2017

"Hail Mary, Full of Haze": Physicalism and the Knowledge Argument

Jesse R. Powell
Georgia Southern University

"Hail Mary, Full of Haze": Physicalism and the Knowledge Argument

An Honors Thesis submitted in partial fulfillment of the requirements for Honors
in the Department of Literature and Philosophy

By
Jesse Powell

Under the mentorship of
Dr. Joe Pellegrino

ABSTRACT

This project aims to provide a clear and compelling reason for rejecting dualism with respect to the mind, by undermining the support dualist positions receive from so called knowledge arguments. In particular, I will show the error present in the many forms of what is variously called the "Mary's Room" or "Mary the Brilliant Color Scientist" thought experiment.

Thesis Mentor:________________________

Dr. Joe Pellegrino

Honors Director:______________________

Dr. Steven Engel

April 2017
Department of Literature and Philosophy
University Honors Program
Georgia Southern University

> Mary is confined to a black-and-white room, is educated through black-and-white books and through lectures relayed on black-and-white television. In this way she learns everything there is to know about the physical nature of the world. She knows all the physical facts about us and our environment, in a wide sense of "physical" that includes everything in *completed* physics, chemistry, and neurophysiology, and all there is to know about the causal and relational facts consequent upon all this, including of course functional roles. If physicalism is true, she knows all there is to know. For to suppose otherwise is to suppose that there is more to know than every physical fact, and that is just what physicalism denies… it seems, however, that Mary does not know all there is to know. For when she is let out of the black-and-white room or given a color television, she will learn what it is like to see something red, say. This is rightly described as *learning*—she will not say "ho hum." Hence, physicalism is false.

This is Frank Jackson's "Mary the Brilliant Color Scientist," or "Mary's Room" thought experiment, in its most polished and consistent form (Jackson, in *There's Something*, 51). Jackson offered this refined version in 1986, intending it to replace his initial formation of the thought experiment, created in 1982. It has proven to be one of, if not *the* single most controversial thought experiment in the philosophy of mind. Why might this be so? There are two specific reasons why this thought experiment has become such a touchstone in the contemporary discussions of the philosophy of mind. First, as is made explicit by Daniel Stoljar and Yujin Nagasawa, the thought experiment "zeroes in" on that aspect of conscious experience that is most puzzling: the quality of experiencing consciousness first-hand, the "what-it-is-like" to have a conscious experience. We are so inadequately equipped to explain this experience to one another that by virtual consensus we have come to consider it to be wholly ineffable (Stoljar and Nagasawa, 1-2). The second is the relationship that Frank Jackson has with his "creation." In the span of three years, from 1993 to 1996, Jackson went from blatantly denying the truth of a solely physical universe to rejecting wholesale the conclusions he initially arrived at through his thought experiment. He explained this change in his understanding in 2004, saying

“[T]he argument... contains no obvious fallacy… yet its conclusion—that Physicalism is false—must be mistaken" (Stoljar and Nagasawa, 23). I am inclined to agree. This reversal of his initial position means that Jackson has experienced both of the immediate reactions I feel most readers have after having been exposed to the Mary thought experiment for the first time; either they find it obviously true, or they find it obviously fishy (responses it elicits by design; we will return to this point later.). For those of you whose reactions are like mine, of the fishy kind, here is the challenge: pointing to where exactly the argument goes wrong. That is precisely what I intend to do here. I will first discuss the broader class of “knowledge arguments.” Then, after a brief discussion of what a thought experiment and an intuition are, I will run over the various forms Jackson’s thought experiment has taken over the years, as to make clear all the relevant features that may elude our grasp were we to consider just one. Specifically, I will be looking at three forms from Jackson 1983, 1986, and 1998. After I’ve propped it up nice and thorough, I will knock it all down by demonstrating a fatal flaw that will plagued any Mary thought experiment know matter what form it may take.

So what is a knowledge argument, and why does it matter here? Let me lead with an example. Consider this example from C. D. Broad: Consider a “mathematical archangel,” a being with domain specific omnipotence, knowing everything about logic and chemistry. If this were all the angel knew, it would be unable, according to Broad, to predict the way that ammonia would smell had it never smelled ammonia. Therefore, there must be some other sort of stuff to know that is not simply chemical or logical or any combination of the two alone (Stoljar and Nagasawa, 6). What this example demonstrates is twofold. First, it shows us an example of what is called the *knowledge* *intuition*. The knowledge intuition is a “prima facie modal truth;” it is an “obvious” statement of possibility. Think about the way roses smell. Compare the *experience* of smelling a rose with all the sorts of things you can learn about roses and smells from textbooks: the chemicals involved, the sensory tissues of the nose and their connections with our brains, etc. No matter how many textbooks you read, it seems *obvious* that you at no point know what roses smell like unless you smell them; unless you experience the smell of a rose. Or similarly, how would you explain to a blind person what a rose smells like? What could you say that would make them understand? Could you do this? For many, it will seem *obvious* that there just is not any way at all to make a blind person experience color through linguistic descriptions of *our* experiences of color alone. It’s just one of those things you have to do for yourself. If you have this feeling, then you feel the knowledge intuition: that it is possible to know physical knowledge without knowing any *phenomenal* knowledge. Another way of putting phenomenal knowledge is “what it is like.” Second, it demonstrates a common conclusion that is drawn from this intuition; if it is possible to have no experiential knowledge, no phenomenal knowledge, when we have nothing but physical knowledge, then it seems there must be more than just what is physical. Thus we have the full definition of a knowledge argument as understood by Stoljar and Nagasawa: any argument that includes the knowledge intuition as a premise, and concludes that Physicalism, the belief that everything that exists is physical, is false. The simplest form of such an argument may have been given by Jackson, when he says: “Nothing you could tell me of a physical sort captures the smell of a rose, for instance. Therefore, Physicalism is false” (Jackson, *Epiphenomenal Qualia*, 127). If these are the two sufficient conditions for being a knowledge argument, containing the knowledge intuition and reasoning that Physicalism must be false, then there have been many historical instances of knowledge arguments much earlier than any of Frank Jackson's work, like Broad's mathematical archangel discussed earlier.

