if he is not, it just turns out to have been unwarranted interference.

The Dualist:

What should we ordinary everyday agents do about the current treatment of animals in industrial farming? As Kantians, is it our duty merely to keep our

hands clean by refraining from personal animal consumption, or should we aspire to some kind of political action?

Christine Korsgaard:
I think it is our duty to support the Animal Legal Defense fund. How's that for a straight answer?

The argument is complicated, so I'll just cheat here and refer you to my forthcoming book *Fellow Creatures* for the details, but I think that animals have natural rights against humanity collectively speaking. These include rights not to be forced to live in painful or stressful conditions and not to be experimented on in painful and invasive ways. So it is a question of justice, not just compassion, and we all have a duty to promote conditions of justice, at least by supporting those who are more actively trying to bring them about. More generally, the law is the only real solution to the problem of cruelty to animals. As long as there are no effective legal protections for animals, and animals can be owned, animals will be at the complete mercy individuals and organizations who can profit or otherwise benefit from treating them cruelly. We can't just wait for all these individuals and organizations to develop enough personal compassion to refrain. It's not acceptable. So we should all be working to get effective laws in place.

The Dualist:
In "Self-Constitution in the Ethics of Plato and Kant," you argue that the project of self-constitution generates internal standards to which our actions must conform if they are to be actions at all. But one question that a moral skeptic could ask is - do we have the freedom to not undertake the constituting project? If not, if it is a necessity that we undertake the self-constituting project, can we have different attitudes towards the project?

Christine Korsgaard:
No, I don't think we have the freedom not to constitute ourselves. We constitute ourselves through action, and we don't have the freedom not to act. Self-constitution is part of the human way of living in the same way that breathing is part of the animal way of living. You can stop, but you can't decide to stop, anyway not for long. Of course people can constitute themselves more or less reflectively, more or less consciously, more or less actively. You can be slovenly about it, although I think once you are really aware that that is what you are doing, it's harder than you might think to do that without a certain amount of either self-deception or despair. On the other hand, people can suffer from a kind of weakness of will about constituting themselves morally for the same reason that they suffer from weakness of will about constituting themselves physically. Just one more drink, just one more piece of cake, just one more sedentary day – that's not going to make you unhealthy, not just that one. But they add up, and one day you find you are an unhealthy person. It's the same thing with moral self-constitution: just one more ungenerous or cowardly or unkind action – that's not going to make you into the kind of person you

despise. But they add up, and one day you find you are a narrower, more cowardly, less interesting, or worse person than you once meant to be.

Resources for Philosophy Undergraduates

This section includes listings of journals, conferences, and contests available to undergraduates in philosophy. If you have comments, suggestions, or questions, or if you would like to be listed here in the next issue, please contact us and we will gladly accommodate your request.

JOURNALS**:**

Information given is as recent as possible, but contact the specific journal to ensure accurate information.

Aporia: Brigham Young University. Submissions are due early fall. Papers not to exceed 5,000 words. Send submissions to: Aporia, Department of Philosophy, JKHB 3196, Brigham young University, Provo, UT 84602. Visit: http://aporia.byu.edu/.

The Bertrand Russell Society Quarterly: Edinboro University of Pennsylvania. Visit: http://www.lehman.edu/deanhum/philosophy/BrSQ/

Dialogue: Phi Sigma Tau (international society for philosophy). Published twice yearly. Accepts undergraduate and graduate submissions. Contact a local chapter of Phi Sigma Tau for details or write to Thomas L. Predergast, Editor, Dialogue, Department of Philosophy, Marquette University, Milwaukee WI 53233- 2289. Visit: http://www.achsnatl.org/society.asp?society=pst

The Dualist: Stanford University. Submissions are due early 2013. 10-30 page submissions. For more information, see https://philosophy.stanford.edu/dualist-journal or contact the.dualist@gmail.com. Check website for information on submitting a paper and updates on the submission deadline.

Ephemeris: Union College. For more information, write: The Editors, Ephemeris, Department of Philosophy, Union College, Schenectady, NY 12308. Visit: http://punzel.org/ephemeris

Episteme: Denison University. Due November 14. Submissions must be at most 4,000 words. Contact: The Editor, Episteme,

Department of Philosophy, Denison University, Granville, Ohio 43023. Visit: http://journals.denison.edu/episteme/

Interlocutor: University of the South, Sewanee. Questions can be addressed to Professor James Peterman at jpeterma@sewanee.edu. Send submissions to Professor James Peterman, Philosophy Department, 735 University Avenue, Sewanee, TN 37383-1000.

