Sanjay Chandrasekharan. *Varieties of Supervenience.* Carleton University Cognitive Science Technical Report 2001-03 (www.carleton.ca/iis/TechReports). November, 2001. © 2001 Sanjay Chandrasekharan.

Varieties of Supervenience

Sanjay Chandrasekharan

Ph.D. candidate (Cognitive Science),
Institute for Interdisciplinary Studies,
Carleton University,
Ottawa, Canada

schandr2@chat.carleton.ca

Pre-doctoral research fellow,
Center for Adaptive Behavior and Cognition,
Max Planck Institute for Human Development,
Berlin, Germany

schandra@mpib-berlin.mpg.de

Supervenience is a fundamental concept for non-reductive physicalist theories of the mind (theories which hold that the physical level is the fundamental level of understanding, but which also hold that the mind cannot be *reduced* to the brain or any other physical level). Most computational theories of the mind belong to this category. In this paper I will outline two different kinds of supervenience relationships and the problems they face in explaining the mind-body problem.

In addition, I will argue that the more plausible version of supervenience is in conflict with multiple realizability (the thesis that the mind can be instantiated by things other than the brain), which is the philosophical basis of all Artificial Intelligence (AI) efforts. I then suggest a possible way of saving both the supervenience relationship and multiple realizability, taking inspiration from the Indian philosophical concept of autoreflexivity of awareness (*svasamvedana*).

Supervenience

Even though supervenience was introduced to the philosophy of mind by Davidson, the notion of supervenience considered here is the one laid out by Kim (1993).

Here's a non- rigorous definition of supervenience: If A supervenes on B, any change in A is a change in B, but not vice versa. Where A and B can be objects, events, or state of affairs.

Applying this to a strictly internalist story about the mind, any change in mental state A is a change in the brain state B, but all changes in brain states are not changes in mental states. Let's call this supervenience-I, for supervenience-Internalist. This is the kind of mind-body supervenience I will be arguing against.

The formal notion of supervenience, as outlined by Kim, has three versions, the weak, strong and the global. I outline only the weak and strong versions below, because Kim equates the global with the strong one.

> *Weak supervenience: A weakly supervenes on B if and only if necessarily for any property F in A, if an object x has F, then there exists a property G in B such that x has G, and if any y has G, it has F.*

> *Strong supervenience: A strongly supervenes on B just in case, necessarily, for each x and each property F in A, if x has F, then there is a property G in B such that x has G, and necessarily if any y has G, it has F.*

Multiple Realizability

The notion of multiple realizability I am using here is the garden variety one; so common that it is taken from the World Wide Web. Multiple realizability is defined as follows (from the online dictionary of Philosophy of Mind):

> *If two events can be tokens of the same mental type (e.g., they are both thoughts about Vienna), while being tokens of two different physical types (e.g., one is a pattern of neural activations in a mammal's brain, one is a certain distribution of electrical currents in a creature made out of silicon), then the mental type in question is multiply realizable.*

Less rigorously, the multiple realizability thesis says that the mind is not dependent on the brain as a physical system, and creatures made of metal, plastic or blue cheese can have a mind.

Two kinds of supervenience

The discussion of supervenience is largely done in the abstract. I would like to bring it down to a more concrete level, using examples. One reason for doing this is to show that there are different kinds of superveniences, and therefore, appeals to supervenience have to specify which of these superveniences they are appealing to. Since this requires picking and choosing between superveniences, it shows that the supervenience relationship is a norm that is applied to the mind-body problem, not a brute fact.

First, a bit of clearing of philosophical ground: I only consider the supervenience of mental content here, the issue of consciousness is not considered. Two, I don't consider the notion of supervenience of explanations, which considers one narrative (say of the mind) supervening on another narrative (say of the brain). Three, my examples are positioned against *internalist* models of the mind, models which hold that all mental states are *just* brain states. And finally, my sympathies are with an interactionist view of the mind, where mental states supervene on a brain-environment complex, instead of just the brain.

The supervenience relation, as seen in the world, can be categorised broadly into two. Let's call them Conventional Supervenience and Physical Supervenience.