So if he didn't come up with the idea himself, and his argument is not the only knowledge argument, what makes Mary special? Why is she different? One obvious difference is the rhetorical effectiveness of the Mary thought experiment as opposed to the simplified form of the knowledge argument given above. Even if they assert the same conclusion, I think it is fair to say that Mary's Room does so in a more convincing way. So why is that? It begins with the fact that it is a thought experiment. What a thought experiment is supposed to is allow us to reason to conclusions about normally observable events when "normal", traditional kinds of experiments cannot be completed, or as a supplement to these normal experiments. They consist of a series of imaginable physical events, that can be imagined in sequence until they have run their course, until they reach an end state, and the conclusions we can infer from the imagining of this end state. These conclusions function as pseudo-observations; what you conclude is not observations proper, but instead are rational *intuitions*; they are things that *seem* right, and are not immediately demonstrable as false. These feelings may fall short of being empirical evidence, but they are usually extremely compelling nonetheless; it will many times just seem to be the only way things could be.

There are a number of things that can be done to strengthen feelings of intuitiveness. One such thing is to set up the thought experiment in such a way that it is *reductio ad absurdum*. This means setting up an argument where the truth of one of the premises leads to a contradictory conclusion, and so we can reject the premise responsible for the contradiction as false. Another tool we can use to strengthen our intuitiveness of our conclusions is to use *eliminative inference*, where every alternative conclusion is shown to be less satisfactory than our conclusion. One thought experiment that puts both to use masterfully is Galileo's Falling Bodies Thought experiment. Galileo presents the thought experiment through a dialogue in his *Discoursi*:

> SALVIATI: If we take two bodies whose natural speeds are different, it is clear that on uniting the two, the more rapid one will be partly retarded by the slower, and the slower will be somewhat hastened by the swifter. Do you not agree with me in this opinion?
>
> SIMPLICIO: You are unquestionably right.
>
> SALVIATI: But if this is true, and if a large stone moves with a speed of, say, eight, while a smaller stone moves with a speed of four, then when they are united, the system will move with a speed of less than eight. Yet the two stones tied together make a stone larger than that which before moved with a speed of eight: hence the heavier body now moves with less speed than the lighter, an effect which is contrary to your supposition. Thus you see how, from the assumption that the heavier body moves faster than the lighter one, I can infer that the heavier body moves more slowly.
>
> And so, Simplicio, we must conclude therefore that large and small bodies move with the same speed, provided only that they are of the same specific gravity (Swan, 346-47).

Thus, it would seem there is simply no need to observe the scenario laid out by Galileo; either a lighter and heavier stone tied together fall faster than the heavy stone by itself or they don’t. If they do, we get a contradiction, so we only have one option left. The dialogue compels the audience’s imagination into considering the conclusion undoubtedly, obviously correct; you can see where this could come in handy.

With a basic understanding of what a thought experiment is, the knowledge intuition and the knowledge argument, we can begin a deeper analysis of Jackson’s thought experiment and knowledge argument combo. Perhaps it would be even better to say *family* of thought experiments, as he has given the same basic argument in a multitude of forms, each highlighting important features of their Wittgensteinian family resemblance, with varying degrees of rhetorical effectiveness. By looking at three versions in particular, I will identify the important aspect of the argument laid bare by each, and then conjoin them into a form of the Mary thought experiment/ knowledge argument that is as transparent as possible.

