Russellian Diagonal Arguments and Other Logico-Mathematical Tools in Metaphysics

Laureano Luna

I.E.S. Doctor Francisco Marín, Siles, Spain

Abstract

In its most general form, a diagonal argument is an argument intending to show that not all objects of a certain class C are in a certain set S, and does so by constructing a diagonal object, that is to say, an object of the class C so defined as to be other than all the objects in S. We revise three arguments inspired by the Russell paradox (an argument against Computationalism, an argument against Physicalism, and a counterargument to the Platonic One Over Many argument), extract its underlying structure, and suggest a criterion to tell the ones that end up at a paradoxical object like the old Russell set from the ones that could actually accomplish a diagonalization. We conclude with the suggestion that the use of logico-mathematical tools, which is a significant methodological contribution of the analytical tradition, opens up a promising line of research in metaphysics.

Keywords: Analytical Philosophy, Computationalism, Physicalism, Platonic Forms, Sets, Diagonalization, Indefinite Extensibility, Russell's Paradox, Axiom of Replacement, Metaphysics.

1. Four Russellian Diagonal Arguments in Metaphysics

In its most general form, a diagonal argument is an argument that shows that not all objects of a certain class C are in a certain set S and does so by constructing (usually by reference to S) a diagonal object, that is to say, an object of class C that is other than all the objects in S. We expound three arguments concerning metaphysics, all of them inspired by the Russell paradox, extract its underlying structure, and suggest a criterion to tell the ones that end up at a paradoxical object like the old Russell set from the ones that could actually be able to deliver a diagonal object.

Luna and Small (2009) have put forward a Russellian diagonal argument against Computationalism. Computationalism is the thesis that all thought acts are computations (executions of algorithms) so that there is a correspondence between thought types and algorithms. Below is a version of the argument. By 'thought' we will mean hereafter 'thought type'. Note that we define a special re-

Argumenta 2, 1 (2016): 107-123
ISSN 2465-2334

DOI 10.14275/2465-2334/20163.LUN

lation and denote it by marking 'about' with an asterisk: 'about*'; this relation is not exactly the one Luna and Small use but it will do the job as well.

ARGUMENT 1

Assume Computationalism.
If Computationalism is true, there is a function f from algorithms to thoughts, representing the correspondence between the former and the latter, such that, for any thought t, there is an algorithm g such that f(g)=t.
Let us say that a thought is about* x if and only if ('iff', henceforth) it asserts a proposition of the form '$\forall y\ (y \in S_1 \rightarrow y \in S_2)$', for some sets S_1, S_2, and $x \in S_1$.
Call an algorithm g normal iff f(g) exists and is not about* g; let S be the set of all normal algorithms.
Let t^* be a thought asserting just '$\forall x\ (x \in S \rightarrow x \in S)$' and let $t^*=f(g^*)$.
t^* is exactly about* all normal algorithms.
Then, by the usual Russellian reasoning,[1] g^* is normal iff it isn't. Contradiction.
Therefore, f does not exist and Computationalism is not true. □

The aboutness* relation involved in the argument may look contrived but all that matters for the validity of the argument is that it be well-defined, and there is no obvious reason to believe it is not. The mention of sets S_1 and S_2 obeys the reason that quantifiers bounded by predicates like 'all P' are usually granted to successfully quantify over all P if P's extension is a set.

Argument 1 invites a parallel argument against Physicalism if by Physicalism we understand the claim that mental states or thoughts are so dependent on brain states ('brainstates', hereafter) that no thought exists without a corresponding brainstate and no two different thoughts can accompany one and the same brainstate-type: this is often called 'type supervenience Physicalism'. By 'brainstate' we will mean 'brainstate-type' hereafter.

