

## The Cogito and the Metaphysics of Mind

*Abstract:* Descartes used the cogito to make two points: the epistemological point that introspection affords us absolute certainty of our existence, and the metaphysical point that subjects are thinking things logically distinct from bodies. Most philosophers accept Descartes's epistemological claim but reject his metaphysical claim. I argue that we cannot do this: if the cogito works, then subjects are non-physical. Although I refrain from endorsing an argument for dualism based on this conditional, I discuss how such an argument would differ from the conceivability arguments pursued by Descartes in the Sixth Meditation and by contemporary philosophers. Unlike those arguments, this argument would not be refuted by the discovery of a posteriori identities between physical and phenomenological properties. In other words, it is possible to argue for substance dualism *even if* phenomenal properties are physical properties.

That there is an epistemological difference between the mental and the physical is well-known. Introspection readily generates knowledge of one's own conscious experience, but fails to yield evidence for the existence of anything physical. Conversely, empirical investigation delivers knowledge of physical properties, but neither finds nor requires us to posit conscious experience. In recent decades, a series of neo-Cartesian arguments have emerged that rest on this epistemological difference and purport to demonstrate that mind-brain identity is false and that consciousness is not even realized by or supervenient on physical properties. Where Descartes argued he could clearly and distinctly conceive mind and body as existing separately, contemporary anti-physicalists hold that the conceivability of worlds in which actual world correlations between physical and phenomenological properties fail shows that these correlations are contingent rather than logically or metaphysically necessary. Together with Descartes, they conclude from conceivability that identity, as well as strong supervenience, is false.<sup>1</sup> If the argument of this paper is correct, however, then there is an argument for dualism that arises from the epistemological distinction, is grounded in the Meditations, and is yet distinct from the

conceivability arguments pursued both by Descartes and contemporary anti-physicalists. Furthermore, the argument is immune to the standard objections to conceivability arguments: its conclusion follows even if there are a posteriori identities between physical and phenomenal properties.

To see this, we will return to where it all began: the Second Meditation. Although Descartes's explication of his argument for dualism does not appear until the Sixth Meditation, after his proof of God's existence assures him that he can trust his clear and distinct perception of the possibility of the separate existence of mind and body, the seed of that argument is in the Second Meditation.<sup>2</sup> Descartes's claim there is that he has found a belief that is immune to error: no matter what else may be true – and, in particular, no matter how the world might be physically – his conscious experience assures him that he is conscious, or thinking, and that he exists. Although the cogito has attracted more attention as an epistemological claim regarding what is certain or immune to doubt and hence an appropriate foundation for a particular kind of epistemological project, it is no less a metaphysical claim, for it purports to demonstrate the existence of a subject.

Few contemporary philosophers think Cartesian dualism is a live option, but not very many doubt that the cogito works in something like the way Descartes said it does. We may not agree on how, exactly, Descartes thought it works, or on how it is best to be understood. Nonetheless, we take it that to know whether we exist, we need not appeal to anything other than our subjective, conscious experience. I shall argue, however, that we are fooling ourselves if we think we can accept this Cartesian conclusion without accepting some form of substance dualism. In particular, I shall argue for a conditional: if

the cogito works, in the sense that one can be certain on the basis of introspection that one exists, then, on pain of accepting a bizarre metaphysics of the physical, subjects are non-physical. Many of us have long found the consequent of this conditional, the metaphysical claim that subjects are non-physical, to be incredible. The problem, however, is that we also take the epistemological claim that is the antecedent of this conditional, that introspection affords certain knowledge of one's own existence, to be indubitable. My thesis is that these metaphysical and epistemological claims stand or fall together: we have to either deny that introspection affords us certain knowledge of our own existence or accept some form of substance dualism.

Before I turn to the argument proper, allow me to give a brief outline of how the paper will proceed. Following a more detailed explanation of what is meant by the antecedent of the conditional ("the cogito works"), I will turn to the argument, which will have the following four steps. In Step 1, I describe two worlds, W1 and W2. W1 is relatively normal physically; i.e., it is a world very much as we take the actual world to be. W2, in contrast, is physically very different – a world such that we would be disinclined to believe that it contains a cognizing subject. In Step 2, I argue that despite the manifest physical differences between the two worlds, they are subjectively indistinguishable, in a sense that I will specify. In Step 3, I argue that, given that they are subjectively indistinguishable, we have to either deny the cogito works in W1, which is tantamount to denying it works in the actual world, or accept that it works in W2, which will entail, on pain of accepting a bizarre metaphysics of the physical, that subjects are non-physical. In the final step of the argument, I consider and respond to a number of objections. In the conclusion of the paper, I discuss how an argument for dualism

developed from this conditional would differ from Descartes's own conceivability argument in the Sixth Meditation and from contemporary conceivability arguments.

As a terminological point, let me note that I will use "subjective experience" and "conscious experience" interchangeably, and in using them I refer to the phenomenological character of mental events generally. I do not mean them to refer only to sensations. Thus I take it that when someone thinks on some particular occasion a cogito or sum thought, there is some subjective experience or phenomenology associated with this, and it is this that I wish to identify by the use of these terms.

