"Physical Theory and the Complexity of Phenomenal Experience"

ABSTRACT: Arguments that subjective experience is resistant to physical theory typically require that a physical theory describe contingent historical information. This is too strong a criterion for physicalism for at least two reasons. First, this confuses the information in a theory with historically contingent information that the theory may manipulate. This is the difference between describing natural laws and natural kinds, on the one hand, and describing some specific event that is an instance of such kinds, on the other. Second, and more important, these arguments confuse the amount of information a person may retain and manipulate in a phenomenal experience with the amount that they may retain and manipulate in any mental capability representing their theoretical knowledge. These errors can be demonstrated with some basic observations from the theory of descriptive complexity, using primarily Frank Jackson's Knowledge argument as an example.

The most compelling and influential arguments that physical theory may fail to account for phenomenal experience have made use of the notion that there is something incommunicable about subjective experiences. Phenomenal experiences, it seems, are essentially private, and there are fundamental impediments to formulating them in a physical theory or in sharing that theory. Thomas Nagel created the most noteworthy form of this argument when he claimed that our subjectivity forms a kind, a product of our own nature, and that we cannot experience or understand radically different kinds of experience, such as that of a bat (1974). But the clearest and most powerful argument that we cannot reduce phenomenal experience to physical theory has been the Knowledge argument, which aims to show that phenomenal information is of a different kind than is physical information (Jackson 1982). The influence of these arguments has not abated, and they are cannonically cited as strong evidence against the physical reducibility of phenomenal experiences. However, these argument can be interpreted to show that phenomenal information is just *more* information than we typically attribute to our understanding of a physical theory. This points to two problems with subjectivity arguments regarding phenomenal experience. First, they assume that a scientific theory includes information not only about natural laws and natural kinds, but also about contingent historical facts. Second, they assume a mind is one homogenous capability with unbounded capacity. Correcting these errors suggests a novel theory of subjective experience and by extension of phenomenal experience: a full accounting of such experiences would be distinguished from standard physical theories only in being significantly more complex.

1. A Formulation of the Knowledge Argument

The Knowledge argument is now as familiar as it is influential: from the premise that phenomenal experience gives a kind of information, and the premise that the relevant sciences cannot predict this information, appears to follow the conclusion that these sciences are incomplete with respect to phenomenal experience. The first explicit formulation,[1] and the most intuitively compelling form, of the Knowledge argument was given by Frank Jackson (Jackson 1982). Jackson illustrates the argument with a thought experiment:

> Mary is a brilliant scientist who is, for whatever reason, forced to investigate the world from a black and white room *via* a black and white television monitor. She specializes in the neurophysiology of vision and acquires, let us suppose, all the physical information there is to obtain about what goes on when we see ripe tomatoes, or the sky, and use terms like 'red', 'blue', and so on….
>
> What will happen when Mary is released from here black and white room or is given a color television monitor?… It seems just obvious that she will learn something about the world and our visual experience of it. But then it is inescapable that her previous knowledge was incomplete. But she had *all* the physical information. *Ergo* there is more to have than that, and Physicalism is false. (1982: 130)

This argument leaves vague what "all the physical information" is. It thus could have two forms, given here as enthymemes:

<u>Knowledge Argument, Form 1</u>

Premise 1a: Mary has a complete (standard, scientific) theory of color vision; call this *T*.

Premise 2: When Mary sees colors she gains new information.

Conclusion: Physicalism (the claim that physical theory can predict all phenomenal information) is false.

Here I am assuming that a standard scientific theory is one that describes both natural kinds and natural laws that relate these kinds. This is meant to be a very general notion of scientific theory, and is consistent for example with either a classical syntactic view of theory or the various richer semantic views (see Suppes 1972, Gieri 1994). The notion that a theory that describes natural kinds and the laws that relate them deserves to be called "standard" might be challenged, but the arguments that follow will only require that *some* theories are like this to make the relevant points.

Alternatively, we could understand premise 1 differently, to be:

<u>Knowledge Argument, Form 2</u>

Premise 1b: Mary has a complete (standard, scientific) theory of color vision *and* she has all the physical information about any particular kind of visual experience she might have when she leaves the black and white room; call this *T+*.

Premise 2: When Mary sees colors she gains new information.

Conclusion: Physicalism is false.