Let’s begin with Jackson’s 1986 article “What Mary Didn’t Know,” where we get the form of the argument I began this project with. The important thing to note here is the description of what counts as “physical”: “Everything in *completed* physics, chemistry, and neurophysiology, and all there is to know about the causal and relational facts consequent upon all this, including of course functional roles” (51). Now what does it mean to know everything in “completed” physics? Well, it means that there is not a single fact about physics, that Mary would be surprised about; she has a perfect physical *theory*. She knows all physical facts and the causal relations between them, themselves being physical facts. What do theories do? They are more than just a collection of facts. Following Van Fraassen (1970, 1989), I am going to here on discuss theories with respect to *models*. A model is called a model for a particular theory *if and only if* the theory is entirely true with respect to that model. Models here are some mathematical structure or another: some geometric figure, some collection of sets and functors between them, etc. In the case of physics, we might think of a probability distribution as our model. Under this interpretation of theory, at any given time, (at least for all those physicists who aren't Mary,) the job of science is to try and produce a theory that is made true by a probability distribution representative of the real world. We might think of the distribution as representing possible events in space-time and the likelihood of them occurring under such and such conditions, and so on. What physicists do is use empirical observations to formulate a *prediction* about the nature of reality, in the form of a probability distribution potentially equivalent with that one representing the real world. They then find ways of testing the model, and they adjust according to the results. In other words, they make a guess as to what the full model of the universe is, using the pieces of the model we have gained knowledge of through empirical data. For Mary, any prediction she made would have to be 100% accurate, if it was indeed the case she had a final, complete theory of physics. If it turned out that her model did not align in every conceivable way with the actual probability distribution that is reality, then there must be some fact about physics that she does not know, and this is exactly what we are asked to assume when doing Jackson's thought experiment. That is enough on theories for now, but we will return.

The original form of the thought experiment is found in "Epiphenomenal Qualia" published in 1983:

> 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. She discovers, for example, just which wave-length combinations from the sky stimulate the retina, and exactly how this produces via the central nervous system the contraction of the vocal chords and expulsion of air from the lungs that results in the uttering of the sentence ‘The sky is blue.’ (It can hardly be denied that it is in principle possible to obtain all this physical information from black-and-white television, otherwise the Open University would of *necessity* need to use color vision.)
>
> What will happen when Mary is released from her black-and-white room or is given a color television monitor? Will she *learn* anything or not? 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* physical information. *Ergo* there is more to have than that, and physicalism is false. (130)

Here the argument is unique in two particular ways that I find significant. Instead of physical “facts,” Jackson describes Mary as knowing all physical *information*. The difference is significant because of the difference in the connotations of the terms. The term *information* I associate with formal information theory, and as such may be a possible aid in discovering a solution to the challenge Mary poses to physicalism. Also, when Jackson discusses introducing Mary to colored television, he raises an important question that is central to the issue: whether she will in fact learn anything or not. But before we have had a real chance to formulate an answer for ourselves, in the next sentence Jackson answers the question for us: “[Upon seeing colored images for the first time] it seems just obvious that she will learn something about the world and our visual experience of it.” Thus we have a second intuition lurking in the premises, tipping the scale even further against physicalism; that a person seeing colored images for the first time *must* learn something new. Notice how effective this intuition is; certainly, I will concede that I instinctively *feel* that Mary, having never seen color, upon doing so will learn something new. I admit the *apparent* correctness of this learning claim, and I will venture to speak for the general populace and say that the way I feel is not uncommon. But if we can find no reason for rejecting our feelings; if it turns out that our intuitions are to be trusted or that we can find no evidence to the contrary, this feeling (one that a great number of us most likely experience when we imagine seeing color for the first time) will lead us to conclude that there is indeed non-physical information to be learned. Whether we can trust our intuitions or not, the two aspects of this argument are here made explicit, where everywhere else they are only implied. Jackson treats as synonymous the terms *facts* and *information*, and implies that we have a “learning intuition” so to speak, that being presented with a new modality of sensory experience--in this case the introduction of full-color visual experiences, as opposed to merely black-and-white experiences--will result in learning.

Returning to Jackson’s 1986 article, there is a third form of the Mary thought experiment that can tell us about the argument family in which Jackson positions his thought experiment. This form allows us to categorize the experiment in ways that are not so easily ascertained by studying any other version. Here Jackson gives the argument in the form of a syllogism, reformulated to deal with the objections to his original formulation in “Epiphenomenal Qualia” offered by Paul Churchland in his “Reduction, Qualia, and the Direct Introspection of Brain States”:

1. Mary (before her release) knows everything physical there is to know about other people.
2. Mary (before her release) does not know everything there is to know about other people (because she *learns* something about them on her release.)
3. Therefore, there are truths about other people (and herself) which escape the physicalist story. (293)

A few differences between the original argument and this revision are immediately clear: the argument in this version claims as necessary a strict ordering of events. That is, the two premises above (numbers 1 and 2) must occur before Mary experiences color, for them to support the conclusion. Each premise is formulated so as to support the conclusion that physicalism is false. This is stated clearly in the original form as well, but to see it set apart by parenthesis and as a part of a sparser argumentative form in general serves to bring to the forefront the temporal component of both the premises and the conclusion. Furthermore, and perhaps more interestingly for my purpose, Stoljar and Nagasawa point out that this argument is a quantificational one; it makes claims about a class of entities:

1. Every physical truth is such that Mary (before her release) knows that truth.
2. It is not the case that every truth is such that Mary (before her release) knows that truth.
3. Therefore, there is at least one truth that is nonphysical. (14)

