Why physicalism seems to be (and is) incompatible with intentionality

Richard Johns

Published in *Acta Analytica*, Volume 35, pages 493–505, (2020).

This is the accepted manuscript version. The final publication is available at https://link.springer.com/article/10.1007/s12136-020-00423-3 (paywalled)

Abstract

There is a long history of philosophical intuition that the human mind must be more than physical or mechanical. I argue that this intuition arises from the perfect "transparency" of physical and mechanical states, in the sense that such states have no obscure or occult elements, but are fully intelligible in (usually) mathematical terms. In the paper, I derive a contradiction from the claim that such a physical system has genuine intentionality, comparable with an intelligent human. The contradiction arises from the fact that, according to physicalism, the physical properties of a brain state determine the narrow propositional content of any conscious thought occurring in that state. This fact allows a physical property of brain states to be defined using Cantor's diagonal construction, and then a contradiction results if a physical system is assumed to form thoughts involving that property.

Physicalism, and the materialist view that preceded it, face a common criticism: we have a strong intuition that nothing fully physical or mechanical could possibly be a mind. In this paper I will investigate the source of this intuition and present a new argument in support of it. My argument differs from previous arguments in this vein, in that it aims to show that physicalism is inconsistent with intentionality, rather than consciousness or phenomenal qualities.

The intuition that nothing physical or mechanical could be a mind is driven, I will argue, by the fact that such systems have states that are entirely clear and intelligible, having precise mathematical characterisations. We shall therefore begin by recounting the historical origins of this idea, and its role in arguments against physicalism.

1. Mechanism, Physicalism, and the Mind

The pioneers of the mechanical philosophy, philosophers such as Boyle, Descartes and Gassendi, saw no prospect of understanding everything in mechanical terms. Some tricky properties, such as colours, were declared to exist only in the mind of the observer. Of course that merely moved the problem rather than solving it, as one must still give an account of the mind. Most followed Descartes in seeing the mind as essentially non-mechanical, but a few materialists (notably Hobbes) dared to suggest that the mind itself is part of the mechanical world.

Nowadays the strict mechanical philosophy, where the physical world consists of particles having geometrical qualities only, is not tenable. Even in Newton's day the gravitational force was problematic for this view, as the force appeared to act magically at a distance without any contact between the particles affected. More recently physicists have discovered that the physical world is subtler than Descartes could have imagined. Fields are needed, as well as particles, and the mathematics needed to describe such fields is highly abstract, and very remote from Euclidean geometry. Thus today there are no mechanists of the 17th century sort. Nevertheless, the spirit of materialism (so to speak) is still alive and well in the 21st century, and goes by the name 'physicalism'. Quantised massless fields may not be mechanical, but they are still fully physical.

The notion of a 'physical' property is not nearly so sharply defined as a mechanical one, however. If a physical property doesn't have to be geometrical, then what are its limits? There is unfortunately no generally-accepted understanding of 'physical property', but in this paper I shall assume that physical properties are all conceptually *transparent*, which here roughly means that they can be perfectly comprehended by minds of sufficient ability (especially mathematical ability). This notion of transparency is discussed in more detail in the next section.

It is significant that physicalism, despite its great differences from the mechanical philosophy, is similarly attacked for an alleged inability to account for the mind. In my view this similarity between mechanism and physicalism results from the fact that both assert the complete transparency of the world. I believe that it is the transparency of physical properties that grounds the intuition (whether correct or not) that intentional states and phenomenal experiences are more than "merely" physical.

2. Physicalism and Transparent Qualities

The pioneers of the mechanical philosophy had no grounds to think that mechanical qualities would be adequate to explain all phenomena. They nevertheless urged that mechanical explanations be sought, where possible, on account of their *clarity*. Boyle (1674), for example, begins his defense of the mechanical philosophy by saying, "And the first thing that recommends it is the clearness and intelligibleness of its principles and explanations". He contrasts corpuscular properties with those used by the Aristotelians and alchemists, since the latter are 'dark', 'obscure', 'occult', etc.