Most responses to the Knowledge argument share the claim that the information or knowledge gained in phenomenal experience is somehow subtly different from the information or knowledge gained or described through physical theory, and so physicalism should not have to explain all of phenomenal experience but physicalism can still be somehow sufficient.[2]

However, one can grant Jackson's assumption that physicalism requires that phenomenal information is physical information of the kind that can in certain situations be predicted by the relevant physical theory, and still show that the knowledge argument is either invalid or assumes too much, even for strong forms of physicalism. For the sake of a demonstration here, I will assume one very strong formulation of physicalism: computational functionalism (the view that all mental states – and even phenomenal experiences – can be sufficiently described as discrete states of a Turing machine, plus the supposition that the brain instantiates something like such a Turing machine).[3] Even with such a strong theory, one can show that for the first form of the Knowledge argument, the premises can be true and the conclusion false; and that the second form of the Knowledge argument offers an unlikely vision of the relevant scientific theory, and also makes a hidden empirical assumption about our mental capabilities which is highly implausible.

I begin with the first form of the argument. This assumes that Mary has a standard scientific theory about color vision, which is a theory which describes natural kinds and natural laws. Fortunately, we do not need to establish what makes a theory physicalist to demonstrate the relevant points. We can keep this notion primitive. Suppose that T is a complete standard kind of physical theory of color vision. To say that T contains a complete standard kind of scientific theory of color vision means at least that if we have a sufficient description of any relevant historical situation of color vision H_{t1}, T and H_{t1} together entail (and, as noted, this can be syntactic or semantic entailment) all the relevant physical facts at some later time 2, H_{t2}. We do not need to explain what H_{t1} and H_{t1} are like, except to illustrate the concept: in the case of physicalism, for example, each H_{tn} might be a complete description at time tn of what is happening in someone's visual cortex during a color vision experience. Many would deny that this is a sufficient description of a phenomenal experience, of course, but as the target of the Knowledge argument is physicalism we need address the premises as they would be formulated for a charitable version of physicalism.

Note that theory T is a kind of information; it is the information of the physical theory that describes natural kinds and their relations through natural laws. H_{t1} is also physical information, but of a particular phenomenon: it is a description of the relevant visual color information that the agent experiences at time t1. Only historical physical theories, such as for example one finds in some parts of evolutionary theory, entail historical information about the environment at any time. There is then no *a priori* reason that a theory of color vision must entail any input to the theory. That is, T is unlikely to entail any H_{tn} for any time tn, just as, say, dynamics alone (without any historical information) does not entail that there is a moon in such and such a position in orbit of the Earth. But then Jackson's argument is simply invalid. Just as a complete theory of dynamics may predict no particular motion, a complete theory of color vision need not predict any particular color experience. Imagine a thought experiment in which Mary knows the complete theory of dynamics, but was locked in a closed environment where she never sees the moon and is never told about the moon. One day she is let out and sees the moon, and so learns that the Earth does have such a satellite. The analog to the first form of the Knowledge argument would then be:

Dynamics Argument

Premise 1a: Mary has a complete (standard, scientific) theory of dynamics.

Premise 2: When Mary sees the Terran moon she gains new information.

Conclusion: The claim that dynamics is a sufficient scientific theory of motion is false.

But this is obviously invalid. We should not conclude that dynamics was incomplete as a theory of motion. Most standard scientific theories will not entail all the relevant historical information but will still be excellent physicalistic accounts of the relevant phenomena. Dynamics does not predict what masses there are and where they are and how they are moving. Chemistry does not predict how many kilograms of hydrogen there are or where and when various chemical reactions take place. And so on. A standard theory tells us not all the relevant historical information but rather what one can predict from the relevant kind of historical information. Thus, premise 1a supposes that we have a theory that does not entail all the information of a phenomenal experience, but rather tells us what one can predict from such information; and so both Premise 1a and 2 could be true but the conclusion of the first form of the Knowledge argument could be false.

2. Descriptive Complexity of Phenomenal Information

A plausible response to this is that a complete theory of color vision should include *some* historically contingent information. This is to turn to the second form of the Knowledge argument, and hold that armed with the complete physical theory of color vision T+, Mary should be able to derive H_{t1} for that time t1 when she finally sees colors. Some plausible arguments can be made that a theory of color vision should include at least some such information. After all, such a theory must have significant connections to disciplines like ethology and evolutionary biology, and garner much historical information through them.