What is meant by the antecedent of the conditional, "the cogito works"?<sup>3</sup> Dispute about *how* it works, and about how Descartes took it to work, is legion. As Jaakko Hintikka noted at the start of his 1962 paper "Cogito, Ergo Sum: Inference or Performance?", "After hundreds of discussions of Descartes's famed principle we still do not seem to have any way of expressing his alleged insight in terms that would be general and precise enough to enable us to judge its validity or its relevance to the consequences he claimed to draw from it." This grim assessment of our grip on the logic of the cogito is no less accurate today, despite a wealth of interesting work by Hintikka and others.<sup>4</sup>

Nonetheless, the fact that we do not understand how it works should not obscure the fact that we have a clear enough sense of what it would mean for it to work. All I wish to capture is that to be absolutely certain that I am conscious or that I exist, I need not look to anything outside of my own subjective experience. What Descartes took to save him from the paralyzing doubts of the First Meditation was his discovery, in the Second, that his conscious experience assures him that he is conscious and that he exists. His point is of course general: it is not that there is anything special about his cogito or

sum thoughts, it is that any conscious experience like that which he is having as he thinks those thoughts is sufficient to ensure the existence of a thinking subject and sufficient to assure that subject of its existence. To accept the antecedent of the conditional this paper defends is thus to accept (i) the Cartesian conclusion that we can be absolutely certain that we are conscious and that we exist; and (ii) that our confidence in this conclusion rests on introspection alone.

It is worth emphasizing that the point of the cogito is not merely that our subjective experience affords us knowledge of the fact that we are conscious or that we exist; it is that it affords us absolute certainty that we do. A Cartesian meditator recognizes, so the story goes, that given the subjective experience she is having it could not possibly be false that she is conscious and that she exists. Her subjective experience as she thinks the cogito or sum thought guarantees the truth of the corresponding indexical proposition.

What I have said about the cogito is no doubt familiar, yet to accept the epistemological claim that the cogito works in this familiar way commits one to serious metaphysical consequences. Consider the following world:

**W1:** W1 contains someone who is thinking “I am”.

To make the example clearer, suppose W1 is very much like the actual world and that the subject thinking the sum thought is someone very much like you. For simplicity, let us suppose that W1 contains only one conscious being – the subject thinking that thought. Glance up from this paper, look around the room, think the sum thought and thereby (or so it would seem) assure yourself that “I exist” is true, and imagine that this is pretty much exactly what is happening in W1.

World W2, in contrast, is somewhat harder to describe even though, once described, it will not be hard to imagine. To describe it, I will first describe W1\*, which introduces a modification to W1, and then W1\*\*, which introduces a modification to W1\*. A final modification to W1\*\* will yield W2.

**W1\*:** W1\* is, in many ways, just like W1; the principal difference is that in W1\*, one billion times a second, all the fundamental particles in the world pop out of existence and are replaced instantaneously by type-identical particles. (We are to suppose that, other than the popping into and out of existence, the particles of this world are governed by the same laws as are those in W1. Thus, the worlds are indistinguishable physically except insofar as the token particles in W2 are constantly changing.)

**W1\*\*:** W1\*\* is, in many ways, just like W1\*; the principal difference is that in W1\*\*, there is a one nanosecond delay between the popping out of existence of the fundamental particles and their being replaced by type-identical particles.

**W2:** W2 is, in many ways, just like W1\*\*: the principal difference is that in W2, the delay is not one nanosecond but is instead 1000 years.

W2 is a very strange world, at least by our standards. It is as if everything that is happening in W1 was cut up into nanosecond long temporal slices which were then spread out, so to speak, with an interval of 1000 years in between each time slice. So W2 is very odd, physically, and very much different from W1, but nonetheless it is quite conceivable. (I will address modal concerns about W2 later.)

As different as W2 is from W1 physically, however, the worlds are subjectively indistinguishable. As I understand this concept, this is not the claim that W1 and W2 do not differ in terms of their subjective properties. From the objective perspective, it is easy to see that the distribution of subjective experiences over times is different in W1 and W2. In W1, every second contains one second of conscious experience; in W2, in every thousand year period there is only one nanosecond of conscious experience. It is

important to emphasize, however, that the judgment that the worlds differ in terms of their subjective properties is a judgment from the objective perspective. The situation is very much different when we consider the subjective perspective – the perspective that looks not on conscious experience, from the outside, but is instead of conscious experience, from the inside.<sup>5</sup>

The notion of subjective indistinguishability is very hard to define precisely without begging substantive questions about whether worlds like W2 contain enduring subjects. Nonetheless, the idea is intuitive enough. We have no trouble understanding philosophers who say, for instance, that two molecule-for-molecule duplicate persons in identical environments are subjectively indistinguishable. The claim is that things seem to one just as they seem to the other. Of course, this locution does posit enduring subjects who have the conscious experiences said to be subjectively indistinguishable. We are also familiar, however, with thought experiments involving the notion of subjective indistinguishability in which the existence of an enduring subject is very much in question. Consider, for instance, a thought experiment in which you have to choose between making a long journey by regular space ship and by teletransporter. According to a plausible view of personal identity, you do not emerge from the teletransporter (someone else does), but holding such a view would not normally prevent us from understanding what it would mean to say that situations in which you are really teletransported (and therefore, on the view held, cease to exist) are subjectively indistinguishable from ones in which you step into the transporter and are sent to the destination via superfast conventional means. Similarly, although we should be careful not to let our diction prejudice the question of whether W2 contains an enduring subject,

we should not have trouble getting an adequate grip on what it means to say that worlds W1 and W2 are subjectively indistinguishable.

Why think W1 and W2 are subjectively indistinguishable? First, I should make clear that I am assuming that consciousness supervenes on the physical: necessarily, if world W has at  $t_1$  conscious experience C, then any world that is physically identical at  $t_2$  to W at  $t_1$  has conscious experience C at  $t_2$ . On this assumption, every nanosecond long bit of consciousness that occurs in W1 occurs in W2, given that every nanosecond long subvenient base of consciousness in W1 occurs in W2 as well. Each of the worlds is like this: there are conscious experiences  $C_1, C_2, C_3$ , and so on. Furthermore, although there is a gap in time in W2 between each conscious experience  $C_1 \dots C_n$ , this gap is not something of which that world contains any conscious experience. (The awkward construction is due to the need to avoid begging the question on whether W2 contains an enduring subject by saying that the gap in time is not something of which the subject in that world is ever aware.) From the objective perspective, the gap is clear enough. From the subjective perspective, however, the conscious experiences in W2 follow each other as seamlessly as do the conscious experiences in W1; the stream of consciousness flows as smoothly in one world as in the other.