Transparency, in the sense used here, is the same as Boyle's "clearness and intelligibleness". But what does it really amount to? Let us begin by asking why mechanical explanations were so clear and intelligible. Such explanations viewed a system as consisting of particles whose properties were just shape, size and motion – geometrical (and kinematic) properties, in other words. Now geometrical qualities – square, triangle, plane, sphere, etc. – are intelligible in the sense that they are, or exactly correspond to, mental concepts, or abstract objects. According to Plato, for example, these geometrical Forms were accessible only through the mind, not the senses. Being accessible to the mind entails that geometrical qualities like *Sphere* have no occult qualities, i.e. ones that are hidden or beyond our ken. When we define a sphere (in the usual way) we specify the concept entirely, leaving nothing vague or indeterminate.

Whether or not Plato was correct in his realism concerning geometrical and mathematical entities, we can agree with him that these entities are fully intelligible. (If such entities are mere creatures of the mind, for example, then this must also be true.) So we see that the transparency of mechanical systems derives from the intelligibility of mathematical objects.

In that case, however, physical properties are also transparent in exactly the same way, and to the same degree. For, as noted above, when physics gave up on a mechanical understanding of the world, it retreated to other forms of mathematical description. The mechanical models of electromagnetic fields that Maxwell devised were later considered superfluous to a physical understanding—as Hertz famously put it, "Maxwell's theory is Maxwell's equations". Furthermore, no present physical theory claims the existence of any property that cannot be mathematically characterized in some way. How would present-day physicists respond to a claim that a certain property of the electron (say) is 'occult', or not amenable to mathematical

treatment? They would certainly dismiss it as invalid, on the grounds that one could not calculate any observable consequences from the theory.

It must be noted that, while the simple transparent properties discussed so far are all ones that humans can grasp clearly, this need not be true of transparent qualities in general. Since human minds are finite and limited, there may be transparent properties that are simply too complex for us to grasp. Such properties are very different from occult qualities, however, in being fully intelligible in principle, by a mind such as Laplace's demon.

Thus, one necessary condition for a property to count as 'physical' is that it be fully intelligible, or transparent to the mind. Physicalism is then the general idea that 'everything is physical', so that all of reality can be described completely, without any residue, in terms of its physical properties. The usual way to make this idea precise is to say that all properties *supervene* on physical properties, so that two systems cannot differ at all, unless they also differ in their physical properties. There are different kinds of supervenience, however, as the modality involved might be logical, metaphysical, or nomic. Which type of supervenience follows from the idea of physical properties as conceptually transparent?

The point to note here is that, if physicalism is true, then all of reality is conceptually transparent, so that if Laplace's demon were provided with a complete physical description of a system, then nothing would be hidden from it, and so it could infer (a priori) all of the other properties possessed by the system. In other words, physicalism entails that all properties (including mental properties) *logically* supervene on physical properties.[1] There are of course physicalists who espouse 'a posteriori physicalism', and claim that the supervenience relation involved here is metaphysical rather than logical, so that Laplace's demon needs more than a complete physical description of the world to infer all of the mental facts. Such a view clearly

---

[1] This claim needs to be understood carefully, since Laplace's demon may not possess (for example) the concepts 'water' and 'lake', and so may not be able to infer the sentence: "lakes (on earth) are filled with water". Nevertheless, once these terms are defined, in the language of physics, Laplace's demon can infer that the sentence is true. In a similar way, even though the axioms of second-order Peano arithmetic are categorical, and hence logically entail all of the truths of arithmetic, to infer a true sentence about prime numbers one first needs a definition of 'prime number'.

denies the conceptual transparency of the world, however, and thus seems to require that some physical qualities are 'occult', or unintelligible. In my view this means abandoning the core idea of physicalism, which it inherited from the mechanical philosophy.

3. Transparency and the Mind

As stated above, I claim that the apparent inability of mechanical (and physical) explanations to account for the mind results from the *transparency* of mechanical (and physical) processes. The role of transparency in this intuition seems to be crucial to much, and perhaps all, opposition to mechanism and physicalism.