On this view, the theory of color vision is more like cosmology than like dynamics: it is explaining what particular organisms do and experience, and therefore must include historical information about particular kinds of organisms (i.e., humans), just as cosmology combines dynamics and other sciences with information about the historical facts of our actual universe. Thus, the Knowledge argument could be reformulated to explicitly endorse some version, perhaps a limited version, of premise 1b. Since by supposition Mary cannot previously be given this information by being shown the relevant information (that is, she first learns the theory T+ without seeing colors), this historical information must instead be conveyed in some kind of historical theory that we presume she reads or is told about. But no matter how well we describe red to Mary, we may believe that she will still learn something when she sees red for the first time. Thus, it seems, physicalism, even coupled with historical physical information, is false. The Knowledge argument, slightly revised, appears to still work.

Here is the second confusion. The Knowledge argument contains the assumption that the amount of information that we can manipulate with our theoretical understanding is sufficient to include the kind of historical information that (if physicalism is correct) constitutes the phenomenal experience of color. But this is neither obvious, nor even plausible. Humans have various mental capabilities, and each of these will have some limit to the amount of information that it can store, recognize, and manipulate. If we are physicalists, such limits are real and must constrain any account of the kind of information we suppose is involved in some mental function. Furthermore, these limits may be different for different capabilities, and these capabilities may not be able to fully share information. Most physicalists will suppose that these are various brain modules, which have varying degrees of interdependence with, and independence from, each other.

The information stored or manipulated in such modules can in principle be measured objectively, using a standard called "Kolmogorov complexity."[4] Kolmogorov complexity defines a measure of the minimal size of descriptions. The Kolmogorov complexity of a description is the smallest computer program that can generate (e.g., print) that description. The size of the relevant program in bits is one metric of Kolmogorov complexity. This measure is then a kind of measure of how compressible some information is. Some very large descriptions will compress down very significantly. An infinitely long string of "10101010…", for example, compresses to a very small program. The infinite extension of pi also compresses significantly. We say then that both have a low Kolmogorov complexity. But some descriptions cannot be compressed any more. Information which cannot be compressed we can call *Chaitin random*; a Chaitin random description that is 1000 bits long, for example, can be reproduced by a program that is no less than a 1000 bits of information. The measure of Kolmogorov complexity can be standardized if we agree upon the use of some particular universal Turing machine as a standard. However, the measure of Kolmogorov complexity is implementation independent, up to an additive constant; this follows directly from the fact that any universal Turing machine can perform the operations of any other such machine, with some finite additional program. For this reason, the measure is objective.

We do not need to speculate too much about what is being measured beyond that the measurement is of physical things, presumably brain states, since we are supposing that physicalism is true for the sake of showing that this second version of the knowledge argument is flawed. There are many physical states that could be measured, but the basic guiding principle is that one would make the coarsest (that is, simplest) measurements that identify and distinguish those states that make a difference to behavior (behavior here is construed broadly, as it must be, to include things like peceptuomotor control and not just linguistic reports). We have been using visual perception and visual experience as our example. If we measure the state of the visual cortex during some red experience, we would not measure molecules and their placement if it makes no difference (if moving these molecules could not possibly alter behavior), but would measure, say, the connectivity and firing patterns of individual neurons if altering them could make a difference (that is, if it follows that simplifying such a state would result in the loss of behavior). This is something of a platitude, but pointing it out should clarify that when we say that some kind of physical brain state is more complex than another, we are comparing similar physical kinds identified by their functional role.