What becomes uniquely clear in this restatement is that, especially in the conclusion, Jackson's argument hinges on the nature and existence of one specific truth; his conclusion is that there is at least one truth that is non-physical. Stoljar and Nagasawa can also be credited with the last pertinent observation: the conclusion of Jackson's original formulation--that at least one non-physical truth exists, and therefore physicalism is false--can be contested, not because of the logic of his argument, but because of a possible interpretation of his terms. Thus, the original form of the experiment cannot ever disprove physicalism. His reformulation of the experiment (and their restatement of it above) closes that loophole. To understand this point, Stoljar and Nagasawa give a definition for the *psychophysical conditional*, a logically deductive conditional concerning the *supervenience* of the psychological on the physical. According to Jackson in “Mind and Illusion” (1998), the intuition that Mary learns something new when seeing a color television set is an intuition that “Mary cannot carry out an *a priori* derivation from the physical information imagined to be at her disposal to the phenomenology of color vision.” It may be the case that the psychophysical conditional is possibly necessary and *a posteriori*, learnable only by some empirical observation or another, but a response to a version of the knowledge argument set up so as to suppose the “a posteriori-ness” of the learning intuition is not an argument equivalent to Jackson’s original family of arguments. So if we wish to formulate responses to the true set of Mary thought experiments, we must understand the psychophysical conditional as being a priori. So then, physicalism, the claim that everything that is true of the world is physical in nature, we assume to imply both that the psychophysical conditional is necessary and also *a priori* (Stoljar and Nagasawa 15).

I have discussed in length the different features of the family of arguments under scrutiny that are important for a proper understanding of what the whole family is really up to, but that are more or less concealed from most of the argument forms in particular, if not by all but one. I now will formulate a version of the argument that makes explicit every important feature of the Mary thought experiment in all its various implementations. I call it the "Hail Mary" version of the argument, because it is a sort of last ditch effort to maintain any version of the argument's validity and soundness, one that is the clearest, and that accounts for as many of the traditional objections that it can through clarification alone, without changing the semantic content of the first form of the Thought Experiment given in 1983. If this version can be defeated, any version can. It is as follows:

1. Assume that there is a possible world where a scientist (one who makes predictions based on facts) named Mary knows the *complete and final physical theory,* that is, all physical information/all physical facts, despite having never been anywhere but the black-and-white room she is currently in nor has she ever had any exposure to colors besides black and white (from Jackson 1983 and 1986).
2. If at time *t* Mary has never seen color, and at any time *t+n,* were she to be presented with a color television, it is *possible* that she learns something she didn’t know beforehand (Jackson 1983).
3. If physicalism is true, then the proposition, “*it is true in every possible world* that the psychological intervenes on the physical*,”* is true *a priori* (Jackson 1998).

4. *Necessarily*, if Mary knows the complete and final physical theory at time *t* and *learns* at any time *t+n*, then there exists at least one non-physical truth (Jackson 1986).
5. Either there is at least one non-physical truth, or physicalism is true (see Jackson in *There is Something About Mary*, pg. xvii).
6. Therefore, it is *possible* that physicalism is false.

In this form, we can see all the features of the family resemblance explicitly stated, and the many systems of logic that would be needed to symbolize a Mary Thought Experiment; the temporal aspect, the alethic modality and Jackson's assumed two-dimensional semantics for it, the quantification, the production of an intuition by the knowledge argument; we can see all these things clearly in the "Hail Mary" form. Thus concludes my attempt at a fair treatment of Jackson and his reasoning. I believe this to be a robust exegesis of what Jackson was trying to get at, and I hope that this final version of my own design is one Jackson would agree is not only fair, but as polemically capable as any he came up with himself.

Mary, Mary, Quite Contrary

So, now that we have had a thorough introduction to the knowledge argument and its parts, and the special class of knowledge arguments that are the family of Mary thought experiments, does any error become apparent? Or is their apparently no error? I suggest that there is indeed an error, although it is far from being apparent. This is precisely why the argument has been so controversial, and continues to generate replies to this day. It is why it is so fascinating, precisely because there is as of yet nothing approaching a consensus on what exactly the problem is, if there is a problem at all.