One sees the assumption of transparency at work in "knowledge arguments", i.e. ones that infer the absence of thought in a physical system from our inability to *know* about such thoughts, given a full physical description. Blaise Pascal (1669, no. 72) seems to have given an early example of a knowledge argument when he states, "... if we were simply material, that would prevent us from knowing anything at all, there being nothing so inconceivable as the idea that matter knows itself. We cannot possibly know how it would know itself." The argument here, as brief as it is, involves *our* inability to know how matter could have knowledge. Now why should our inability to know some proposition *p* entail that *p* is false? Pascal is apparently assuming that a 'simply material' being is an open book – it is transparent, or intelligible, so that its properties are readily apparent to us. If a transparent substance were to "know itself", then we humans should be able to see how it knows itself. (Also, Pascal seems to take it for granted that no such knowledge will be forthcoming.)

In a similar vein, Leibniz (1714, Sec. 17) held that "Perception, and that which depends upon it, are inexplicable by mechanical causes, that is to say, by figures and motions." To support this claim Leibniz imagined enlarging an alleged thinking machine to the point that one could walk around inside it, as in a mill. The purpose of the enlargement is not made clear in Leibniz's text, but by making all the parts of the machine clearly visible, I believe it draws attention to the full intelligibility of the machine. In such a transparent system no properties are obscure or mysterious, so that any thoughts that were occurring there would be open to our inspection, yet Leibniz says that we would see only "pieces working upon one another".

The same assumption of the transparency of physical states is also at work in 20th century arguments against physicalism. In Frank Jackson's (1982) knowledge argument, for example, we are invited to consider the case of Mary, a neuroscientist who has never had a colour experience, yet who has complete physical information about the physiology of human colour vision. Jackson clearly supposes that physical qualities are fully intelligible, since he states that, according to physicalism, Mary ought to know *what it is like* to have (say) a red colour experience. If the physiology of vision involved occult qualities of some kind, opaque to the intellect, then how could Mary make an inference of this sort?

Jackson's thought experiment is intended to pump our intuition that the known properties of matter are logically independent of properties of conscious experience. One cannot, for example, define the colour red in terms of geometrical properties, or any other mathematical properties. Thus the overall scheme of arguments such as Pascal's, Leibniz's and Jackson's may be summarized as:

1. The intelligible properties of physical systems are logically independent of mental properties like colour experiences.
2. Physical systems are transparent, so that all their properties are intelligible.

∴ Physical systems do not have mental properties

My argument below is rather different from this pattern, in that I do not use premise 1. Also, my argument is focused on intentionality rather than phenomenal experiences.

I thus agree with the claim of BonJour (2010) that conscious intentional content is a serious problem for physicalism. Referring to Jackson's fictional neurophysiologist, BonJour reports (p. 17) having a very clear intuition that Mary could not deduce the propositional content of a person's conscious intentional state (such as fearing that the train will be late) from a physical description of their brain at that moment. Now, a physicalist who is also an externalist about mental content might agree with this point, since the content of an intentional state will depend on states of affairs outside the head, such as the brain's causal connections to external objects. BonJour's response to this, which seems exactly right, focuses on a person's awareness of their own intentional content. When an ancient person thought about water, for

example, they were certainly thinking about H_2O, yet this was not "part of their conscious, internal grasp of what they are thinking about" (p. 17). BonJour notes that, "All that I normally have any sort of direct access to, if materialism is true, is my own internal physical and physiological states, and thus my conscious understanding of what I am thinking about at a particular moment must be somehow a feature or result of those internal states alone" (p. 18).

Here we will use the term 'conscious intentional content' to refer to a person's "conscious, internal grasp of what they are thinking about". Physicalism is committed to the claim that Laplace's demon (though perhaps not Mary[2]) could infer a person's conscious intentional content from their physical brain state. Let us call this consequence of physicalism the *mind-reading thesis*.[3] As mentioned above, BonJour finds it "utterly clear" (p. 17) that the mind-reading thesis is false, yet he provides no argument to support this intuition. In Section 6, however, it will be shown that the mind-reading thesis leads to contradiction.

4. The Cantor Diagonalisation

The contradiction I derive from physicalism uses the diagonal construction that Cantor developed to prove that even an infinite set, such as the set of natural numbers, has more subsets than members. Generally speaking, this sort of argument rules out certain kinds of correspondence between objects and properties. Suppose that there is a class of objects, all of a certain type, and a class of properties that can apply to objects of that type. Thus, for any object x, and any property P, either $P(x)$ or $\neg P(x)$ will be true. The Cantor

[2] I do not assume that a human being such as Mary could perform these inferences, since they may be too complex.