To see that this is so, suppose you are a brain in a vat and mad scientists hook you up to a sophisticated DVD player such that (i) when 'play' is pressed, your sensory input is exclusively of the movie, which is being piped directly into your consciousness, and (ii) you have no sensory experiences except when 'play' is pressed. Were this so, your total sensory experience on occasions upon which the entire movie is played without pause would be the same as on occasions upon which 'stop' is pressed several times.

(And, obviously, how long 'stop' is pressed would make no difference at all.) Although your total sensory experience would be identical, your total conscious experience would not be, for you would be conscious at times when 'stop' is pressed (and wondering, no doubt, what happened to all your sensory experience). However, suppose the following modification to the described case: in addition to (i) and (ii) above, (iii) is true: you are conscious if and only if 'play' is pressed. In this case, your total conscious experience on occasions on which the entire movie is played without pause will be indistinguishable, from your perspective, from your total conscious experience on occasions on which 'stop' is pressed several times. (And, again, how long 'stop' is pressed makes no difference at all.)

The analogy between this DVD case and the case of W1 and W2 should be obvious. In the DVD case, the having of consciousness supervenes on the state of the DVD player (there is consciousness if and only if 'play' is pressed). In W1 and W2, I am supposing, consciousness supervenes on the physical: there is consciousness if and only if there are fundamental particles that comprise an appropriate subvenient base). If we add that in the DVD case the total character of your conscious experience (not only of your sensory experience) is delivered by the DVD player, then the analogy deepens. The character of your conscious experience at any instant supervenes on the state of the DVD player at that instant, just as in W1 and W2 the character of your conscious experience at any instant supervenes on the physical state of the world at that instant.

There may be a certain intuitive resistance to accepting that W1 and W2 are subjectively indistinguishable because of how peculiar W2 is. First, one might worry about the 1000 years that come in between each bit of consciousness, either because it is so long an interruption or simply because it is an interruption at all. Second, one might

doubt it makes sense to speak of one nanosecond long bits of consciousness. Concerning the first worry, I would ask how periods of non-consciousness could ever make a difference to periods of consciousness. Since non-consciousness is not something anyone could ever experience, non-conscious gaps between conscious episodes are not something of which one could ever be aware, and therefore, whether the non-conscious gap is one nanosecond or 1000 years would not matter. Furthermore, in the example of W1 and W2, we know the non-conscious gap makes no difference because we have stipulated that W2 contains only those subvenient bases of consciousness present in W1; since there is no noticing of a gap in W1, there cannot be any such noticing in W2. In regard to the second worry, which doubts the coherence of one nanosecond long bits of consciousness, two things should be said. First, we should be careful not to confuse the claim that it is hard to imagine a one nanosecond long conscious experience with the claim that it is inconceivable that there are or could be such conscious experiences. The former claim is probably true, at least for beings like us, but it does not entail the latter. Second, if the notion of a one second conscious experience is coherent, then the notion of a one-nanosecond long conscious experience must be, given that a one nanosecond long conscious experience is a temporal concatenation of a billion one-nanosecond long conscious experiences. (To put the point another way, to feel pain for one second, you first have to feel it for one billionth of a second, for two billionths of a second, for three billionths of a second, and so on.)<sup>6</sup>

Although I do not want to belabor the point, perhaps a final example that closely parallels W2 will help. Suppose you are a brain in vat, and that the vat has an on/off switch. When it is on, you are fully conscious, when it is off, you are fully unconscious,

and there are no intermediate positions. Furthermore, the vat is designed such that whenever the switch is flipped on after some period of being off, your brain state picks up exactly where the prior ‘on’ state left off. In these conditions, whether the switch is flipped off and then on is not something that would show up in the character of your conscious experience.<sup>7</sup> If whether or not the switch is thrown makes no difference, then how long the switch is off (one nanosecond or a billion years), and how often it is thrown (merely once or a billion times), makes no difference as well. At this point, the brain in a vat case is analogous to W2 except in two regards: In the W2 case, a numerically different brain subserves each subsequent episode of consciousness and no brain exists during the periods of unconsciousness. Yet suppose the brain in the vat were replaced with a molecule for molecule duplicate whenever the switch was in the ‘off’ position, and that the vat sat around empty for a while whenever this replacement occurred. Neither of these factors (the replacement of the unconscious brain with a type-identical unconscious brain, and the vat’s sitting around empty for a while whenever this exchange is executed) would make any difference to the character of the conscious experiences had when the switch was moved to the ‘on’ position.<sup>8</sup>

If W1 and W2 are subjectively indistinguishable, in the sense I have indicated, then it follows that if the cogito works in W1, then it works in W2 as well. Let us call whatever it is that the subject in W1 experiences as it thinks the sum thought  $SE_{sum}$ .<sup>9</sup> The claim is that if a token subjective experience of the type  $SE_{sum}$  at W1 guarantees the existence of a subject who thinks a sum thought, which has to be true if one can be *certain* on the basis of introspection that one exists,<sup>10</sup> then the presence at W2 of subjective experience that is subjectively indistinguishable from that token of  $SE_{sum}$

guarantees the existence of a subject who thinks a sum thought. This is because any subjective experience that is subjectively indistinguishable from a token of the type  $SE_{sum}$  is itself a token of the type  $SE_{sum}$ . To put the point another way, if W1 and W2 are subjectively indistinguishable, then introspection cannot reveal whether W1 or W2 is actual (or whether the world that is actual is like W1 or W2). Given that to affirm that the cogito works is to affirm that one can be certain, in virtue of one's subjective experience, that one exists, to affirm the cogito works is to affirm that one can be certain one exists regardless of whether W1 or W2 is actual. Importantly, the point is not merely that one need not *know* whether W1 or W2 is actual to know that one exists. It is that whether or not W1 or W2 is actual, one knows and is certain that one exists.