ARGUMENT 2

Assume Physicalism.
If Physicalism is true, there is a function f from brainstates to thoughts such that, for any thought t, there is a brainstate b such that f(b)=t and t is the thought that accompanies b.
Let us say that a thought is about* x if and only iff it asserts a proposition of the form '$\forall y\ (y \in S_1 \rightarrow y \in S_2)$', for some sets S_1, S_2, and $x \in S_1$.
Call a brainstate b normal iff f(b) exists and is not about* b; let S be the set of all normal brainstates.
Let t^* be a thought asserting just '$\forall x\ (x \in S \rightarrow x \in S)$' and let $t^*=f(b^*)$.
t^* is exactly about* all normal brainstates.
Then b^* is normal iff it isn't. Contradiction.
Therefore, f does not exist and Physicalism is not true. □

[1] Assume g^* is normal; then t^* is about* g^* (for it is exactly about* all normal algorithms) and this makes g^* not normal. Assume g^* is not normal; then t^* is not about g^* and this makes g^* normal.

Nothing is evidently wrong in the argument. However, the fact that it can be easily parodied should cast the shadow of a doubt upon it. For consider the following reasoning, to which we have given an unnecessarily complex form to mirror the structure of arguments 1 and 2 (this is why some phrases are in parentheses):

ARGUMENT 3

Assume there are thoughts.
If there are thoughts, there is an identity function f from thoughts to thoughts such that, for any thought t, (there is a thought t such that) f(t)=t.
Let us say that a thought is about* x if and only iff it asserts a proposition of the form '$\forall y\ (y \in S_1 \rightarrow y \in S_2)$', for some sets S_1, S_2, and $x \in S_1$.
Call a thought t normal iff f(t) (exists and) is not about* t.
Let θ^* be a thought asserting just '$\forall x\ (x \in S \rightarrow x \in S)$' (and let $\theta^* = f(\theta^*)$).
θ^* is exactly about* all normal thoughts.
Then θ^* is normal iff it isn't. Contradiction.
Therefore, f does not exist and there are no thoughts. □

This reasoning has the same structure as Russell's paradox, of which the following is a version:

ARGUMENT 4

Assume there are sets.
If there are sets, there is an identity function f from sets to sets such that, for any set s, (there is a set s such that) f(s)=s.
Call a set s normal iff f(s) (exists and) $s \notin f(s)$.
Let s^* be the set of all normal sets (and let $s^* = f(s^*)$).
Then s^* is normal iff it isn't. Contradiction.
Therefore, f does not exist and there are no sets. □

Russell's famous paradox uses the set theoretical membership relation instead of the aboutness* relation above defined. Of course, arguments 3 and 4 can be easily simplified: function f in each of them serves the unique purpose to make apparent that they can be given the same structure as arguments 1 and 2.

The underlying structure of these Russellian diagonal arguments is a *reductio*:

1. We assume there is a surjective function f: $A \rightarrow B$.
2. We define a relation R relating members of B and members of A.
3. We define a member b^* of B having R to exactly all members of A not related by R to their images by f (i.e. to all *normal* members of A).
4. b^* is the image by f of some member a^* of A.
3. b^* has R to a^* iff it doesn't. Contradiction.
5. Therefore, f does not exist. □

The object b^* is the diagonal object: we use it to diagonalize out of the range of f, that is to say, to construct a member of B that is the image by f of no member of A, showing f is not surjective.

The structure of these diagonal arguments responds to Priest's Inclosure Schema (Priest 2002: 134). Priest defines a set $\Omega = \{x: \varphi(x)\}$, for some property

φ, and assumes ψ(Ω) for some property ψ; then, for each x⊆Ω which has property ψ, he defines a diagonalizing function δ such that δ(x)∉x and δ(x)∈Ω. Obviously, this schema leads to the contradiction that δ(Ω)∈Ω and δ(Ω)∉Ω. Priests is inclined to endorse the contradiction; if we are not, we can infer that if δ exists, then there is no set Ω of all φ-objects that has property ψ. In our pattern, property φ would be 'being a member of B' and property ψ would be 'being the image under f of some subset of A', and δ(x) would be the diagonal object b^* we produce by means of relation R. Thus, the arguments conclude that B is not the image under f of a subset of A. The situation could also be depicted in the terms of Shapiro and Wright (2006) by saying that property φ is *indefinitely extensible relative* to property ψ, which as before implies that there is no set of all φ-objects that has property ψ.[2] The contradiction arrived at on each occasion depends on a first order validity sometimes called 'Thomson's theorem' (Thomson 1962): ~∃x ∀y (Rxy ↔ ~Ryy).