1. Conventional supervenience

Roughly put, this supervenience relation needs observers — A supervenes on B by the setting up of a convention by a group of people. A good example for this kind of supervenience is the dollar bill, usually put forward as an example of the token-identity theory, which is closely associated with functionalism and multiple realizability. This is the kind of supervenience that naturally supports multiple realizability and AI.

According to token-identity theory, each mental state is identical to a particular physical state (it is nothing more than that state), but there is no identity between types of mental states and types of physical states. The value of the dollar note in your pocket could be constituted by a piece of paper (a one dollar bill) or a piece of metal (a coin). Note the following points.

- The bill and the coin belong to the same monetary type.
- The bill and the coin do not belong to the same physical type.
- Hence there is no identity of type between the monetary (being a dollar, having a value) and the physical (paper, metal).

The supervenience of software on a physical machine is similar to this. Notice that the above description does not say *how* there can be something like a monetary property, or the physical mechanism of the monetary property.

In fact, both the dollar bill and software presuppose a set of observers for whom the supervenient property (being a dollar bill and being, say, ASCI characters) makes sense. The function of the dollar bill and the function of the ASCI character set are established by convention, not by nature. Once this function is established, by a convention which says that this function supervenes on a physical object or state, the same property can be realized by different physical objects or states (by coins in the dollar case and by different kinds of machines in the ASCI case).

This kind of supervenience is not a valid one to explain the mind-body problem, because it presupposes an observer, or a set of observers. For the mind to be like a dollar bill, there has to be a convention by which people agreed on psychological states supervening on physical states[1].

<u>Objection</u>

For there to be a convention that property A supervenes on property B, there need not be an observer. The convention can be established by a meaning-giving system.

The notion of a meaning-giving system can be illustrated using a counterfeit coin. What makes a coin counterfeit is the fact that it does not originate from the government mint. And the government mint is not an observer, it is a "face-less" entity[2].

To treat this objection, let me first outline a real-life version of multiple realizability. In the early 1980s, India had a problem with producing coins. So there was a shortage of coins in circulation, and bus conductors in Delhi had a hard time giving people change. The conductors hit on the following solution: instead of returning change, they issued passengers blank tickets of smaller denominations. These tickets could be used for travel in other buses, or turned in at a local transport office to collect money.

Now comes the interesting part. Shopkeepers started accepting these tickets as currency. So you could buy cigarettes with bus tickets!

Notice that a bus ticket not issued by a conductor will be considered as counterfeit. This is an instance of multiple realizability of value — a kind of twisted version of the dollar/coin example. But the twist is interesting.

What the example shows is that the “value-carrier” (coin or bus ticket) is not dependent on a “face-less” meaning-giving system. It is dependent on a group of people agreeing to an instrument of transfer. So the observers are back, and therefore conventional supervenience is not good enough for solving the mind-body problem.

There is more to the example, though. Notice that in the bus ticket example a *full transfer* of value (content) takes place from one physical base to another.

So the question is: If the bus ticket is a bonafide case of supervenience created by a “face-less” entity, how does the transfer of content take place? How can a bus ticket suddenly become currency without its physical base changing? Projecting to the mind-body problem, this implies that my mental state can suddenly change into something else without any change whatsoever in my physical state. So supervenience-I doesn’t hold.

However, as we shall see, the conventional supervenience relation need not be thrown away, it can be used in a way to support the multiple realizability thesis. Let’s look at a more plausible supervenience candidate now.

2. Physical Supervenience

There is no convention associated with this kind of supervenience. It is a naturally occurring physical relation, something that exists independent of observers. In this relation, A supervenes on B *by virtue of B's physical properties*. However, as we will see, not all physical supervenience relations strictly follow the definition supervenience-I needs. I present two examples below that do not meet the requirements of supervenience-I. The examples are presented as analogies as well as illustrations. I encourage readers to think of the mind-body relation in analogy while considering them.

Physical Supervenience I: Consider magnetism. It is a natural supervenience relation. The property of being a magnet (the magnetic force) is supervenient on a physical substance. Notice two things, though.

1) A change in the magnetic force *need not come from* a change in the underlying physical substance. It can be caused by electric currents, by the presence of another magnet, or the change in the magnetic poles of the earth. In other words, if we consider the base B as the magnetic material, the supervenient property A can change without a change in the properties of the base.