Allow me to put forth my proposal for at least *a* problem with the thought experiment, making no claim about whether it is “the” problem. Let’s begin with the synonymy of fact with information that Jackson makes. As I said before, I believe that formal information theory is helpful here. In Information Theory, one standard definition of *information* is given by Claude Shannon and Warren Weaver in *The Mathematical Theory of Communication*. They present *information* strictly in terms of the probabilities of events, demonstrating how it can be quantified with absolute precision. For example, think of a message in a sealed envelope that you’ve just received in the mail. Reading the return address, you learn that the letter was sent to you from your grandmother back home. Now knowing your grandmother, you know she is monolingual, specifically knowing only that language that is your first language (let us just use English). Now two things are true of the message within this envelope: Knowing your grandmother, we know that there are only a limited number of messages that she could have wrote us, and that the message within the envelope could be any one of them, but it can only possibly one of them. So, before we open the envelope, there is a great deal of *uncertainty* about what it contains inside, and if we were to open it and begin reading, with each word we read, we get closer and closer to knowing exactly what message we have been sent. With each consecutive word, the number of possible messages gets smaller and smaller, until finally we have read the note in its entirety and have learned *exactly* what message we have been sent. Now with respect to the example just given, we can see that information functions to reduce *uncertainty*. Before we have opened the envelope: we know there are only so many things that the message within *could* be, or in other words, we have a rough *probability distribution* of possible messages, and even before reading anything we can put varying levels of probability on each possible message from background information about our grandma; how her health is doing, what hobbies she has picked up recently, how close it is to Thanksgiving, and so on. In its unread state, the note from our grandma has the potential to *greatly* inform us as to what she is communicating to us, and different words will inform more or less as we read them (for example, if we were to read "Thanksgiving" it would greatly increase the chance that the letter is about a Thanksgiving dinner, and it would greatly reduce the chance of the letter being about her latest game of bridge, etc.) but as we read more and more of the letter, the chance that the next word we read will clue us in as to what the letter is about, that is, that whatever we think it is about up unto that word is mistaken, slowly decreases; it is highly unlikely that we could read all the way to the last sentence, and have not come extremely close to knowing what the letter is about; there isn't much chance that grandma wrote an entire letter where we can be completely mislead as to what she is writing us for *until* we read the very last sentence. So to recap, think of information as having the ability to reduce uncertainty, and as being carried by some sort of "message", or if you like, a symbol. Relative to just one set of possibilities, where only one alternative is the "true" state of the natural world, each consecutive piece of relevant symbols must contain less and less information, so that if you have amassed say 99% if the relevant messages related to a probability distribution, that other 1% is highly unlikely to let you in on anything that will drastically change the likelihood of alternatives. Now knowing this, we can make a crucial inference: were you to arrive at the *exact* distribution of probabilities, that is, the "one true" distribution, there is nothing you could afterwards learn that could provide you *any* relevant information *at all*. No symbol after having arrived at the true distribution of probabilities could suggest a different distribution. Information can only exist when there is space between your understanding and reality. Whenever you know the “true” probability distribution, then you know *all* the messages that could possibly inform about said distribution; all the relevant information.

Next, let’s take a closer look at what exactly a physical theory accomplishes. Mary is a color scientist, one who knows all the physical facts. It is fairly uncontentious to say that a crucial part of any science is the testing of theories through experimentation to see which ones can be tossed out, and which ones are the closest to the truth. How experiments test theories is usually understood in terms of *prediction.* Theories make predictions about what would happen given a particular set of initial conditions and a certain amount of time. If we know the initial conditions required, and what outcome our theory predicts, in principle all we need to do is recreate the initial conditions, observe what happens, and see if it matches up with what our theory said would happen. If it doesn’t, it’s possible we did not do the experiment well enough; for example, we could have forgotten to *control for*, to try and minimize the influence, of a condition that could alter the expected outcome. But incase we can verify all the proper controls were done, our theory can be regarded as incomplete, or not totally correct. When the latter is said to have occurred, our experiment has lessened the chance that whatever theory we took our initial conditions from is correct. Earlier I mention Van Fraassen, and I think his definition of theory is helpful here, as is the concept of a *model*. To Van Fraassen, a scientific theory is a sort of formal system, containing a set of axioms, and a related model. Now there are many different sets of axioms, and many different models, but for each set of axioms, there is one *unique* model that it has a special relationship with. He writes:

> A model is called a model of a theory exactly if the theory is entirely true if considered with respect to this model alone. (Figuratively: the theory would be true if this model was the whole world.) (1989, 218)

Let me put it another way: consider some set of propositions, call it *A*. Now, starting with *A*, let us say that the set of all propositions in *A* and all those propositions that we can *infer* from *A* make up a second set, call it *T*. For any *T*, if *T* forms a bijection with any other set, call it *M*, that is, if for every element in *T* there is one and only one element in *M* such that the two form a pair and vice versa, then if we call *T* our “theory”, we can call *M* the model of our theory *T*. The illustration Van Fraassen uses is geometric: given some set of axioms, the simplest geometric figure for which each axiom and anything we could infer from them (we call these inferences *theorems*) is true of the figure, that figure is the model for our theory, made up of our axioms and their inferences. Considering scientific theories in this set theoretic semantics, we can understand the job of the real world scientist to be identifying that special set of axioms and theorems that is made true by a special model: the model equivalent to *the real world*. The model that is equivalent to the real world is unique in that we gain knowledge of the model in a piecewise manner. Every observation we make shows one particular piece of the model. If theory testing is comparing real world outcomes that we observe to the predictions of our theory, under a set-theoretic semantics, what predictions amount to are inferring theorems from our axioms. To test if our theory is made true by the real world, we try and recreate the necessary preconditions (The relevant axioms and theorems) needed to create the prediction our theory makes. If we do our experiment correctly, the outcome will give us a true part of the model, and whether or not our axioms and theorems correctly predicted the part of the model we would observe determines how true the theory is; how close it is to the *final theory*, the theory made true by the entire model of the real world.