Arguments 3 and 4 are obviously unsound but it is not obvious what is wrong with them.

2. Diagonal Arguments and Paradox

Let us first address argument 4, which is essentially the Russell paradox.

Certainly, a number of authors, when dealing with Russell's paradox, limit themselves to the conclusion that the diagonal object does not exist on pain of contradiction. Even if the conclusion is true, acknowledging its truth does not provide us with an explanation thereof, hence also not with a solution to the paradox. Simply stating that there is no set of all non self-membered sets because the extension of the concept of non self-membered set is too large to form a set is no explanation at all. If s^* (i.e. the old Russell's set) does not exist, there must be an explanation of why and how its definition fails to define a set. We will offer the two main narratives that aim at an explanation of the facts; we will call them the conventional (the term implies no pejorative connotation) and the alternative narrative.

Let us address the conventional approach in the first place. How the definition of s^* in argument 4 (namely, 'set of all normal sets') fails to define a totality is incomprehensible if normality of sets is well-defined. So, we should explore the possibility that normality is not well-defined. But for normality to be ill-defined, set membership must be ill-defined. And in fact, we usually admit it is ill-defined in a sense, namely, in the sense that there is no definite totality (i.e. no set) of all sets. The multiplicity of sets does not make up a definite totality; it is *indefinitely extensible* or *open ended:* there are sets beyond any set of sets.[3] Hence, sets are not given all at once but come in stages or levels; one cannot have a definite totality of them all; rather, for any definite totality of sets one can define or think of, there are sets beyond that totality, at a higher level. Each time we de-

[2] I thank an anonymous reviewer for suggesting the convenience of mentioning Priest's Inclosure Schema in this context. We are using a simplified version of Shapiro and Wright's *relative indefinite extensibility.*

[3] See Russell 1905 for an early discussion of the topic, though an expression translatable by 'indefinite extensibility' have I found nowhere before Zermelo's 1930 "schrankenlose Fortsetzbarkeit".

fine a totality of sets, we rise to some level of a hierarchy of such totalities, with more sets showing up at further levels.[4]

Accordingly, there must be levels of membership and levels of normality. But the definition of s^* does not specify any level of normality. In failing to distinguish levels of normality, the definition of s^* may be overlooking a necessary hierarchy and committing vicious circularity: after all, it would be defining membership in s^* in terms of self-membership of all sets, s^* included (we will dwell below on the circularity in the definition of s^* and the ensuing necessity to distinguish levels of set membership and normality). These reasons would explain why the definition of the diagonal object fails and the object itself does not exist. Such is the conventional narrative in its bare bones.

There is an alternative account. Some logicians believe our quantifiers can only range over objects that are previously available, so that they never range over indefinitely extensible multiplicities but only over set-sized portions of them that are, if not otherwise, determined by context.[5] If this is actually so, the level of normality of s^* is implicit in the definition of s^* and determined by context in such a way as to avoid circularity. s^* just stands at a higher level than all the sets its definition is about. The definition of s^* would refer to normal sets that are normal at a level of normality that is below the level of normality at which s^* could be or fail to be normal. Then, s^* could be normal without containing itself because it would possess normality at a higher level than its members.[6] This reading of the definition of s^* dissolves (rather than solves) the paradox: the appearance of a paradox would arise from an incorrect reading of the definition of the diagonal object. In this interpretation, s^* is not at all paradoxical; it just diagonalizes out of the sets its definition is about, extending the universe of discourse of its definition by one set, namely, s^*. This alternative approach is typically preferred by those logicians who believe that absolutely unrestricted quantification is impossible and the universe of discourse is always restricted to a set-sized multiplicity, which may be indefinitely extended by diagonalization (for an overview of positions concerning the possibility of unrestricted quantification, see Rayo and Uzquiano 2006, Introduction).[7]

What about argument 3? Is there a related way out? It seems so, since, arguably, there is no set of all thoughts. Patrick Grim (1991, ch. IV) has displayed a number of arguments to show that there is no set of all truths and some of them can easily be transferred from truths to thoughts. One of them is just a Russellian diagonalization argument (Grim 1991: 110-13), of which this is a version:

[4] This is of course the *iterative conception* of sets, according to which sets come in stages forming an indefinitely extensible hierarchy in which sets are always posterior in the hierarchy to their members.