[3] If it turns out (contrary to my view) that such narrow content does not exist, then the mind-reading thesis can be defined as the claim that Laplace's demon could infer a person's conscious intentional content from their physical brain state, together with all relevant parts of their environment. This alternative definition does not affect the argument.

diagonalization requires two more ingredients: (i) there is a surjective function[4] *f* that maps each object *x* to a property *f*(*x*), and (ii) the properties in the class can involve the function *f* itself. One can then define an object *x* to be 'inclusive', i.e. self-applying, just in case *x* has the property *f*(*x*), and non-inclusive when *x* satisfies ¬*f*(*x*). There is still no contradiction here, unless we make the further assumption that the class of properties includes the property non-inclusive. In that case, non-inclusive will equal *f*(*d*), for some object *d*, and then *d* itself is inclusive if and only if it isn't.

Contradictions involving the Cantor diagonalization are sometimes judged to prove theorems, (e.g. Cantor's own theorem) by the method of *reductio ad absurdum*, but in other cases the contradiction is thought to result from an improper definition. After deriving such a contradiction from physicalism, the question will then arise whether this is a *reductio* of physicalism, or merely the fruit of an improper definition. To get clear on the differences between these cases, I will present one example of each kind.

First, I will prove the rather trivial result that, for any formal language of arithmetic, some properties (or sets) of natural numbers are not expressed by any well-formed formula (with one free variable) of that language. Let *L* be a formal language of arithmetic with a finite number of basic symbols, and where each formula is a finite sequence of these symbols. Gödel showed that the formulas of such a language can be assigned unique code numbers, so we suppose that such an encoding has been chosen for *L*. Then, for any given *L*-formula (with one free variable), it is always meaningful to ask whether it is satisfied by its own Gödel number. For example, the formula that expresses '*z* is an even number' might be $\exists y\ z = 2\times y$, and have the code number 456772, which is even. We can then define a natural number *n* to be *inclusive* (w.r.t. *L*) just in case: (i) *n* is the code number of a formula F with one free variable, and (ii) *n* satisfies F. It is apparent that, while inclusive and non-inclusive are well-defined properties of natural numbers, they are not expressed by any formula of *L*. For suppose that non-inclusive is expressed by the formula D. Then D will have a code number, say *d*, and it is clear that *d* satisfies D if and only if it does not, which is a contradiction.

In this argument, the "objects" are the natural numbers, and the "properties" are properties of natural numbers. The surjection *f* has two components. First, the Gödel encoding maps some natural numbers to formulas of *L*. Second, the semantics of *L* determines the property (or set) of natural numbers expressed by

[4] In other words, every property in the class is the image under *f* of at least one of the objects.

that formula. The contradiction then results from assuming that *L* expresses a certain property of natural numbers that involves the Gödel encoding and semantics of *L*.

One might summarise what is learned from this proof by saying that there are *more* arithmetical properties than *L*-formulas, but this (although true) isn't the whole story. For the proof tells us about a *particular* property that cannot be expressed in *L* – one that involves *L*'s own semantics and Gödel numbering. Thus, even a highly restricted class of properties, with only countably many members, might contain properties that can't be expressed in *L*. The cardinality of the class is not the central issue; what matters is the specific properties it contains.

The argument above proves that, for any formal language *L* of the type described, some *properties* (or sets) of natural numbers are not expressed by any *L*-formula. This reasoning can be very slightly modified, however, to show that some *propositions* about natural numbers are not expressed by any *sentence* of *L*. Corresponding to every *L*-formula F with one free variable there is a sentence, ∃xF(x), i.e. its existential closure. (Note that the Gödel encoding assigns code numbers to *all* formulas, including sentences.) We can then define a natural number *n* to be *inclusive* (w.r.t. *L*) just in case: (i) *n* is the code number of a sentence of the form ∃xF(x), and (ii) *n* satisfies F. This meaning of 'inclusive' is different from the one above (i.e. the set of <u>inclusive</u> natural numbers is different from before) but it yields a contradiction in exactly the same way if we suppose that there is some formula D that expresses it, and hence a sentence ∃x¬D(x) that says, "some states are non-inclusive". That sentence must have a code number, *d* say, and the definition then tells us that *d* satisfies D if and only if it doesn't, as before.