2) The magnetic force cannot supervene on all substances. It is multiply realizable, but there are restrictions on what materials can realize it.

The same story can be told about electric charges.

Magnetism could be made to fall in line with Kim’s definition by enlarging the base property B to include everything in the environment that changes property A, including electric currents, other magnets, and the poles of the earth. But this would make supervenience a non-local relationship and not useful for the internalist.

In the case of the mind-body problem, a non-local supervenience relationship implies that mental content supervenes *physically* on the brain plus the environment. This conclusion is strengthened by the following example.

Physical Supervenience II: Consider an image reflected on a mirror. This is another naturally occurring supervenience relationship — the property of being an image supervenes on a physical substance, namely the mirror.

However, it is hard to tease out what exactly the image is supervenient on. The image is dependent on a number of factors – nature of object reflected, position of object, the position of light, the nature of the light, and the nature of the mirror. A change in any of these will change the image. Notice the following:

1) If we consider the mirror as the physical base, a change in the image is *not* a change in the underlying base.

2) If we consider the whole system of mirror, object and light as the base, then the image is supervenient on a complex of things, involving objects, properties (of being a mirror and being an object that reflects light), a fundamental force (light), and a particular configuration of the objects, properties and the force involved. This means supervenience is not a local relationship.

3) Given 2, it is an open question whether the *same* image can be multiply realized using physical supervenience, because it will require reproducing the *same physical relationship* everytime. Even if such reproduction is possible, it still needs substances that have a set of properties (see below).

<u>Objection</u>

What if the reflection on the mirror is treated as supervening on refracted light rays?

This assumes that the mirror and the object are not relevant to the image. All changes in the image are changes in refraction patterns. However, this does not help in saving the supervenience-I relation. Different objects can create the same refraction patterns, but the images would be different. So the image would change, but the physical base (the refraction pattern) wouldn't. So supervenience-I doesn't hold.

To save supervenience-I, we can assume that all refraction patterns are unique. Then a different object will create a different refraction pattern, which will result in a different image. However this solution suggests that we can't have multiple realizability, because all patterns are unique, and therefore, are not reproducible. So it looks like if we subscribe to physical supervenience, we can’t have both supervenience-I and multiple realizability.

Mental material

There is another, more important, reason for this: even though physical supervenience as described in the examples are multiply realizable to some extent, they can be realized *only by specific classes of substances*. For instance, magnetism can supervene only on *magnetic* materials and reflections only on *reflecting* materials. These classes of substances have *properties* that allow the supervenience relationship to exist.

By analogy, if you believe in both multiple realizability and physical supervenience in the case of the mind, then for AI to obtain, you have to postulate a class of materials on which the mind can supervene. In other words, you need some sort of mental material to recreate the mind. The only mental material we know of right now is the brain.

So postulating multiple realizability using physical supervenience is postulating a new *physical* property shared by some (or all) materials[3]. Would the physicists agree to this?

Objection

Information processing material can be viewed as mental material.

By this hypothesis, we have to agree on information processing as a *physical* property that can be instantiated by a certain category of materials, just like magnetism. So, for instance, silicon could instantiate the mind, just as iron does magnetism.

The problem is that the category of materials that instantiate a physical property is *identified by the capacity to possess that property*. Magnetic materials are such because they can possess magnetism, reflective material are such that they can reflect. Therefore, by this hypothesis, information processing materials should first be *identified as having a mind*.

However, since the multiple realizability hypothesis does not identify such materials by *pointing them out as possessing minds*, the notion of information processing as a physical property does not help the thesis[4].

Supervenience and Indian Philosophy

Let me recap the points I've made:

P1: The supervenience relation that naturally supports multiple realizability is one that uses conventions. However, since this supervenience needs an observer, or a set of observers, it is not useful in solving the mind-body problem.

P2: There's another supervenience relationship, what I call physical supervenience, which is a naturally occurring relation, and doesn't need conventions. But physical supervenience poses two problems: one, it is not a strictly local relationship. For this relation to apply to the mind-body problem, the environment has to be brought into the picture. In other words, the environment has to be part of the base on which mental content supervenes. Two, it needs us to postulate some sort of mental material on which the mind can supervene.