[5] This theory is congenial to the usual model theoretic principle that all universes of discourse are sets.

[6] Note that s^*, if it exists, is normal at some level of normality; otherwise, it would be a member of itself, and only normal sets are members of s^*.

[7] The alternative approach faces this objection: why can normality-at-any-level not be used to reproduce the paradox? But the answer seems straightforward: 'normality-at-any-level' fails to quantify over all levels because the hierarchy of levels is indefinitely extensible. The alternative account does little more than assuming that the hierarchy proposed by the conventional account is in fact always implicitly present in our discourse.

> Let T be any thinkable (i.e. definable, constructible or susceptible of specification) set of thoughts; define some aboutness* relation from thoughts in T to any objects; construct a diagonal thought that is exactly about* all thoughts in T not about* themselves; if the diagonal thought were in T, it would be about* itself iff it were not; hence, it is not in T; rather, it diagonalizes out of T; as the diagonalization procedure can be achieved for any thinkable set of thoughts, there is no thinkable set of all thoughts; but if the set of all thoughts existed, it would be thinkable, since it could be thought of as the set of all thoughts; thus, it does not exist.[8]

Here is a related argument based on an idea by Russell 1903 (par. 500: 538-39) that employs no aboutness relation:

> For any definable set of thoughts x, let $\pi(x)$ be its product, which is the thought that all thoughts in x are thoughts. For any such x, $\pi(x)$ exists (even if $x=\varnothing$ and $\pi(x)$ is vacuously true). Let s be any definable set of thoughts. Let R be the (possibly empty) set of all products $\pi(x)$ in s such that $\pi(x)\notin x$. R is definable because s is; hence, $\pi(R)$ exists. Assume $\pi(R)\in s$. Then $\pi(R)\in R$ iff $\pi(R)\notin R$, which is a contradiction. Hence, $\pi(R)\notin s$ and there is a thought not in s. As s was any definable set of thoughts, there is no definable set of all thoughts. But if the set of all thoughts existed, it would be definable as the set of all thoughts. Therefore, the set of all thoughts does not exist.[9]

This is exactly how Russell's paradox is customarily used to prove there is no set of all sets: for each set s of sets, there is a set not in s, namely, the set of all non self-membered sets in s. In these arguments we apply the diagonalization procedure only to sets and we assume that the diagonal objects cannot be ill-defined; indeed, as we deal only with multiplicities that are sets, we can no longer argue that the aboutness or membership relations are ill-defined because thoughts or sets should come in levels: those thoughts or sets that can be put in a set can be given all at once because they do make up a definite totality.

So, we can escape argument 3 in the same way we avoided argument 4. In the conventional narrative, since it is incomprehensible how t^* could fail to exist if normality of thoughts is well-defined, we are compelled to assume that normality is not well-defined; and it is not because the multiplicity of thoughts is not a definite totality; it is indefinitely extensible: there are thoughts beyond any set of thoughts. Thoughts come in levels. Accordingly, there must be levels of aboutness* and levels of normality; then θ^*, the purported thought about* all normal thoughts, is defined with vicious circularity because its definition does not specify a particular level of normality.

In the alternative narrative, the level of normality is implicit in the definition of θ^*, and θ^* is about* normal thoughts that are normal at a level of normality that is below the level of normality at which θ^* could be or fail to be normal. If so, θ^* can be normal without being one of the thoughts θ^* is about and no paradox exists: θ^* simply diagonalizes out of the thoughts it is about*.

[8] See also Luna and Small 2009: 88-89.

[9] One may be tempted to think that the definable objects must form a set because there are at most a countably infinite number of definitions; however, Richard's paradox strongly suggests that natural language is indefinitely extensible. Luna and Taylor 2010 propose that some syntactical expressions must define different objects in different logical contexts due to inevitable ambiguity in the range of quantifiers. See also Luna 2013.