The Cantor diagonalizations above prove theorems, but now let us turn to a structurally similar argument that is merely paradoxical. To generate the Grelling paradox, an adjective is defined as 'heterological' just in case it satisfies the negation of the property it represents. (E.g. 'long' is a heterological adjective, since it is not a long word, and hence satisfies the property <u>not long</u>.) When we ask whether 'heterological' is itself heterological, we obtain a contradiction, since that particular word is heterological if and only if it is not heterological.

In this case, the "objects" are adjectives, and the "properties" are properties of adjectives. The surjection *f* is defined by the semantics of English, and maps each adjective to the property it expresses. Here there is no explicit hypothesis about the class of properties expressible in English, for example that every

property can be expressed in English. This is apparently not necessary, since 'heterological' becomes an adjective of English as a result of its definition, and so is guaranteed to be expressible in English. It should also be noted that the Grelling paradox does not need to involve the word 'heterological'. Instead of saying that 'long' is heterological, we can say that 'long' is *an adjectival word or phrase that satisfies the negation of the property it represents*. Then a contradiction arises concerning whether that italicised adjectival phrase itself satisfies the negation of the property it represents.

It is generally agreed that arguments like the ones above involving a formal language are cogent, whereas those like Grelling's paradox, using natural language, merely involve a vicious circle of some sort. Let us investigate why this is so, beginning with Grelling's paradox. An important feature of the definition of 'heterological' is that it is *indirect*, in the sense that it provides a meaning of 'heterological' in terms of the *existing* meanings of other words (e.g. 'long', 'monosyllabic', etc.) To know whether a word X is heterological, we must already know what X means. (By contrast, I can know that the word 'subdermatoglyphic' has certain syntactic properties – it is six-syllabled, contains the letter 'm', etc. – without having any idea of its meaning.) This indirectness is fine as long as X already has an assigned meaning, for in that case the meaning of 'heterological' is ultimately grounded. But when X is the word 'heterological', then we are trying to define 'heterological' in terms of itself. We have a closed circle, and a contradictory one, due to the negation operator in the definition.

We saw that Grelling's paradox can be constructed without using the word 'heterological', but the basic circularity is the same. We know that vicious circles can be created by declarative and imperative sentences as well as by definitions. For example, you ask Alan what the date is, and he replies, "It's the date that Beth says it is." Then you ask Beth, and she says, "It's the day after what Alan says." Or suppose that Alan and Beth are eating at a restaurant. Alan tells the waiter, "I'll have what she's having", while Beth says, "Give me something different from him." The moral is that indirect assignments of meaning have to be carefully restricted to ensure that they are ultimately grounded, and do not form a circle.

The definitions of 'inclusive' in the arithmetical arguments above are also indirect, since to know whether *x* is inclusive you first have to know the formula (if any) that *x* is the code number for, and then you need to know the meaning of the formula. No circularity results, however, since every formula of *L* already has a well-defined meaning, provided by the formal semantics. (These formulas are composed entirely of

logical and arithmetical symbols, such as $\forall$, $\rightarrow$, +, $\times$, 0, 1, etc., and do not include the symbol 'inclusive'.) Moreover, the contradiction results only when one assumes that the property inclusive is expressed by one of the formulas of *L*, so the contradiction refutes that assumption.

5. Languages with intrinsic semantics

The notion of a physical system with conscious intentional states (similar to those of a human being) is implicitly contradictory, as I will argue in the next section. The contradiction arises from the more general notion of what we might call a 'self-interpreting' language, or one with intrinsic semantics. Let *L* be such a language. The sentences of *L* are physical symbols of some kind, with fully transparent syntactical properties, but there is no need for semantical rules since the propositional meaning of each sentence is logically determined by its syntactical properties. Now I realise that such a language may seem to be an impossibility, and indeed I will show that (under certain assumptions) it *is* impossible. However, the existence of a language with intrinsic semantics is a consequence of physicalism. This will be shown in the next section, but the rough idea is that the possible physical states of a thinking machine can be regarded as the sentences of such a language, and that the mind-reading thesis entails that these sentences have intrinsic meaning.