We know that human individuals have at least the following two capabilities: they can understand, use, and manipulate theories; and they can see objects and respond to them or otherwise deal with them in various ways. But we cannot assume that these are the same capability, nor that they are fully integrated with each other so that the information in one can be freely and fully manipulated by the other. The kind of description that we might measure with Kolmogorov complexity could be a theory, such as the complete standard theory of color vision T, or it could be a historical description (perhaps incomplete) of the state of an individual visual cortex at time tn, H_{tn}. Note that T+, the richer historical version of a theory of vision, is T plus some historical information of the same kind as H_{tn}. Let I(T) be the Kolmogorov complexity of T. By supposition of Jackson's thought experiment, our subject can at least know T, so I(T) is less than or equal to the amount of information that the agent can recognize and manipulate with the cognitive capabilities that constitute the relevant theoretical understanding. Call the upper limit of the information that can be manipulated by theoretical understanding N (in this case, N can be the storage capacity of the theoretical understanding of Mary). Let H_{t1} be a description of all the relevant features of the state of the brain (presumably, the visual cortex) of Mary at time t1 that are required to fully describe the relevant (physical features of the) color experience at time t1. Let $I(H_{t1})$ be the Kolmogorov complexity of H_{t1}. Note that there is no reason to believe that $I(H_{t1}) \leq I(T)$, nor that $I(H_{t1}) + I(T) \leq N$, nor even that $I(H_{t1}) \leq N$. In fact, it is highly plausible that $I(H_{t1})$ is significantly greater, perhaps vastly greater, than N. For example, the portion of the human neocortex devoted to visual processing is very large, and appears to be significantly larger than that devoted to mathematical or linguistic reasoning. We are able to store and manipulate and use visual information in various ways, and such manipulations could well exceed our abilities to manipulate such information in the form of theoretical information or describe it linguistically. Thus, it could be that (1) T and H_{t1} is a fully sufficient physical account of color vision because together these predict the relevant subsequent state of the agent in question (T and H_{t1} entails H_{t2}); and (2) T and H_{t1} explain H_{t2} in such a way that if we did have a theoretical grasp of both T and H_{t1} we would find no reason to believe that something was "left out" of the theory (that is, we would not accept the conclusion of the knowledge argument); but (3) $I(T) < N << I(H_{t1})$, and so $N << I(T+)$. In such a case, the agent could have a complete standard theory of color vision T, but the agent could lack the theoretical form of the information that constitutes the actual having of the experience of color vision (namely H_{t1}), because in practice she cannot manipulate the phenomenal information H_{t1} with her theoretical understanding. This is consistent with the claim that the subject can herself experience color vision and learn something new thereby, since being aware of some information, being able to recognize it, being able to act in particular ways with that information, and so on, do not require that the agent's ability to theorize are sufficient to be able to grasp that information as part of a theory. Visual memory, to stick with our example of phenomenal experience, need not have the same kind of information storage and manipulation capabilities as our theoretical abilities. These are different kinds of abilities, and they can be instantiated in different parts of the brain, be significantly independent of each other, and have different capacities.

An analogy may help. The Kolmogorov complexity of Peano arithmetic is obviously small enough for human beings to memorize and work with this theory, but there are large numbers which are Chaitin random that a mathematician will not be able to memorize or practically work with.[5] Suppose now that Mary knows Peano arithmetic, Zermelo-Frankel set theory with the Axiom of Choice, and how to construct the natural numbers using set theory; let us call this combination of theories, *basic arithmetic theory*. A Knowledge argument that arithmetic was not reducible to basic arithmetic theory would go something like this. There are very large Chaitin random numbers that Mary cannot memorize, recognize, or manipulate, even though she knows basic arithmetic theory; this follows simply from the facts that Mary must have a finite theoretical memory, and that there are infinitely many numbers of Kolmogorov complexity exceeding the information capacity of that memory. Let B be such a number. Even though Mary knows basic arithmetic theory, she cannot "in her head" add B to itself. Therefore, B and its properties are not expressible or derivable in basic arithmetic theory. Namely, basic arithmetic theory is incomplete in respect to addition with B. The analog of form 2 of the Knowledge argument then is something like:

Basic Arithmetic Theory Argument

Premise 1: Mary knows basic arithmetic theory *and* she has the ability to store, recognize, and manipulate all numbers in which she may have an interest.

Premise 2: Mary cannot predict the relevant properties of B.

Conclusion: Basic arithmetic theory is not a sufficient theory of numbers and their arithmetic properties.

This argument has obviously gone badly wrong. The problem is with premise 1, which mixes human capabilities with the limitations of theory. All of the properties of B that interest us may be expressible or derivable in basic arithmetic theory. But B is just too large for Mary to handle, and so numbers she may have an interest in may be beyond her theoretical abilities to remember and manipulate. Just so, there is no reason to believe that the information in our visual systems, which we use effectively to guide ourselves through the world, can be compressed enough to be manipulated with those capabilities (a physicalist might suppose, with those brain modules) that enable theoretical reason, nor to be communicated in the bandwidth of human speech. But this does not show that this information is not the kind of information that could be fully expressed and used in predictions by a physical theory of color vision, just as Mary the mathematician's limitations show nothing about the relative completeness of basic arithmetic theory. Thus, we have not shown that phenomenal information is not physical information.