Janua Sophia: Edinboro University of Pennsylvania. Submissions and inquiries sent to Janua Sophia, c/o Dr. Corbin Fowler, Philosophy Department, Edinboro University of Pennsylvania, Edinboro, PA 16444. Visit: http://www.sshe-iaprs.org/januasophia.htm

Princeton Journal of Bioethics: Princeton University. Visit http://www.princeton.edu/~pjb/

Hemlock: University of British Columbia. Visit http://philosophy.ubc.ca/community/philosphy-students-association/prolegomena or write prolegom@hotmail.com or Prolegomena, Department of Philosophy, 1866 Main Mall, Buchanan E370, University of British Columbia, Vancouver, B.C., Canada. V6T 1Z1.

Prometheus: Johns Hopkins University. Prometheus strives to promote both undergraduate education and research, and looks for submissions that originate from any scholarly field, as long as those submissions clearly demonstrate their applicability to philosophy. Visit http://prometheus-journal.com. Write prometheusjhu@hotmail.com or Prometheus, c/o Philosophy Dept., 347 Gilman Hall, Johns Hopkins University, Baltimore, MD 21218.

Stoa: Santa Barbara City College. For more information, write The Center for Philosophical Education, Santa Barbara City College, Department of Philosophy, 721 Cliff Drive, Santa Barbara, CA 93109-2394. Visit: http://www.sbcc.edu/philosphy/website/STOA.html

The Vassar College Journal of Philosophy: Vassar College. Dedicated to both quality and accessibility, it seeks to give undergraduate students from all disciplines a platform to express and discuss philosophical ideas. Questions about *The Vassar College Journal of Philosophy* can be directed to

philosophyjournal@vassar.edu. Visit: http://philosophy.vassar.edu/students/journal/

The Yale Philosophy Review: Submissions due February 14. Visit: http://www.yale.edu/ypr/submission_guidelines.htm

CONFERENCES:

There are many undergraduate conferences, so contacting the philosophy departments of a few major schools in a particular area or researching on the web can be quite effective. The conferences below are by no means an exhaustive list.

American Philosophical Association: The APA website, http://www.apaonline.org/ contains an extensive list of conferences.

Butler Undergraduate Research Conference: Butler University. The conference is held in mid-April. See http://www.butler.edu/urc/ for details.

National Undergraduate Bioethics Conference: Visit http://www.asbh.org/meetings/nuc/national-undergraduate-bio-conf.html or write bioethic@nd.edu.

Pacific University Undergraduate Philosophy Conference: Pacific University. The conference is held in early April. Visit http://www.pacificu.edu/as/philosophy/conference/index.cfm for details.

Rocky Mountain Philosophy Conference: University of Colorado at Boulder. Visit: http://www.colorado.edu/philosophy/rmpc/rmpc.html

Undergraduate Women in Philosophy Symposium: University of California at Berkeley. Email: mapwomen@gmail.com

ESSAY CONTESTS:

The essay contest listed below aims at a broad range of undergraduates, but there are many other contests open to students enrolled at specific universities or interested in particular organizations.

Elie Wiesel Essay Contest: open to undergraduate juniors/seniors with faculty sponsor. Questions focus on current ethical issues. Submissions are due in late January. The top prize is $5,000. For more information, visit:
http://www.eliewieselfoundation.org/entercontest.aspx

The Dualist would like to thank the following contributors from Stanford University:

The Philosophy Department

Special Thanks to:
R. Lanier Anderson
Nadeem Hussain
Krista Lawlor
Tamar Schapiro
Teresa Mooney
Eve Scott
Ai Tran
Erika Topete

THE DUALIST is a publication dedicated to recognizing valuable undergraduate contributions in philosophy and to providing a medium for undergraduate discourse on topics of philosophical interest. It was created by students at Stanford University in 1992 and has since featured submissions from undergraduates across North America.

If you would like to receive an issue of THE DUALIST or to submit a paper, please contact us at the address below. We prefer that submissions be formatted according to the Chicago Manual of Style guidelines.

Papers should be submitted in electronic form only.

Visit our website for submission information:
https://philosophy.stanford.edu/dualist-journal

Please email us with any inquiries:
the.dualist@gmail.com

Or write to:
The Dualist
Philosophy Department
Stanford University
Stanford, CA 94305