3. Arguments 1 and 2 vs. Arguments 3 and 4

Let us reckon what the fate of arguments 1 and 2 would be if we could apply to them what the conventional or what the alternative narrative has to say about Russell's paradox. In the conventional approach, we would accuse the diagonal objects of being ill-defined and we would declare them nonexistent, which would indeed refute the arguments. In the alternative narrative, we would claim that the diagonal objects are not comprehensive enough in their scopes to bring about the contradictions they seem to provoke (for they would fail to be about* all normal objects) and this would all the same refute the arguments. Obviously, if one of these criticisms is not applicable because the argument's framework is not of the proper kind, neither is the other, since those criticisms are alternate treatments of one and the same type of situation, namely, multiplicities that can never be entirely given.

So let us first examine whether we can level against the diagonal objects in arguments 1 and 2 the same type of counterargument the conventional narrative employs against the existence of s^* and θ^*. Can the diagonal objects in these arguments be ill-defined for the same reason as s^* and θ^* are in the conventional narrative?

As regards argument 1, Luna and Small deny that possibility. They argue that, at least under the Church-Turing thesis, the set of all algorithms not only exists, it is effectively enumerable: it is the set of all Turing machines; hence, according to the model theoretical principle that any nonempty set is a possible domain of discourse, we should be able to refer to them all in order to diagonalize out of them; the authors assume that principle and call it the *principle of semantic clarity*. In the terms of our approach here, we would say that contrary to what happened with sets and thoughts, if algorithms can all be put in one and the same set, that is to say, if they are all given at once and need not come in levels, concerns regarding levels of normality and circularity in the definition of the diagonal object are out of place. One has to be a strict finitist, which is an extremely radical position to adopt in philosophy of mathematics, to deny the existence of the set of all Turing machines. And even if one rejects the Church-Turing thesis and believes that algorithms exist that cannot be represented by Turing machines, one will most probably believe that algorithms, whatever they are, do form a set. Thus, as the central claim of Luna and Small seems plausible, one has to avow that there does seem to be a difference between this case and arguments 3 and 4.

If we approach argument 2 in this spirit, the relevant question is whether there is a set of all possible brainstates. We have no better reason to believe that there is no such set than we have to believe there is no set of all possible (types of) earthquakes or (types of) atoms, for instance. Brainstates are much the same kind of objects as types of earthquakes or as elements in the periodic table: they are types of physical objects and these are not the kind of objects we expect to constitute indefinitely extensible multiplicities. Usually, if a class C of objects is indefinitely extensible, it is the case that we can employ the definition of an arbitrary set of C-objects to define a new C-object not in the set, so diagonalizing out of the set. But possible types of physical objects, like types of earthquakes, do seem to be possibly given all at once because their givenness appears to be absolutely independent of our definitions of them: we can hardly make new types of earthquakes emerge by diagonalizing once and again out of sets of types of

earthquakes into an indefinitely extensible hierarchy of levels. Be it as it may, if the set of all possible brainstates does exist, brainstates and normality need not come in levels, and the proposition stated by thought t^* is, for all we know, well-defined; if so, there is no evident reason to deny that t^* is a possible thought that diagonalizes out of all thoughts in the range of f. One must grant at least that there does *seem* to be a relevant difference between t^*—in arguments 1 and 2—and the diagonal objects s^* and θ^* that, according to tradition, are paradoxical.

If the set of all algorithms and the set of all brainstates actually exist, the alternative approach to Russell's paradox has nothing to say about arguments 1 and 2, since it only deals with cases involving indefinitely extensible multiplicities.

Let us take this as a preliminary approach to our subject. In order to gain additional insight, we need to examine in some detail the topic of circularity in the definition of s^* in argument 4. Laurence Goldstein (2009), reasoning within the conventional narrative, believed that it is possible not just to prove the non-existence of s^* by *reductio* but also to explain and render it intuitive by showing why its definition fails to define a set. He pinpointed circularity in the definition of s^* in the following way.