The mere possibility of this kind of modularity, and of these kinds of information capacity differences between human capabilities, means that the historical information of a particular visual experience may be beyond the theoretical scope of an individual working with a successful complete theory of vision. We should no more expect Mary to have all the information of a red experience than we should expect a great mathematician to be able to remember and divide "in her head" two incompressible (Chaitin random) numbers that are 1000 numerals long.

3. Subjectivity as History

A third form of argument from subjectivity is that the relevant theory may be too complex for humans to understand with their theoretical capabilities. The clarity of the Knowledge argument allows us to identify this implicit problem in the more vague claims about subjectivity that form other subjectivity arguments against the possibility of a physicalist theory of phenomenal experience. Nagel's paper, "What is it Like to be a Bat?" is one of the most influential examples of such an argument. Nagel argues that we could never get inside the subjective viewpoint of a bat, and grasp the experience of echolocation. He recognizes that this may have something to do with human limitations:

> My realism about the subjective domain in all its forms implies a belief in the existence of facts beyond the reach of human concepts. Certainly it is possible for a human being to believe that there are facts which humans never *will* possess the requisite concepts to represent or comprehend. (1974: 441)

The talk about concepts here is very misleading. After all, we can easily have a concept like, that-which-humans-will-never-understand, and so have a concept of the phenomenon in question. Rather, we can make more sense of such limitations if we recognize them as regarding theory and our abilities to hold and use such theories. To use the notation above, this would be to entertain that the information of a theory standard theory T of a kind of phenomenal experience, perhaps even if just a standard scientific theory without reference to historical states, may be less than our theoretical grasp can handle: $N < I(T)$. But, as already noted, this would not show us anything about whether T was an adequate theory of phenomenal experience.

This is of course possible; until we have such a theory we have little guidance to allow us to predict its complexity. But we can take as moot the supposition that any successful standard theory of a phenomenal experience is going to be beyond our grasp because of its complexity, since it is far more likely that the relevant historical information is going to be beyond our grasp. The wonder of good theories, after all, is that they are relatively simple, whereas historical facts are not. This suggests an alternative and more conservative interpretation of the difficulty that the subjectivity arguments struggle to articulate: subjectivity is history.

It is plausible that there are good reasons why our brains would have modules that generate and use sensory, autonomic, and other kinds of information which exceeds in complexity the information capacity of our theoretical abilities. First, if there is benefit to having the ability to recognize and otherwise use a great deal of complex sensory information, there may be some evolutionary pressure to increase that information capability or at least keep it large. There is no obvious reason that we would have a significant or even equal benefit from having theoretical information handling capabilities that can manage that same information. If we are realists and assume the environment is shared, for example, then there is likely a great deal of parsimony to be had in letting the world speak for itself. We can leave our theoretical abilities not for describing in complete detail shades of red but rather, based on the assumption we both see the red, we can use them instead to make and communicate relevant inferences. Second, it seems unlikely that language could ever manage to convey the information in a visual experience, for example, in a timely way. Given the bandwidth constraints of this medium, which we inherit and do not choose, it is likely that it is better used for other purposes. Such purposes are facilitated, again, by the assumption of realism and the consequent assumption that we share the relevant sensory experiences.

The complex internal states of some human having a red color experience, or a bat having an echolocation experience, are likely extremely complex. It is even possible that if we take all the features of such an experience into account, it might sometimes be unique. But now we are on the verge of a platitude. For standard scientific theories do not predict historical events, but strive only to predict what follows from such events. Thus, it could well be that some future standard physical sciences of mind will be incomplete with respect to phenomenal experiences, but only in the same way that other successful sciences are incomplete with respect to, say, the actual composition of a wetland of such dimensions at such a place and time.

The view of mind that underlies these subjectivity arguments about phenomenal experience is based on, or rather assumes without reflection, a naïve conception of a single, homogenous mental capability with unbounded storage capacity. Of course, scientific evidence marshals an overpowering case for some form of modularity in mental capabilities, and no physicalist can allow that any such module has unbounded storage capacity. We cannot then assume, as do these subjectivity arguments, that the information in one such module is available to or could even "fit" in another.