P3: Because of the second reason above, physical supervenience is in conflict with multiple realizability, and hence AI. If A supervenes on B *by virtue of physical properties*, then obviously A is tied to B's physical properties, and therefore A is not multiply realizable.

Now we can ask the question whether we can save both the supervenience relationship and multiple realizability. To begin, let us ask: what kind of model can save both?

1) Obviously, given the examples, and my sympathies, a model that includes the environment.

2) But because of P3, we need a model where mental content supervenes on physical structure not by virtue of physical properties, but by virtue of mental content itself. The model should not postulate mental material, and should allow for the agent's internal physical mechanism to be anything (or at least silicon), not just the brain.

3) This implies moving away from a strictly physical supervenience relation to a more convention-like supervenience relation, where content can move freely between

bases. However, we cannot use the conventional supervenience as in the dollar-coin case, because we need a model that gets rid of the need for an observer.

To suggest such a model, I take inspiration here from the Buddhist notion of *Swasamvedana* or autoreflexive awareness (as laid out in Matilal, 1986), the notion of *awareness events* that are self-aware[5]. According to the model laid out by Matilal, an awareness event is like a lamp/light-bulb, it has the property of illuminating itself; so an awareness event in an agent doesn't necessarily need an observer/observation module to become an awareness. Though it is not specifically laid out in the model, I assume here that an awareness event is akin to a representation, and, therefore, has a link to the environment (broadly construed, i.e. including bodily states etc.). I'm encouraged to treat it this way because of the agent-environment interaction inherent in the constructivist metaphysics advocated by Buddhism, even though I am not particularly keen on traveling all the way up the metaphysical path with the Buddhist.

Now here's why I think this model could solve the problem of supervenience:

- An awareness event A supervenes on the brain+environment physical state B, so the non-reductive physicalist model of supervenience holds -- any change in the mental state A is a change in the physical state B.

- The idea of these events being autoreflexive, or being "self-illuminating", gives them the ability to be multiply realizable, realizable on any base, because *the property of being a mental event is part of the event itself, and not the physical system.*[6]

In some ways, the model is similar to conventional supervenience; the difference is that while conventional supervenience needs an external observer and a convention to set the mental content, the autoreflexivity model, as I see it, doesn't, because the awareness and content is internal to the *event*.

In effect, the awareness event model marries physical supervenience with conventional supervenience in an elegant way. It extends physical supervenience by bringing in the environment into the supervenience relationship, but gets rid of the need for an external observer from conventional supervenience, by postulating awareness as "self-illuminating".

The awareness event model works better because it "empowers" mental content, so to speak. It considers mental content to be a bit more "powerful", having more intrinsic properties, than the passive representations used in classical AI models. Or the epiphenomenal representations postulated by eliminativist models, where the power is vested with the physical system, and the mental is a stub, an epiphenomena generated by the physical system. The awareness event model essentially moves some of the "power" to the abstract entity "event", which brings together the environment, body and the mind.

Admittedly, this is a move towards a kind of idealism, where an abstract entity "awareness event" is postulated with some internal properties. However, I consider this a plausible move -- if numbers as abstract entities can have properties, awareness events can have properties as well. More closer home, an awareness event is not a lot more idealistic than the abstract inner representations postulated by most computational models. Also, I think some sort of an abstract entity is a requirement in this case, because if the property of being a mental state is a physical property, we wouldn’t have needed the supervenience relation in the first place.

More worrying is the nature of the “self-illumination”. How can an awareness be “self-illuminating”? It looks as if the introduction of autoreflexive awareness only substitutes a problem with a mystery. However, this need not be so. An awareness event can be viewed as something similar to action-oriented representations, as suggested by Andy Clark (1997).

Clark argues against idealised, centrally stored representations of classical AI and suggests that representations are closer to the Gibsonian notion of affordance. An affordance is “an opportunity for use or interaction which some object or state of affairs presents to a certain kind of agent.” Thus, for a human a chair affords sitting, but to a woodpecker or termite it may afford something different. Thus, representations are not action-neutral, as envisioned by classical AI, but they are action-oriented. Representations are created by agents on the fly, to execute an action, by exploiting cues that exist in the environment.