Goldstein points out that if s^* existed, its definition would fail to define a set because it would be viciously circular; and it would be viciously circular because the expression that defines s^*:

$$\forall x\,(x \in s^* \leftrightarrow x \notin x)$$

can be developed into the infinite conjunction of one sentence of the following form for each set s:

$$s \in s^* \leftrightarrow s \notin s.$$

If s^* existed, one of these sentences would be

$$s^* \in s^* \leftrightarrow s^* \notin s^*$$

which would render the definition of s^* clearly circular, besides inconsistent. Since s^* is specified by its definition, it can only exist if its definition succeeds in defining a set. But if s^* exists, its definition fails and s^* does not exist; as a consequence, s^* does not exist.

Goldstein's idea suggests the necessity of distinguishing levels of set membership in order to avoid viciously circular definitions. This in turn suggests another approach to the problem of the circularity in s^*. Note that normality is defined upon the set membership relation. For normality to be well-defined, the set membership relation should be determinate when normality is defined upon it. But that is not the case if s^* exists because membership in s^* depends on normality, for s^* is defined to have precisely all normal sets as members. So, if s^* exists, normality is defined upon set membership and set membership (in s^*) is defined upon normality. In order to disentangle our definitions, we should distinguish alternate levels of membership and normality: membership_0 is membership before any normality has been defined; normality_0 is normality defined upon membership_0; membership_1 is membership after normality_0 has been defined; normality_1 is normality defined upon membership_1, etc.

One can indeed obtain a set of all normal sets at each level of normality but one can extend the levels through all ordinals, which go beyond sethood. Hence, the necessity of distinguishing levels of normality implies that the objects for which normality is defined do not make up a set but an indefinitely extensible

multiplicity. *If the objects for which normality is defined do form a set, then the hierarchy of normality levels is not indefinitely extensible because there is a level at which all normality levels are already available, namely, the level at which those objects form a set; this should permit to use a definition of normality simultaneously valid for all levels, if such levels exist at all.*

It is easy to see how the alternative narrative would have it here; it would contend that, since normality can only be defined upon what is already determinate, it is *in fact* not defined by reference to membership in s^* itself; so, s^* stands at a higher level of normality than all the normal sets its definition is about. This happens, so to say, in an automatic way. Hence, the definition of s^* cannot but diagonalize out of all the normal sets it refers to, and reading it otherwise makes no sense. Obviously, this approach only applies when the objects for which normality is defined do not form a set, for if they do, there is a highest level that contains all levels and out of which it is impossible to diagonalize. Both analyses of argument 4, that of the conventional and that of the alternative approach, are easily applied to argument 3 after the suitable replacements; essentially, one must substitute thoughts for sets and the aboutness* relation for the set membership relation.

In the following paragraphs, we will deal simultaneously with arguments 1 and 2 though explicitly referring solely to argument 1: applying to argument 2 what we will say about argument 1 is straightforward.[10]

The question is whether normality and the diagonal object in argument 1 are circularly defined for the same reason as they are in argument 3. t^*, the diagonal object in argument 3, is exactly about* all normal algorithms but if g^*—such that $f(g^*)=t^*$—exists, then aboutness* seems ill-defined for t^* because it is defined through these clauses: for each algorithm g,

$$t^* \text{ about}^* g \leftrightarrow f(g) \text{ not about}^* g$$

among which

$$t^* \text{ about}^* g^* \leftrightarrow t^* \text{ not about}^* g^*.$$

It is clear that if there is a g^* such that $f(g^*)=t^*$, then normality, as it occurs in the definition of t^*, is ill-defined. This seems to leave us with a disjunction: *either* normality in the definition of t^* is well-defined and g^* does not exist *or* normality in the definition of t^* is ill-defined (so that t^* does not exist and the existence of g^* is not even an issue). But this disjunction proves nothing. One can choose the first disjunct if one wishes to save the argument or the second if one prefers to keep Computationalism. So, to rescue the argument, the second disjunct must be shown false or, at least, it must be shown that it is false if Computationalism is true. We will argue for the latter, that is, we will argue that Computationalism would imply that normality in the definition of t^* is well-defined.