The catch, of course, is that if the cogito works in W2, then W2 must contain a subject. Moreover, it must contain a subject who thinks a proposition ("I exist") and thereby assures itself of its own existence. But when does this subject exist, and what is it?

The first question constrains the answer to the second. If W2 contains a subject who is thinking a sum thought, then this subject must exist at every instant at which the subjective experiences that comprise the thinking of the sum thought (or the subjective aspect of the thinking of the sum thought) do. It follows that either the subject must exist at times when no fundamental particles exist, or the subject pops out of and into existence many, many times, or the subject is a temporally scattered whole. If we assume that in W1, the thinking of the cogito thought takes 1 second, then the thinking of that thought in W2 takes 1000 billion years, in which case either a subject endures for 1000 billion years, even though in that 1000 billion years matter exists for only a total of 1 second, or the

subject pops into and out of existence every 1000 years, a billion times, existing for a grand total of one second, or the subject is temporally scattered over 1000 billion years, with 1000 years separating each of its temporal parts. If the cogito works, then one of these options has to be right. I do not know which it would be, but in none is the subject something we should be happy to call physical.

In regard to the first option, it seems a contradiction to hold that some physical thing can exist when no fundamental particles do.<sup>11</sup> In regard to the second option, it seems very hard to accept that one and the same physical object can pop out of existence and then back into existence 1000 years later, made of different token particles. There are two reasons this is implausible. First, there seems to be a general principle against anything going out of and coming back into existence – the intuition that one and the same object cannot have two distinct origins of existence. Second, even if we accept that some entities can go out of and come back into existence, in the case at hand, in which there is a total change in token particles and no temporal or causal continuity, it is very hard to see in virtue of what a physical object present at one nanosecond is numerically identical to the physical object present at any other nanosecond. Finally, in regard to the third option, it seems hard not to balk at the thought that W2 contains a physical object with temporally scattered temporal parts, for there seems to be nothing in virtue of which the various temporally displaced collections of matter are parts comprising a whole. Not everyone will find this option implausible, but most of us are hesitant to embrace a freewheeling universalism. On pain of accepting one of these options as a reasonable metaphysics of the physical, however, we have to either accept that subjects are non-

physical or deny that the subjective experience we have as we think a sum thought guarantees the existence of a subject.

If that subjective experience does not guarantee the existence of a subject, however, then we cannot be certain on the basis of introspection that we exist. We can think the sum or cogito thought, but as we do we have to grant that for all the subjective experiences we are having the world could be such that no one exists and no one is thinking. If the world is in fact like W1, then we would be genuinely thinking that no one exists and that no one is thinking. If the world is in fact like W2, however, then no one would be thinking that no one exists and that no one is thinking – it would merely seem, to a subjective perspective that no subject occupies, that someone is.

I am not arguing that subjects are in fact non-physical or, conversely, that our confidence in the cogito is mistaken. I am arguing only for a conditional: if introspection affords certain knowledge of one's own existence, then subjects are non-physical. I do not know how to respond to this conditional, for it is no easier to deny the antecedent than it is to accept the consequent. When I introspect it seems, as Descartes rightly noticed, that I can be absolutely certain that I exist. Yet how could I possibly be non-physical? How could anything be only a thinking thing?

Before I discuss how the present argument differs from the conceivability arguments offered by Descartes and by contemporary anti-physicalists, I will consider a few of avenues of objection.

A first objection might be to the metaphysical possibility of W2. Perhaps it is not metaphysically possible for fundamental particles to pop out of and into existence, or for a world to be empty of such particles, or for there to be time in such a world. I do not see

any reason to suppose W2 is metaphysically impossible,<sup>12</sup> but it is not an issue we need to resolve, for the thought experiment can be amended to avoid the objection and yet achieve the same objective. Consider world W2\*:

**W2\*:** W2\* is just like W2 except that instead of popping out of existence, the fundamental particles combine to form a few large blocks of cement. 1000 years later, they instantly return to the state they were in before they became blocks of cement, then one nanosecond later they instantly combine to form blocks of cement again, and so on.

W2\*, on the assumption that blocks of cement are not conscious, is subjectively indistinguishable from W1 and thus an argument essentially similar to the one I gave using W2 is possible. To accept the subject in this world is physical, we would have to accept either that physical objects can pop into and out of existence, be such that they take human form, have a speck of conscious experience, then sit around for 1000 years as a block of cement, then take human form for another speck of conscious experience, and so on, or have temporally scattered temporal parts.

We could also consider world W2\*\*:

**W2\*\*:** W2\*\* is just like W2\* except instead of forming large blocks of cement, the fundamental particles simply fly around in various ways for 1000 years, forming anything they please, so long as they never form anything that is the subvenient base of conscious experience. Then, every 1000 years, they form – for exactly one nanosecond – a temporal slice of the cognizer in W1. After the nanosecond is up, they fly around again for another 1000 years before forming the next temporal slice, and so on.

I take it that this world, too, will be subjectively indistinguishable from W1 and that, therefore, if the cogito works in W1 it works in this world as well. Once again, we have to deny the cogito works or accept a metaphysics according to which W2\*\* contains an enduring physical object and some subject identical to that object.