This suggests a novel theory of consciousness, and of why subjectivity arguments about phenomenal consciousness have such appeal. Some phenomenal states may be in principle explicable with physical theory, perhaps even a theory we can understand, but the composition of their actual instances may be of such great complexity that they cannot be given a humanly manageable description. This historical information is relevant – it is right to call it information of such a complexity and to not reduce it further – because it guides our actions, can be recognized as such, and so on. But it appears to us as resistant to physical theory because it is rather resistant to our ability to manipulate it as theoretical information.[6]

REFERENCES

Carnap, Rudolph (1967) *The Logical Structure of the World and Psuedoproblems in Philosophy*. Translated by Rolf A. George. Chicago: University of Chicago Press.

Chaitin, G. J. (1966) "On the Length of Programs for Computing Finite Binary Sequences." *Journal of the Association of Computing Machinery,* 13: 547-569.

----- (1982) "Gödel's Theorem and Information." *International Journal of Theoretical Physics,* 22: 941-954.

Churchland, Paul (1985) "Reduction, Qualia, and the Direct Introspection of Brain States." *Journal of Philosophy,* 82 (1): 8-28.

Giere, Ronald, N. (1994) "The Cognitive Structure of Scientific Theories." *Philosophy of Science,* 61: 276-296.

Jackson, Frank (1982) "Epiphenomenal Qualia." *Philosophical Quarterly,* 32: 127-136.

Kolmogorov, A. N. (1965) "Three Approaches to the Quantitative Definition of Information." *Problems of Information Transmission,* 1 (1): 1-7.

Lewis, David (1999) "What Experience Teaches." In *Papers in Metaphysics and Epistemology, Volume 2:* 262-290. Cambridge, UK: Cambridge University Press.

Loar, Brian (1990) "Phenomenal States." *Philosophical Perspectives 4, Action Theory and the Philosophy of Mind,* J. Tomberlin (ed.): 81-108. New York: Ridgeview Publishing Company.

Nagel, Thomas (1974) "What is it Like to be a Bat?" *The Philosophical Review,* 83: 435-50.

Solomonoff, R. J. (1964) "A Formal Theory of Inductive Inference, Parts I and II." *Information and Control* 7: 1-22, 224-254.

Suppes, F (1972) "What's Wrong with the Received View on the Structure of Scientific Theories?" *Philosophy of Science* 39: 1-19.

NOTES

[1] Arguably some of the presuppositions of Jackson's knowledge argument were around for some time. For example, the early Carnap argued that instances of phenomenal experience are not scientific knowledge, since they are unverifiable, and also because science accounts for structural relations between objects which are inferred from sense data, but science does not account for the nature of sense data themselves (see Carnap 1967: 19ff).

[2] A few recent examples of discussions of the knowledge argument that grant that there are different kinds of knowledge involved include: Churchland uses the distinction between knowledge by acquaintance and knowledge by description (Churchland 1985); Loar proposes that phenomenal experience is knowledge that concerns objects under a different description than they are given in theory (Loar 1990); Lewis argues that phenomenal experience is a kind of knowing how, whereas theory is a knowing that (Lewis 1999).

[3] If one adopts an alternative view of mind to computational functionalism but remains a physicalist, the arguments that follow should still work. However, the application of standard notions of descriptive complexity might require substantial additional mathematical work to generate relevant measures for continuous states. Since this is not necessary to make the points of this essay I will not extend the argument to other kinds of views of mind.

[4] "Kolmogorov Complexity" is a common name for this measure of descriptive complexity, but the measure was independently developed by three mathematicians: G. J. Chaitin, A. N. Kolmogorov, and R. J. Solomonoff. See (Chaitin 1966, 1982), Kolmogorov (1965), Solomonoff (1964).

[5] A simple counting argument shows that there are many Chaitin random, or incompressible, numbers of each size. Suppose for simplicity we are using a binary coding of numbers, and

have fixed a universal Turing machine to use as a concrete measure. At most about half of the strings n bits long can be compressed, since there are 2^n such strings and at most 2^{n-1} shorter strings that they could be compressed into.

[6] Special thanks are due to XXXX XXXX for some helpful observations.