I will now show that the existence of a self-interpreting language *L* leads to contradiction, if *L is assumed to be able to express properties that refer to its own intrinsic semantics*. (To paraphrase Pascal's *pensée* quoted above, I will show that it is impossible for syntax to talk about itself.) The proof again uses the Cantor diagonalization, and is structurally identical to the second argument above concerning languages of arithmetic.

Let us suppose that some sentences of *L* express existential thoughts, of the form $\exists x\mathrm{F}(x)$, where F is a *purely syntactic* property of some *L*-sentences. If S is such an existential sentence, then it is always a clear-cut matter whether S itself has the syntactic property F. Let S be an *inclusive* sentence if and only if (i) S expresses such an existential, and (ii) S has the embedded syntactic property F. Since the syntactic properties of a sentence S determine whether it is inclusive, it follows that inclusive (and its negation, non-inclusive) are themselves syntactic properties.

The usual Cantor construction then shows that the property non-inclusive can't be expressed in *L* itself. If it could, then there would be an *L*-sentence (D, say) which says that some *L*-sentences are non-inclusive. This sentence D is an existential, and the property involved, non-inclusive, is a syntactic property, so D is inclusive if and only if D satisfies that property. This is a contradiction, so by *reductio ad absurdum* there is no such sentence D.

In the next section we will consider what this result means for physicalism, but first we should make sure that the definition of 'inclusive' is not circular. As stated in the previous section, a definition is at risk of circularity when it is indirect – that is, the definition appeals to the meanings of other words – for then those other words might include the term being defined. On the face of it, this definition of 'inclusive' is indirect, since it appeals to the meaning of 'F' that is embedded in the existential. There is no risk of circularity, however, due the fact noted above that inclusive is a purely syntactic property. To make the matter clear, it may help to suppose that the *L*-sentence S is (say) Nf33i!4/#09Jl*2, and intrinsically means 'there exists an *L*-sentence containing exactly 15 characters'. Then S is inclusive, since it expresses an existential involving a syntactic property, and S itself consists of 15 characters. But remember that the semantic properties of S logically follow from its syntactic properties, so the fact that S means 'there exists an *L*-sentence containing exactly 15 characters' is logically equivalent to a *purely syntactic* fact about S. (Let's say it's the fact that it contains the sequence '33i!4' somewhere.) What about the fact that S is *inclusive*? What kind of fact is that? It follows from the conjunction of two syntactic facts: that the sequence contains '33i!4', and that it consists of exactly 15 characters. Every other inclusive sentence is similarly inclusive in virtue of a purely syntactic fact about it. Thus 'inclusive' itself is equivalent to the disjunction of these syntactic facts, and so is also purely syntactic. Being a purely syntactic property, there is no risk of circularity.

This point is really central to the argument in this paper: the definition of 'inclusive', in the case of a self-interpreting language, is purely syntactic. If Laplace's demon were to survey the sentences of such a language, he would fully understand the syntactic properties of each sentence and be able to infer its semantic properties. Then inclusive, from the demon's perspective, would be just another syntactic property. The demon could even express it in ordinary syntactic terms, although perhaps not very tidily. The definition of inclusive cannot be circular, therefore.

6. Thinking Machines

In this section we will show that physicalism is committed to the existence of a self-interpreting language for each thinking being, and consider the consequences of that fact.

Suppose there exists a physical system M (for 'machine') that literally has conscious occurrent beliefs and other intentional states. Like many science fiction characters, M is assumed to have an intellect roughly equal to (or greater than) that of an intelligent human. As argued in Section 3, physicalism entails the mind-reading thesis, that these intentional states logically supervene on the complete physical state of M. In other words, if Laplace's demon is supplied with a precise physical description of M at some time[5], the demon could then infer the propositional content of the machine's conscious thought, at that time.