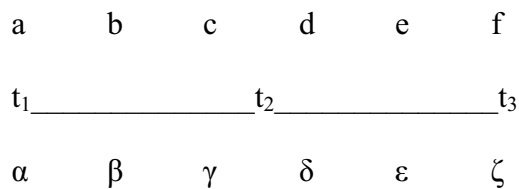
A second objection would begin with the claim that, for us, the sum thought actually happens at an instant and argue that, therefore, to say it works in W2 is to say only that W2 has to contain a subject at that instant at which the sum thought is thought. From here the objection could develop in one of two ways. Perhaps the sum thought occurs in just one of those nanosecond long durations in W2. If this is so, then granting the cogito works does not commit one to the existence of a subject who is present at more than one of those nanosecond long durations in which matter exists. Alternatively, the objection could say that although the thinking of the sum thought occurs at an instant, more than one of those nanosecond subjects is thinking the thought – perhaps several hundred, several thousand, or perhaps even a billion are. If this is right, then again, to say the cogito works in W2 does not commit one to the existence of a subject who exists at more than one of those nanosecond durations. Developed in the first way, the objection says that W2 contains one subject who proves with certainty its own existence; developed in the second way, it says that W2 contains a great many subjects who do.

Although there is something aesthetically appealing, at least to those with a taste for irony, about subjects whose entire lives consist in nothing but the thinking of a sum thought, this view has little else to recommend it. The objection is right that if the sum thought occurs at an instant, then to think this thought these tragic heroes need exist only for an instant. The problem with the objection, however, is that it does not seem right to say that the sum thought happens at an instant. Although I can naturally only speak for myself, the thinking of ‘I exist’ definitely seems to have temporal extension. For one thing, I seem to think the ‘I’ part before the ‘exist’ part.<sup>13</sup> As I think that thought, I seem to be conscious at every moment, but the thinking of that thought, or the apprehension of

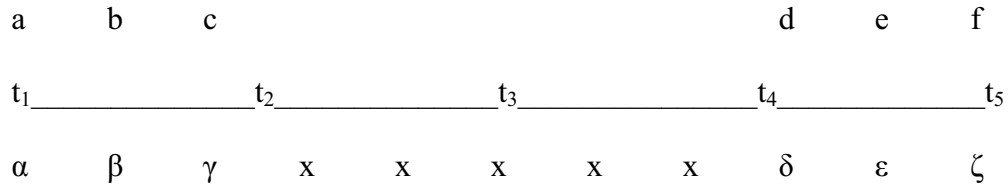
that judgment, seems to bridge these moments. It is articulated, or has parts, in much the way my grasp of the thought this sentence expresses does, and those parts seem to be temporally ordered. Furthermore, insofar as I am able to frame the judgment that I exist without a mental mouthing of the words, I am able to hold this awareness of my existence before my mind, as it were, for at least a few moments.

A third objection would insist that something must be wrong with saying that W1 and W2 are subjectively indistinguishable. This objection is similar to the two objections I considered in my exposition of the argument, which denied the possibility of nanosecond long conscious experiences and worried about how the many 1000 years gaps could make no difference to conscious experience. Although I think I have adequately answered those objections, I do recognize that there is something very strange about the claim that W1 and W2 are subjectively indistinguishable. In light of this, I think it is appropriate to offer a simpler version of the argument which achieves the same objective but does not require that we think about nanosecond long durations of consciousness separated by 1000 years.

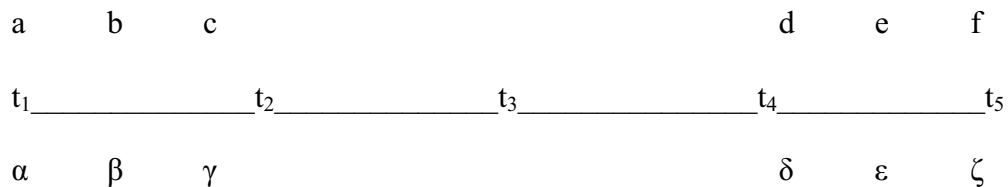
All that is required is that we imagine a world in which halfway through the thinking of the sum thought, so to speak, the physical matter in the world undergoes the kind of change that happens in W2. Suppose we represent some actual world thinking of the sum thought as:



where the upper line represents the subjective experience and the lower line represents the corresponding brain state. Let us suppose further that the  $t_1$  to  $t_3$  duration is one second – it does not really matter how long it is, but thinking of it as one second will be helpful. Now imagine a world like the following:



Again, the upper line represents subjective experience and the lower line represents the corresponding brain state. The x's that occur between  $t_2$  and  $t_4$  are brain states that do not subserve consciousness. In this case, we are not supposing that the second world is one in which the thinking of the cogito thought is sliced up, so to speak, into a billion pieces, but rather that it is simply split in two. In this simplified version of the argument, it is perhaps easier to see that one could not tell whether one was in the first world or the second world, or, to put the point another way, one could not tell, from introspection, whether when one has the subjective experience of the cogito thought, what is happening is like the first world or the second. Finally, we can imagine a third world:



Here we have simply gotten rid of the brain states that fail to subserve consciousness – the world is, we can suppose, empty between  $t_2$  and  $t_4$ , or at least empty of anything that resembles a brain. Those brain states were not making any difference to conscious experience anyways, so their absence makes no difference either. Whether the actual world is like the first, second or third of these worlds when we think a cogito thought is thus not something that the subjective experiences we have as we think that thought reveal. Thus, unless the cogito works in the third world just as in the first, the subjective experiences we have as we think a cogito thought cannot afford us certainty of our existence. Yet it is no easier to grant the third of these simplified worlds contains a physical object identical to the subject thinking the cogito thought than it is to grant the more complex world W2 does. In both worlds, the interruption rules out, on pain of a bizarre metaphysics of the physical, that one and the same physical object exists both before and after the interruption.