So if we stick to a set-theoretic semantics for theories, we find ourselves in a situation of *uncertainty* with respect to the theory true of the real world, but also to the complete model that is the real world. Given some subset of the complete model of the real world, let's call them *W*, we can see how more or less true all the competing theories we have are. In this way, given any subset of the model of the real world, each additional piece of our model *informs us* about the probability distribution over possible events in the real world, the rules governing the behaviour of things in the real world, and the probability that any particular theory will be made true by the whole model of the real world. We gain more pieces of our model by running *experiments*, where the outcome is some piece of the model or another that we then compare to what each of our theories says *should* have happened. The theory that gets it the most right we then consider to be more likely the true theory, and so on.

Mary's Lamb Gone Astray

So what does this have to do with Mary? Well think about what Frank Jackson claims she is in possession of before watching the color television: all physical facts/information. If Mary has any facts at all, they are of the world, meaning they are pieces of the model that is the real world, where here “real” is an indexical, the facts are a part of whatever world is real to the Mary we are thinking about, the possible world we imagine her to be in. And if Mary has all physical information, the only thing the information could *inform* about are the physical truths of the world. What all this means in the light of information theory and set theoretic semantics for scientific theories is that A) Mary could receive no message at this point that would *inform* her about anything physical, and B) Mary can learn no new fact at this point that was a *physical fact.* But if she has all the physical facts, we can say she has the complete *physical* model of the world. If she has the final physical model, she has the final physical theory, because nothing she can learn at this point can inform her; she has 100% of the relevant information, and so knows the exact distribution of probabilities of competing physical theories, in this case, one theory in particular, the true final theory, has a 100% probability of being correct, and all other theories have a zero percent probability. Remember this is all before she has ever seen any color beyond black and white. So concerning the color television: we agreed that *if* she were to see the colored television having complete physical knowledge, that it's possible she learn a new fact which would necessarily be non-physical.

But Mary is a scientist after all, let's see her do some science! Suppose she made a *prediction* about what she will experience when viewing the colored television, i.e. predicting all the physical facts that will be true under the initial conditions represented by her viewing color on the screen. Either she will have predicted the *complete set of facts* that is equivalent to the outcome; everything that will happen we she views the colored images, or she will not. Given her knowledge of only those things that are physical or that supervene on the physical, she will be unable to predict anything nonphysical happening, that is to say, that if it were the case that she learned something, that which she learned would inform her about her theory of reality, specifically, that the *final* model of the real world is not equivalent to the *physical* model of the real world. So she makes her prediction, she infers a theorem, and this theorem represents what she thinks will happen when she sees the colored images on the television screen. She will only be able to predict a *fully* physical outcome, and whether it is equivalent to the actual outcome or not, she will believe beforehand that the predicted outcome will be a *complete prediction* of what will happen. Now at this point, there are only two consistent possible outcomes:

1. Her prediction is complete, and she learns nothing.
2. Her predictions is incomplete, and she will learn a new fact.

If it is the case that 1. occurs, then there are no facts, physical or nonphysical, to be learned by Mary from watching a color tv screen, and if it is the case that 2. occurs, there must have been at least one nonphysical fact that Mary learned by watching color television. Now let us make clear that her physical theory of reality is *still* a theory of reality first, and a *physical* theory of reality by happenstance, since it includes only physical axioms and theorems. Whether or not it is a true theory of reality depends on if it is complete when including only physical facts. It makes no sense to say that her final and complete physical theory is either final of complete if it turns out there are non-physical facts about reality. For the purely physical theory would be true with regard to some other model, that is *not* the actual world, but is instead some possible world with at least one less fact than the actual world, and including only the physical facts in her theory.