This model of representation is very similar to the idea of a self-aware awareness event, as sketched above. The “self-illumination” property of an awareness-event is not a mystery in such a model, *it is part of the event's action-orientedness*[7]. An action-directed awareness event is aware of itself, within the context of the action. So, the event of my hand encountering the coffee cup is aware of itself, which is why I move to the next event, raising my hand, and start drinking from the cup. Notice that this is much weaker than the claim that an agent has full-blown awareness of all her awareness–events.

Future Work

This model is not intended as a direct translation of an Indian philosophy model into Cognitive Science. It is a model that just draws inspiration from an Indian philosophy model (mostly Buddhist). From an Indian philosophy standpoint, I'm treating a mental event here as an instance of consciousness. The relationship between consciousness and mental events is quite nebulous in Indian philosophy and fitting those debates into Cognitive Science is almost an impossible task. This model is just intended as a bridge – to connect Indian philosophy and Cognitive Science. Also to connect the largely functionalist idea of supervenience with the situated cognition (and Indian philosophy) notion of mental content being episodic -- i.e. mental content as being tied to the environment, instead of being passive, stored data, as suggested by classical AI models. Since the episodic, environment-directed, character of mental content is a recurring theme in Indian philosophy, I think it could be explored further to build bridges with situated cognition models in current Cognitive Science.

Endnotes

[1] There is of course the argument made by externalists like Millikan (1994) that the function can be "selected for" by evolution. However, in my view, "being selected for" is also a norm. The crucial difference is that the observer for whom it is a norm, namely nature, is an entity that is not considered as a subject, but is considered as being "neutral". So it is still a convention, but one apparently sanitised. I return to this point below.

[2] There is always the question of how the faceless entities come to be.

[3] Let me briefly consider Millikan's point here. Even if we grant Millikan that a representation is "selected for", she has to agree that only a certain class of materials, namely organic matter, has gone through the selection process. So I think she, too, cannot make the move to multiple realizability.

[4]Theories about the atomic structure of magnetic substances come *after* they are identified as magnetic, not before. So information processing materials should be identified as mental materials first, *before* we launch into discussions about them having an atomic structure, if any, that makes them information processors.

[5] *Swasamvedana* is a thesis held by various Indian schools, but I find the Buddhist model interesting because it attributes autorefelexivity of awareness to discrete *events*, not to a central controller. Tom Yarnall has pointed out to me that not all Buddhists accept the *Swasamvedana* thesis. Also, questions have been raised about the suitability of using Matilal as a reference here.

[6] For my purposes, I am treating self-illumination as a property intrinsic to the abstract entity event. There could be other ways to construe self-illumination, including purely physicalist approaches. Many thanks to Dr. Ram-Prasad Chakaravarthi for pointing this out to me.

[7]Another way to look at self-illumination is to consider it a case of emergence, where high-level properties are considered to "emerge" out of interactions of low-level properties. A phenomena is considered to be "emergent" if the categories used to describe it are radically different from the categories possessed by its underlying processes. A good example is temperature, which emerges out of interactions between atoms. But atoms themselves do not have the property temperature. Here, interactions with the environment is the underlying process, and self-illumination is the emergent property.

References

Block, N. (1980). Introduction: what is functionalism? Readings in philosophy of psychology. Ed. N. Block. Cambridge, MA, Harvard University Press. 1: 171-184.

Clark A. (1997). Being There: putting brain, body, and world together again, Cambridge, Mass., MIT Press.

Dictionary of Philosophy of Mind, available at http://artsci.wustl.edu/~philos/MindDict/

Kim, J. (1993) Supervenience and the mind. Selected philosophical essays. Cambridge Studies in Philosophy. Cambridge University Press.

Matilal, B.M. (1986) Perception: An essay on classical Indian theories of knowledge. Oxford, Clarendon Press.

Millikan, R. G. (1994) Biosemantics, in Mental representation, a reader, by Stich, S. and Warfield, T. A. (eds.), Blackwell.