Having considered these objections, I will turn to the question of how an argument for dualism developed from the conditional this paper defends would differ from the conceivability arguments advanced by Descartes in the Sixth Meditation and by contemporary anti-physicalists. The argument is of course obvious:

1. if the cogito works, then subjects are non-physical.
2. the cogito works.
3. therefore, subjects are non-physical.

In considering the differences between this argument and conceivability arguments, I will suppose that someone advancing this argument would support the first premise as I have in this paper and the second premise by appeal to intuition.

What is Descartes's argument? He writes:

First, I know that everything which I clearly and distinctly understand is capable of being created by God so as to correspond exactly with my understanding of it. Hence the fact that I can clearly and distinctly understand one thing apart from another is enough to make me certain that the two are distinct, since they are capable of being separated by God....It is true that I may have (or, to anticipate, that I certainly have) a body that is very closely joined to me. But nevertheless, on the one hand I have a clear and distinct idea of myself, in so far as I am simply a thinking, non-extended thing; and on the other hand I have a distinct idea of body, in so far as this is simply an extended, non-thinking thing. And accordingly, it is certain that I am really distinct from my body, and can exist without it. [ATVII 78]

Descartes's argument rests on two premises: that God can bring about whatever Descartes can clearly and distinctly understand and that he can in fact clearly and distinctly understand mind and body existing apart from one another. (Descartes interprets the cogito insight in the Second Meditation to ground the second premise.)<sup>14</sup>

Modern anti-physicalist arguments from conceivability, although they are framed and developed differently, typically involve a similar intuition: the conceivability of worlds physically identical to the actual world, but phenomenologically different, demonstrates the possibility of worlds in which identity, realization or supervenience fails (the contemporary focus is, of course, usually on phenomenological properties rather than on mind as a kind of substance). As Katalin Balog recently put it, the argument from conceivability

begins with the premise that we can conceive of *any* physical or functional facts obtaining without there being any phenomenal experience at all. This is sometimes expressed by saying that zombies (that is, beings that are our physical and functional duplicates, but that possess no phenomenal experiences) are *conceivable*. From this assertion of conceivability it is inferred that zombies are genuinely possible. [498]

Again, there are two premises, and they closely parallel Descartes's: conceivability entails possibility, and zombies (or worlds physically identical to the actual world but phenomenologically different) are conceivable.

In examining the differences between the argument one could develop from the conditional I have defended and arguments from conceivability, there are two factors of both the Cartesian argument and its contemporary variant that I wish to isolate. First, the conceivability alleged to demonstrate possibility rests on a putative failure of a priori entailment between the way the world is physically at a time and the way the world is mentally, or phenomenologically, at a time.<sup>15</sup> Second, the conclusion is not merely that identity fails, but that supervenience does as well: worlds can be physically identical yet phenomenologically different. Both points are also found in Descartes. In regard to conceivability resting on a failure of a priori entailment, he says, for instance, “Let this supposition [that his body or any thin vapour which permeates it is nothing] stand; for all that I am still something.” [ATVII 27] He makes clear that he takes the test of possibility to be whether he can conceive it without contradiction. [ATVII 71] And he says that “the concept of body includes nothing at all which belongs to the mind, and the concept of mind includes nothing at all which belongs to the body.” [ATVII 225] In each case, he points to an a priori gap between propositions or concepts and concludes the second point on the basis of it – that, as he puts it, “the mind can, at least through the power of God, exist without the body; and similarly the body can exist apart from the mind.” [ATVII 170]

What is different about the argument currently considered, however, is that although it clearly arises from the epistemological difference, it differs from the conceivability argument on both of these points. Nothing said in defense of the first premise, and in particular not the conceivability of W2 and the argument that W1 and W2 are subjectively indistinguishable, rests on a failure of a priori entailment between

physical and phenomenological properties. Further, in defending that premise I assumed, rather than denied, that phenomenological properties supervene on the physical.

Let us consider this second point first. Both Descartes and contemporary proponents of conceivability take the conclusion of their arguments to be that supervenience fails. This is a different claim than that type identity is false. (A pre-established harmony view might hold, for instance, that mind and body are non-identical but that, nonetheless, supervenience obtains).<sup>16</sup> The current argument, in contrast, rests on the assumption of supervenience: this was a premise in the argument that W1 and W2 are subjectively indistinguishable. They are subjectively indistinguishable because they contain the same subjective experiences given that they contain the same subvenient bases of consciousness (the only difference being, of course, that from the objective perspective these subvenient bases, and therefore, subjective experiences, are temporally displaced from one another, a displacement which is imperceptible, or makes no difference, from the subjective perspective). This argument from the cogito, therefore, would conclude non-identity but not that supervenience fails.<sup>17</sup>

Turning to the second point, suppose, contra proponents of conceivability arguments, that via some mix of neuroscience and conceptual evolution we get to the point where we can deduce a priori the phenomenological history of the universe from a complete physical history of the universe, and vice versa. Alternatively, suppose that we get to the point where a complete physical history of the universe simply contains, as part of it, the phenomenological story.<sup>18</sup> If this were our understanding of the physical and the phenomenological, we could, from knowledge of what the world is like physically at a time, deduce a priori what the world is like phenomenologically at a time, and vice versa.