When we consider the possibility that she learns from watching color TV, what effect does this have on her physical theory? If it *informs* her of anything in her physical theory, then it was not final, since there was a non-zero probability on more than one model that could possibly make the theory true. That is to say, no matter if a fact, or information, etc. is physical or nonphysical, a final physical theory meant to represent the real world will be informed by any and all information relevant to the model that makes the theory true or false, that is, nature. So if she learns from watching color tv, she would beforehand believe that she has the final theory of reality, and afterward have been informed that her original theory is made true by a model that isn't reality, and have been given a new piece of the model that *is* reality. But she has *all* physical information; nothing more can inform her about which model is the one that makes her physical theory true. So even after watching the colored television, it makes no sense to say that any nonphysical fact could *inform* her about her physical theory, which is coincidentally her theory of reality. So even if she were to become aware of some nonphysical fact, she would not change her theory of reality. But how does that make sense? What this seems to suggest is that there is a third possible outcome: That her prediction is complete, *and* she will learn a new fact. But that is blatantly contradictory. She cannot both be surprised *and* know what is going to happen. It seems we have reasoned out a contradiction. Either Mary predicts everything true of her future color tv viewing, and learns nothing by viewing color tv for the first time, which is equivalent to saying that her complete physical theory is a complete theory of reality, or, she learns something which is a nonphysical fact, but it does not affect her theory of reality. Let us consider further some consequences of a final physical theory: If nothing can inform us about our theory and the model that makes it true, there cannot exist any unpredicted outcome. If the outcome is not exactly as predicted, then our theory isn't final. So Mary's theory, if truly final, *must* predict the complete set of theorems that constitute the outcome of her watching colored tv. But this does not mean that all the facts must be representative of positive facts. A theory makes as many claims about what *is not and cannot be the case,* as it does what *will* be the case. So think about the negative predictions Mary could make: everything her theory predicts will *have* to come to pass, because her theory is complete and final, and an incorrect prediction means the theory is neither complete nor final, and is so not a possibility. So if we maintain our assumption that a nonphysical fact is possible, meaning there is at least one part of the model representative of the real world that is nonphysical, then a completely physical theory of this same model made true by the real world would have to predict the existence of a nonphysical fact. If we are asked to assume that Mary has a final physical theory, that means there is nothing she can learn that will inform her as to which model makes her theory true, and that so far, everything in her theory has been true of the real world. Remember that a theory is made true by the *simplest model* for which all its axioms and theorems are true. So that means that there can be nothing that escapes Mary's physical theory that is a part of the real world, if it is *precisely* the real world that makes her theory final. Herein lies the reason for our contradiction: it just makes *no sense* to say that a theory is final and complete relative to a model that includes a part that escapes theory. Either the theory is made true by the model, the *whole* model, or it isn't made true by that model at all.

In "The Price of an Ultimate Theory," Nicholas Rescher makes an argument against the possibility of the traditional conception of a final theory. He starts by assuming what is traditionally called *the Principle of Sufficient Reason*. It simply states that everything that is true has a reason for being so, or in Rescher's formulation, for any fact *f*, there is at least one fact *f'* that explains *f*. Next, he goes on to postulate some features of final theory. It would seem that if a theory were final, it would be what Rescher calls *comprehensive*. This is a binary relation between our theory and all facts, where for any fact *f*, the explanation of *f* is a part of our theory, *T**. So that no fact is left unexplained by the axioms and postulates of our final theory. A final theory would also have what he calls *finality*, that is, since our theory *T** "affords" all explanations, and the truth of *T** is itself a fact, then the only thing that could explain *T** were it truly the final theory would be *T** itself. But we normally forbid this sort of explanation; a fact does not explain itself, a principle that Rescher calls *non-circularity*, that there does not exist a fact such that that fact explains itself. So it seems, that we cannot maintain all four of the following propositions as consistent: The Principle of Sufficient Reason, Comprehensiveness of a final theory, Finality of a final theory, and Non-Circularity. At least one of these four must be rejected as false, or we must accept that we cannot have a final theory. If we cannot have a final theory, then Jackson's argument does not prove that physicalism is false, because the first premise in the Hail Mary form, that Mary has the final physical theory, would be false.

I can think of no compelling reason to reject any of the four properties required for a final theory, so it seems to me at least that it would be better to reject the notion that there could even be a final theory. Furthermore, we seemed to be getting a contradiction from the premises of the Hail Mary form of the argument. If we reject any of these four principles, we can maintain the validity of Jackson's the knowledge arguments and the truth of all the premises, but if this were are response we would continue to infer a contradiction. For these two reasons, it seems the best choice to reject the truth of premise one. But let us consider the alternatives.

We could reject premise two, but I *do* feel that it is intuitive that Mary would learn by seeing colored tv for the first time, so I will maintain this intuition if at all possible. Let us try three then. If we reject three, we are asserting that it is not the case that if physicalism is true that the psychophysical conditional is necessary and a priori. P then Q is equivalent by way of material implication to asserting that either Q or not P. The negation of which is equivalent to saying that not Q and not not P by De Morgen's Theorem, and this to not Q and P by double negation. So the negation of premise for is the same as asserting that "it is not the case that the psychophysical conditional is necessary and a priori, *and* physicalism is true". This means that either it is possible for physicalism to be true and that the psychophysical conditional is not necessary but is a priori or that physicalism is true and the psychophysical conditional is necessary but a posteriori. If the former, since not all psychological facts are supervenient on the physical, physicalism can be true while there is simultaneously some fact about other people's mental states that is nonphysical, and so the modus tollens in the Hail Mary form does not hold; Mary could learn something nonphysical and physicalism could be true. But this is contradictory, because physicalism is exactly the assertion that there *is no nonphysical fact.* If the later, the a posteriori nature of the psychophysical conditional means that the knowledge argument is not a valid argument against physicalism; Mary could not know the complete and final physical theory before ever having left her room if at least some of those things supervenient on the physical could not be known a priori. If we reject four, the knowledge argument again loses its validity; it is equivalent to asserting that “there is not at least one nonphysical fact and Mary knows all physical facts” by the same method used on the psychophysical conditional.

So if we reject premise three, we get a contradiction, but if we maintain *all* of our premises, we likewise get a contradiction. We have but three alternatives:

1. Mary does not know a complete physical theory before leaving her room.
2. It is not possible for Mary to learn something new by watching color tv.
3. There is not at least one physical fact and Mary knows all physical facts.