It would seem that argument 1 contains the same vicious circularity as arguments 3 and 4: in argument 1, normality is defined upon the aboutness* relation; hence, for normality to be well-defined, this relation should be determinate before we define normality; but it seems it is not, because aboutness* is defined by means of the expression '$\forall y\ (y \in S_1 \rightarrow y \in S_2)$' where S_1 can be the set of all normal algorithms, as it is in fact assumed to be for t^*. Thus, at least if t^* exists,

[10] In order to apply to argument 2 what refers to argument 1 in the following paragraphs, just replace 'argument 1' by 'argument 2',' algorithm' by 'brainstate', 'g^*' by 'b^*', 'g' by 'b', 'Computationalism' by 'Physicalism', and 'syntactical' by 'physical'.

normality involves aboutness* and aboutness* involves normality. So, it appears that to avoid circularity, we should introduce alternate levels of aboutness* and normality: aboutness*$_0$ is aboutness* before any normality is defined; normality$_0$ is normality defined upon aboutness*$_0$; aboutness*$_1$ is aboutness* after normality$_0$ has been defined; normality$_1$ is normality defined upon aboutness*$_1$, etc.

But this hierarchy of levels would take us too far if Computationalism is true and all thoughts are algorithms; it would take us beyond sethood, which is impossible if algorithms form a set, as they seem to do. If the set of all algorithms exists, normality cannot come in an indefinitely extensible hierarchy of levels: there must be a level at which normality of algorithms is fully determinate (namely. the level at which the set of all algorithms becomes available) and at that level we can profit from that determinateness to carry out the diagonalization procedure successfully. This situation makes it dubious that we can escape argument 1 by applying the same strategies we used against arguments 3 and 4: the fact that, in all evidence, algorithms do form a set would stay in our way.

As regards the determinateness which we claim algorithms possess and which should render normality well-defined in argument 1, consider that algorithms are syntactical in nature and syntactical facts are always determinate: Gödel showed they are equivalent to arithmetical facts. It is precisely this difference between semantical and syntactical properties that makes the whole difference between Gödel's famous self-referential sentence G and the Liar. G only involves the syntactical property of provability within a formal system and this makes of it a definite mathematical statement that cannot be paradoxical.[11] That syntactical facts (and most probably physical facts too) are determinate in a sense in which others are not can also be illustrated by this example: note that one can easily produce a paradox by referring to semantical properties of thoughts or of propositions as in 'what I am now thinking is false' or 'this proposition is not true' but it is hard to figure out how one could produce a paradox if one refers only to syntactical features as in 'this sentence has five words' or only to physical properties as in 'my current brainstate involves at least one billion synapses'. There is surely a relation between determinateness understood as the property of adding up to a definite totality and need not come in (an indefinitely extensible hierarchy of) levels and determinateness understood as the property of being so definite as to preclude paradox. This relation, however, requires further research to be developed elsewhere.

If Computationalism is true, the circularity in the definition of normality in argument 1 can only be apparent. If f exists and thoughts are linked by f to algorithms, any property or relation of thoughts is as determinate as syntactical facts are. Furthermore, if algorithms make up a set, as they seem to do, they need not come in levels; they may be given all at once. Confessedly, the case for the determinateness of physical facts is less conclusive than the case for the determinateness of syntactical facts. However, it would be really strange if brainstates were unable to form a set, for the objects in indefinitely extensible classes seem to depend on our capability to construct new objects by diagonalization and

[11] As an anonymous reviewer points out, there are indeed proofs of (versions of) Gödel's theorem that involve semantical notions (such as 'truth' or 'model'). However, the relevant fact is that the self-referential G only involves the syntactical predicate of formal provability: this is why it cannot be paradoxical. The nature of the proof of G's undecidability is irrelevant for this purpose.

physical states of affairs do not appear to be the kind of things whose existence would depend on our constructions.

The response to the criticism of argument 1 in the alternative account would be as follows: there cannot be normal algorithms at higher levels of normality than t^* can be about, because algorithms form a set, so that all of them can be simultaneously given, making up a possible universe of discourse.