The thinking machine M will have some class of possible physical states. Also, for each proposition *P* that M is capable of conceiving, there will be at least one possible state in which the machine is consciously entertaining (i.e. reflecting on or considering) *P*. For convenience, I will assume that the machine is capable of consciously considering only one proposition at a time, so that each physical state is associated with one proposition at most. In that case, a state in which a proposition *P* is being considered is effectively a sentence that expresses *P*, in the sense that it is a collection of physical properties that precisely specifies the proposition *P*. Moreover, the set of physical states of M that express propositions in this way is effectively a language, so we will call this set *L*. Note that this 'language' *L* has intrinsic semantics in the sense defined above, since the propositional content of each 'sentence' of *L* is a logical consequence of its physical properties.

Now that we have a self-interpreting language, the term 'inclusive' can be defined exactly as in the previous section: An inclusive physical state S is one in which the machine M is entertaining an existential proposition, "Some state has the property F", where F is a purely physical property, and S itself has the

---

[5] Or, if narrow content doesn't exist, then Laplace's demon also needs a physical description of some features of M's environment. This point applies every time I refer to the 'possible states' of M.

property F. For example, if M is a computer made of logic gates, which are numbered, then there may be a state S in which the computer is considering the idea that "Logic gate #76352 is off in at least one state", and logic gate #76352 is off in S itself. Also, since the logical content of each state is a consequence of its physical properties, inclusive is a purely physical property. As shown in Section 5, a contradiction then arises if we suppose that M is able to grasp the notion of an inclusive physical state, as defined here, and form the thought "Some states are non-inclusive".

Could physicalism be saved simply by saying that M cannot grasp this particular (rather convoluted) property? I do not see how, given the fact that the concept of an inclusive physical state is very natural, once one sees that physicalism entails the mind-reading thesis, which leads to there being a language with intrinsic semantics. It should be noted that forming this concept does not require any detailed knowledge of M's internal constitution, but only other general concepts – just as forming Cantor's diagonal property requires no specific knowledge of the surjection *f*. A thinker who could not form the concept of an inclusive state could hardly understand the meaning of 'physicalism' itself. If physicalism entails that no mind can understand what 'physicalism' means, then physicalism is self-referentially incoherent.

7. Why is physicalism to blame?

It is clear that a contradiction is derived in the previous section, using a Cantor-style diagonal construction, but perhaps it is less clear what it is about physicalism that might be responsible for it. In fact, one reviewer has suggested that this contradiction might result from 'any view where the vehicles of intentionality have their content essentially'. In this section, therefore, I will explain why physicalism, and in particular its claim that the world is conceptually transparent, is needed for this contradiction.

Consider a 'vehicle of intentionality' (S, say) that has its propositional content essentially, yet which is conceptually opaque. Since S is conceptually opaque, Laplace's demon cannot understand its nature, but instead finds it 'dark' or 'obscure', to borrow Boyle's terms. In other words, S is not defined by a set of transparent syntactic properties from which its content can be logically inferred. A language whose sentences are things like S will not therefore be a self-interpreting language in the sense used in Section 5, and the

contradiction derived in Section 5 is not valid for such a language. In particular, it is stated there that the Cantor construction requires that the self-interpreting language is 'able to express properties that refer to its own intrinsic semantics'. The need for this requirement becomes clear when 'inclusive' is defined, since an inclusive sentence S must possess the syntactic property F that is part of S's meaning. If a vehicle of intentionality S is conceptually opaque, on the other hand, such self-reference cannot occur, and the diagonal property cannot be constructed.

To get a better philosophical understanding of why this contradiction arises for physicalism in particular, it might help to examine another view – a variant of Berkeleyan idealism – that suffers from the same problem. According to this view, which we can call *ultra-idealism*, every created entity, including human minds, is a mathematical idea in the mind of God.[6] (Thus ultra-idealists reject Berkeley's view that human minds are finite mental substances, and regard them as mere ideas, just like everything else in creation.) Now, since minds clearly have conscious intentionality, ultra-idealists claim that (some) *mathematical ideas themselves have conscious intentionality*.