But if you still feel that two is intuitive, as I do, we can say instead that either Mary does not know a complete physical theory before leaving her room, and this is precisely why she can learn by seeing color for the first time, or that there is not any nonphysical facts, and Mary knows all physical facts before leaving her room, and she learns when she sees color tv for the first time. But this exactly the conjunction that generated a contradiction earlier! We then have no choice but to accept that either if Mary learns from seeing color for the first time, then she does not know all the physical facts beforehand, or we get a contradiction. It therefore seems that Mary cannot both know all physical facts before leaving her room, *and* learn something new by watching color tv for the first time. This is only the case if there are no nonphysical facts to be learned. Therefore, if Mary knows all physical facts, then physicalism is true. Since we cannot have a final theory, it is impossible that Mary know all physical facts, meaning that the rejection of two to try and salvage the argument fails. If Mary learns nothing new from seeing color tv, then she knows a complete theory of reality that is purely physical, and there is not even one nonphysical fact. But since she cannot have a final theory of this type, rejecting two means only that Mary knows beforehand whatever it is she would have learned from watching colored tv, with whatever these facts may be neither necessarily physical nor nonphysical.

Before we even have a knowledge argument, I believe Jackson to have also made a mistake in the understanding of his knowledge intuition. I share his feeling, that were Mary to see color for the first time, it would be totally different from any sort of fact she could learn from a textbook, and she would indeed learn something the first time she smelled a rose no matter how much she knows beforehand about the physiology of smell and the chemical properties of a rose, or whatever. The mistake is when he considers the latter sort of knowledge, and the actual knowledge gained from smelling the rose, to be distinct. The separation of phenomenal knowledge and physical knowledge is a false dichotomy. Phenomenal knowledge *just is* physical knowledge. The false dichotomy here, in the very beginning, is what messes us up in the thought experiment; yes, it *is* intuitive to think that if Mary had never seen color, watching a colored television program would teach her something new. But this new knowledge we only believe to be necessarily nonphysical if we accept the divide between physical and phenomenal knowledge. If they are one and the same, she will learn something new, but in a new *modality* than she is used to, without any need for it to be nonphysical. Were she to know the complete theory of physics and still learn afterwards something new by watching colored tv, it would appear to be the case that whatever she learns is nonphysical; but a final theory has been demonstrated to be impossible, and only if she knows a final theory does her learning from colored tv point to anything nonphysical, otherwise, she just learned something new that was physical, but of a new modality of experience.

In conclusion, we might sum up my objection to Jackson’s thought experiment thusly: A theory is always a theory *of* something. What makes a theory true is the model equivalent to that thing of which we are trying to theorize. To say that we have a final theory, is to say we know everything there is to know about that which we are trying to theorize. Saying that the theory is a x-theory of y, is irrelevant to whether or not it is a *final* theory of y. If it is final, nothing can change it, we know all the relevant information. What Jackson gets wrong in his thought experiments is that he trusts his intuitions too readily, something we have all been guilty of. If anything, Mary can teach us one thing: that we must think and speak carefully and skeptically about even those things we may trust most, our intuitive introspections.

Works Cited

Churchland, Paul. "Reduction, Qualia, and the Direct Introspection of Brain States" *The Journal of Philosophy*, vol. 82, no. 1, pp. 8-28.

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

---. "Mind and Illusion." reprinted in Minds and Persons: Royal Institute of Philosophy Supplement 53, A. O'Hear, editor, Cambridge UP, 2003, pp. 251-271.

---. "What Mary Didn't Know." *Journal of Philosophy*, vol. 83, no. 5 (May 1986), pp. 291-295. Rpt. in *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*, Peter Ludlow, Yujin Nagasawa, and Daniel Stoljar, editors, MIT Press, 2004, pp. 51-56.

Ludlow, Peter, Yujin Nagasawa, and Daniel Stoljar, editors. *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*, MIT Press, 2004.

Nagel, Thomas. "What Is It Like to Be a Bat?" *The Philosophical Review*, vol. 83, no. 4, (October 1974), pp. 435-450.

Rescher, Nicholas. "The Price of an Ultimate Theory." *Nature and Understanding: The Metaphysics and Method of Science*, Oxford UP, 2003.

Shannon, Claude E. and Warren Weaver. *The Mathematical Theory of Communication*. U of Illinois P, 1949.

Stoljar, Daniel, and Yujin Nagasawa. "Introduction." *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*, MIT Press, 2004, pp. 1-36.

Swan, Liz Stillwaggon. "Galileo's Falling Bodies." *Just the Arguments: 100 of the Most Important Arguments in Western Philosophy*, Michael Bruce and Steven Barbone, editors, Blackwell, 2011, pp. 346-347.

Van Fraassen, Bas. “On the Extension of Beth’s Semantics of Physical Theories.” *Philosophy of Science*, vol. 37 no. 3, 1970, pp. 325–339.

---. *Laws and Symmetry*, New York: Oxford University Press, 1989.