Standard conceivability arguments would be a wash – zombies and zombie worlds would be inconceivable, and we would know that they, along with worlds with disembodied minds or inverted qualia, are metaphysically impossible. But this would change nothing about the argument we are considering. Consider again the defense of the first premise, which I assume is the controversial one. Being able to deduce, from knowledge of the physical at *t*, the phenomenological at *t*, and vice versa, would not change whether *W2* is conceivable, nor whether *W2* is subjectively indistinguishable from *W1*. We would be able to consider the physical state of any possible world, including *W1* and *W2*, at any time, and deduce the phenomenological character of the world at that time. Similarly, if we knew any possible world, including *W1* and *W2*, contained a certain phenomenological experience at a time, we would be able to deduce the physical character of the world at that time. But even if we had this ability, introspection would not reveal whether the actual world is like *W1* or like *W2*. The argument for the subjectively indistinguishability of those worlds would have been no different if we had supposed that the subject thinking “I am” in *W1* also knew every a posteriori necessary proposition expressing supervenience or identity relations between physical and phenomenological properties. Whatever differences this additional knowledge would make to that subject’s subjective experience or consciousness would show up in the subjective experiences in *W2*, given how that world has been defined.

In light of these differences between the argument for dualism that arises from the conditional I have defended and standard conceivability arguments, standard responses to those arguments do not work against it. Consider the claim that propositions expressing identities between physical and phenomenological properties are a posteriori necessary.<sup>19</sup>

Mind-brain identities cannot be discovered a priori, it is granted, but neither can the identity of water and H<sub>2</sub>O. We cannot deduce phenomenology from the physical, but we are no better at deducing, from our concept of water, that water is H<sub>2</sub>O. In neither case, so the objection goes, should this lead us to doubt the metaphysical necessity of the identity. The issues here quickly become sophisticated and subtle, but there is no need to explore the debate between physicalists and anti-physicalists in detail. We need only notice that this physicalist response is irrelevant to the argument we are considering. That argument neither rests on the failure of a priori entailment nor denies that correlations between the physical and the phenomenological hold with metaphysical necessity. It is a different problem.

There is much about the argument for the conditional, and the argument for dualism one could base upon it, that I cannot discuss here. I want to close the paper, however, with two points. First, most philosophers take the alleged irreducibility of qualia to be the biggest challenge to physicalism and so the contemporary focus is on phenomenological properties. Yet however this debate turns out, there is still an issue, little discussed of late, concerning substance, that which, as Descartes puts it, “doubts, understands, affirms, denies, is willing, is unwilling, and also imagines and has sensory perceptions.” [ATVII 28] This is because it is one thing to ask whether phenomenal properties are physical properties, and quite another to ask whether thinking things are physical things. To put the point another way, consideration of the differences between standard conceivability arguments and the argument from the conditional shows us that even if physicalists are right that statements expressing identity relations between the

physical and the phenomenological are a posteriori necessary, substance dualism might be true.

Finally, I want to emphasize that although it may sound in the final pages of this paper as if I am advocating substance dualism, this is not my intention. The central argument of the paper is only that we have to choose between denying the cogito works and accepting that subjects are non-physical. This is not a choice I am happy to make. That introspection affords certain knowledge of one's existence seems indubitable; I find it absurd to suppose otherwise. Yet I find it no easier to accept that subjects are non-physical, for this is a claim that I can barely understand. Thus I take the upshot of the paper to be chiefly that there is a problem here. Unfortunately I have little to offer now by way of a solution.

---

<sup>1</sup> Descartes, of course, did not talk about supervenience as such. For archetypal contemporary arguments from conceivability, see Kripke 1972 and Chalmers 1996. Although less often acknowledged as such, Putnam's multiple realizability argument [1967] is a version of the conceivability argument insofar as the premise that beings with quite different physical properties could nonetheless instantiate some of our mental properties is supported by appeal to conceivability.

<sup>2</sup> For a good discussion of the relation between the Second and Sixth Meditations, see Wilson 1976.

<sup>3</sup> My use of the term 'the cogito' is not shorthand for Descartes's famous dictum "cogito, ergo sum". Rather, I am referring to a more general insight, dating back at least to Augustine, that Descartes expresses in various, sometimes competing, ways (and about which he affirms different things, at times denying it is a syllogism, at other times claiming it is, and so on). The 'cogito, ergo sum' expression occurs in Principles of Philosophy [I.7: ATVIII A 7]; his fullest treatment of the issue is in the Second Meditation, where he expresses the insight as "this proposition, *I am, I exist*, is necessarily true whenever it is put forward by me or conceived in my mind." [ATVII 25] Here as elsewhere in the paper, I cite Descartes by volume and page number in *Oeuvres de Descartes*, eds. Ch. Adam and P. Tannery (revised edition, Paris: Vrin/CNRS 1964-76), relying on Cottingham, John, Robert Stoothoff, and Dugald Murdoch, trans. 1985 *The Philosophical Writings of Descartes*. Vols. I and II. Cambridge: Cambridge University Press, for translations.

<sup>4</sup> The literature is far too extensive to reference adequately. It continues to grow, as Husain Sarkar's recent book length treatment of the issue testifies (*Descartes' Cogito: Saved from the Great Shipwreck*, Cambridge, 2003).

<sup>5</sup> Another way of putting the claim is this: W1 and W2 are objectively distinguishable in terms of their subjective properties, but they are not subjectively distinguishable in terms of their subjective properties.

<sup>6</sup> See [Author, \_\_\_\_\_] for a more developed defense of this claim. Briefly, one argument is this: assume for reductio that conscious experience has a minimum duration of one second. Suppose Smith and Jones are identical in all respects and that (i) Jones has conscious experience from noon until one second past noon, and (ii) Smith goes out of existence at half a second past noon. Given what we have supposed for reductio,

---

Smith never has conscious experience. Yet it would then follow that between noon and half a second past noon Jones but not Smith is conscious, despite the fact the two are throughout that duration identical in all respects. Since this is absurd, the supposition that conscious experience has a minimum duration of one second is false. A parallel argument could be run using any duration and so it follows that conscious experience has no minimum duration.

<sup>7</sup> One might be tempted to think you would notice a blip, but this is not right. You would only notice a blip if those in charge of the vat programmed your brain to be in whatever state subserves such a noticing.