It is easy to see that ultra-idealism entails the existence of a self-interpreting language, and hence allows the construction of the Cantor diagonal in Section 5. (Each mathematical idea that has intrinsic intentionality, in the sense of entertaining a proposition, can be considered a sentence that expresses that proposition.) In fact, in all respects relevant to constructing the Cantor diagonal, physicalism is identical to ultra-idealism. Even though physicalism regards material objects as mind-independent substances, and not ideas as such, only the mathematical structures within material objects are ever used to explain anything. The concrete aspect of a material object is unintelligible, and of no practical use to physicists, so that physicalists generally pay no attention to it.[7]

---

[6] I am not sure if anyone endorses ultra-idealism, but the 'Mathematical Universe Hypothesis' of Tegmark (2008), according to which, 'Our external physical reality is a mathematical structure', is at least a very similar claim, and it also seems to be refuted by the argument in this paper.

[7] Russellian monists do of course appeal to alleged non-mathematical aspects of the human brain to explain consciousness and other mental phenomena. On the definition of 'physicalism' used here, however, Russellian monism is not a physicalist view since it denies the conceptual transparency of the world.

According to ultra-idealism, some mathematical ideas have conscious intentionality, and physicalism makes practically the same claim.[8] Moreover, these mathematical ideas don't just have thoughts, they can even have thoughts *about themselves*. This is the key point. Constructing the Cantor diagonal requires that "the snake bites its own tail", so to speak. The self-interpreting language must be capable of talking about its own sentences, and its own intrinsic semantics. For this to be the case, it is not enough that the vehicles of intentionality have their content essentially; in addition, those vehicles must be completely intelligible – like mathematical structures for example. If the vehicles of intentionality are 'obscure' or 'dark', like alchemical concepts, then one cannot define any diagonal property like inclusive on them.

8. Conclusion

Physicalism and the mechanical philosophy share a common assumption that the world is fully transparent to a mind of sufficient power. For centuries, many philosophers have regarded such transparency as impossible in the case of perceiving subjects, possessing consciousness and intentionality, though to my knowledge no one has previously demonstrated this impossibility.

The conceptual transparency of physical properties means that physicalism entails the mind-reading thesis, described in Section 3. This in turn means that the physical states of a thinking being are like sentences of a language, except that they have their propositional meaning intrinsically, rather than through arbitrary convention. In Section 5, I showed that such a "self-interpreting" language must be limited in the properties that it can express: in particular, it cannot refer to its own intrinsic semantics. However, given that we humans *can* understand physicalism, and its consequences such as the mind-reading thesis, a contradiction results from the supposition that we are purely physical beings.

The diagonal argument presented here, like the familiar knowledge argument, aims to refute *a priori* (or Type-A) physicalism[9]. However, my argument is quite different from the knowledge argument, and does

[8] To be precise, physicalism entails that some concrete entity, whose state at some time *t* is perfectly represented by a mathematical structure, has conscious intentionality.

[9] See Chalmers (2003) for definitions of Type-A and Type-C materialism.

not face the same objections. For example, the response that Mary acquires a new ability rather than knowledge, or that Mary gains a new way to conceptualise old facts, are not possible objections to my argument. The response of Type-C physicalists, that Mary *could* infer what colour experiences are like, even though *we* are unable to see how, is also inapplicable. Further, while one could simply claim that qualia are some kind of illusion, to deny the reality of intentional states is much more problematic, as it is seems to be self-refuting. Thus, whatever one thinks of the knowledge argument, my argument should be of independent interest.

Conflict of Interest Statement

The author states that there is no conflict of interest.

References

BonJour, L. (2010) "Against Materialism", in Koons, R. C. and Bealer, G. (eds.) *The Waning of Materialism*, Oxford: Oxford University Press.

Boyle, R. (1674) *The Excellency and Grounds of the Corpuscular or Mechanical Philosophy*.

Chalmers, D. (2003) "Consciousness and its Place in Nature". In Stephen P. Stich & Ted A. Warfield (eds.), *Blackwell Guide to the Philosophy of Mind*. Blackwell. pp. 102-142[10].

Jackson, F. (1982) "Epiphenomenal Qualia". *Phil. Quarterly*, 32(127), 127-136.

Leibniz, G. W. (1714) *Monadology*.

Pascal, B. (1669) *Pensées*, Paris: Hachette, 1950.

Tegmark, M. (2008) "The Mathematical Universe". *Foundations of Physics*, 38(2): 101–150.