<sup>8</sup> Although I moved from W1 to W2 via a description of two other worlds because I thought that was the easiest way to describe W2, it is worth asking whether there is any intuitive resistance to thinking that W1, W1\*, and W1\*\* are subjectively indistinguishable (see p. 5-6). If there is not, then we should ask what the relevant difference between W1\*\* and W2 could be such to account for a subjective difference between the worlds. The only difference between the worlds is that rather than a gap of one nanosecond, as in W1\*\*, there is a 1000 year gap in W2. But it seems inconceivable that this could make any subjective difference – just imagine that the extra 1000 years is inserted, so to speak, in the middle of the non-conscious nanosecond.

<sup>9</sup> SE<sub>sum</sub> is the experience from the subjective point of view, or as experienced from the subjective point of view. See Nagel 1974.

<sup>10</sup> The Cartesian meditator recognizes that the proposition “I exist” could not be false given the subjective experience, SE<sub>sum</sub>, that accompanies the thinking of the thought “I exist”, and takes it that whether it is indeed put forward is revealed by conscious experience. See ATVII 25.

<sup>11</sup> It does not help, of course, to suppose a physics of gunk. The point is that if the first option is true, and subjects are physical, then a physical object exists when nothing physical exists.

<sup>12</sup> See in particular Shoemaker 1969, p. 369 for an argument defending the possibility of time without change.

<sup>13</sup> Although I cannot think in Latin, I doubt doing so would change anything of significance – to think sum still seems to have temporal extension, at least in so far as one thinks the thought by thinking the word. Besides, it would be an odd view that held the cogito works if you speak Latin but not if you speak English.

<sup>14</sup> For Descartes’s professed view of the relation between the Second and Sixth Meditation, see his Fourth Replies [ATVII 226].

<sup>15</sup> Examples in the literature are ubiquitous. For instance, David Chalmers: “The facts about experience cannot be an automatic consequence of any physical account, as it is conceptually coherent that any given process could exist without experience. Experience may *arise* from the physical, but it is not *entailed* by the physical.” [1995] The absence of a priori entailment is usually accepted even by those who do not believe it demonstrates metaphysical possibility. Brian Loar says, for instance, “Knowing that p, if p is conceived in physical-functional terms, never a priori suffices for knowing that q, if q is conceived in psychofunctional terms.... This has been denied, but it seems to me correct; it is the fundamental anti-physicalist intuition and I accept it.” [83] Nagel says: “I believe that there is a necessary connection in both directions between the physical and the mental, but that it cannot be discovered a priori.” [1998, 337] And Joe Levine confesses: I am prepared to maintain that materialism must be true, though for the life of me I don’t see how.” [475] See especially Chalmers 1996 for arguments supporting the view that there is no a priori entailment from the physical to the mental.

<sup>16</sup> As Jaegwon Kim points out, “mind-body supervenience is consistent with a host of classic positions on the mind-body problem; it is in fact a shared commitment of many mutually exclusive mind-body theories”. [189]

<sup>17</sup> Furthermore, its conclusion is that thinking subjects are not identical to bodies, not that phenomenological properties are not identical to physical properties.

<sup>18</sup> The fact that we cannot do this makes it hard to imagine what it would be like to be able to do this or what the physical history or description of the world would have to be to allow this. See Nagel 1974 and 1998 for interesting discussions of this issue.

<sup>19</sup> See, for instance, Hill and McLaughlin 1999 “It is false that if one can in principle conceive that P, then it is logically possible that P; ... Given psychophysical identities, it is an a posteriori fact that any physical duplicate of our world is exactly like ours in respect of positive facts about sensory states”. [46]

## Bibliography

Balog, Katalin. 1999. "Conceivability, Possibility, and the Mind-Body Problem", *Philosophical Review* 108: 497-528.

Chalmers, David. 1996. *The Conscious Mind: In Search of a Fundamental Theory*, Oxford: Oxford University Press.

Hill, Christopher and Brian McLaughlin. 1999. "There are Fewer Things in Reality Than Are Dreamt of in Chalmers's Philosophy", *Philosophy and Phenomenological Research* 59: 446-454.

Hintikka, Jaakko. 1962. "Cogito, Ergo Sum: Inference or Performance?", *Philosophical Review* 71: 3-32.

Jackson, Frank. 1982. "Epiphenomenal Qualia", *Philosophical Quarterly* 32: 127-36.

- 1998. *From Metaphysics to Ethics*. New York: Oxford University Press.

Kim, Jaegwon. 1997. "The Mind-Body Problem: Taking Stock After Forty Years", *Noûs* 31: 185-207. (Supplement: *Philosophical Perspectives 11: Mind, Causation and World*.)

Levine, Joe. 1998. "Conceivability and the Metaphysics of Mind", *Noûs* 32: 449-480.

Nagel, Thomas. 1974. "What Is It Like To Be A Bat?" *Philosophical Review* 88: 435-50.

- 1998. "Conceiving the Impossible and the Mind-Body Problem", *Philosophy*, 285: 337-352.

Putnam, Hilary. 1967. "The Nature of Mental States." Originally appeared as "Psychological Predicates" in W. H. Capitan and D. D. Merrill, eds., *Art, Mind and Religion*. Pittsburgh: University of Pittsburgh Press. Reprinted (1975) in Putnam, *Mind, Language and Reality: Philosophical Papers, Volume 2*. Cambridge: Cambridge University Press, 429-40.

Shoemaker, Sydney. 1969. "Time without Change," *The Journal of Philosophy* 66: 363-381.

Wilson, Margaret. "Descartes: The Epistemological Argument for Mind-Body Distinctness", *Noûs* 101: 3-15.