4. Arguments 1 and 2 in Set Theoretical Terms: A Soundness Criterion

So far, our analysis has revealed that thoughts do not seem able to make up a set whereas algorithms do.[12] This permits to recast argument 1 in set theoretical terms. If thoughts do not form a set and algorithms do, then there can be no such surjective function as f from algorithms to thoughts: the set theoretic axiom of Replacement would prohibit its existence. The axiom of Replacement states that, if the domain of a function is a set, its range is a set too. Therefore, if algorithms form a set and f exists, then also thoughts form a set, which seems implausible. So, on the plausible assumption that there is a set of all algorithms but no set of all thoughts, we can use the axiom to argue that f does not exist and Computationalism is false. This is not the place to revise and discuss the justifications of the axiom of Replacement. We will only remark that it is so widely accepted, be it for its mathematical fruitfulness or its philosophical plausibility, that showing it incompatible with the thesis that each thought corresponds to some algorithm is enough to make a case against Computationalism.[13]

The fate of a Russellian diagonal argument seems to depend on whether the members of the multiplicity for which normality is defined form a definite totality and can be given all at once or they form an indefinitely extensible class and come in levels. If argument 1 succeeds as a diagonal argument, its success depends crucially on these facts:

1. There is a set of all algorithms but there is no set of all thoughts: thoughts form an indefinitely extensible multiplicity.
2. If Computationalism is right and f exists, then by the set theoretical axiom of Replacement, there is a set s_f of thoughts that is the range of f; but since thoughts form an indefinitely extensible multiplicity, it is possible to diagonalize out of any definable set of thoughts, hence also out of s_f; so, we can produce a thought that is not in s_f but diagonalizes out of it, thereby proving Computationalism false.

Consider this simplified argumental blueprint: there is no set of all thoughts but, since a set of all algorithms does exist, if Computationalism were right and f existed, Replacement would imply the existence of the set of all thoughts; thus, Computationalism is wrong. This fact would on its own refute Computationalism as defined but it would not refute a weaker form of Computationalism, namely, the thesis that all possible *human* thoughts are so related to algorithms

[12] An anonymous reviewer reminds me of the fact that some axiomatics (e.g. NBG) admit classes too big to be sets, usually called *proper classes;* so another way to express the difference would be: the class of all thoughts is a proper class while the class of all algorithms is a set.

[13] A *locus classicus* for the discussion of the rationale of the axiom is chapter 10 of Parsons 1983.

that f exists with respect to them; it would not, because computationalists could easily raise the counterargument that those thoughts that are not in the range of f, even if they are possible in some abstract sense, may not be possible *human* thoughts; that is, they might be impossible for creatures whose thoughts are linked through f to algorithms. But argument 1 goes a decisive step further and constructs one diagonal thought—hence one possible *human* thought—that is not in the range of f; thus, the argument, if successful, refutes also that weaker form of Computationalism.

The weaker form of Computationalism can also be argued against by means of a slight modification of the argumental blueprint above: there is no set of all possible *human* thoughts but, since a set of all algorithms does exist, if Computationalism were right and f existed, Replacement would imply the existence of the set of all possible *human* thoughts; thus, Computationalism is wrong. The problem is that so far, we have argued against the existence of a set of all possible thoughts but not against a set of all possible *human* thoughts. Here is an argument against the existence of such a set, framed along the lines of the Russellian argument on page 5:

> For any definable set of possible thoughts x, let $\pi(x)$ be its product, i.e. the thought that all thoughts in x are thoughts. For any such x, $\pi(x)$ exists as a possible human thought (because x is definable), even if $x=\varnothing$ and $\pi(x)$ is vacuously true. Let s be any definable set of possible human thoughts. Let R be the (possibly empty) set of all products $\pi(x)$ in s such that $\pi(x)\notin x$. R is definable because s is; hence, $\pi(R)$ exists as a possible human thought. Assume $\pi(R)\in s$. Then $\pi(R)\in R$ iff $\pi(R)\notin R$, which is a contradiction. By *reductio*, $\pi(R)\notin s$ and there is a possible human thought that is not in s. As s was any definable set of possible human thoughts, there is no definable set of all possible human thoughts. But if the set of all possible human thoughts existed, it would be definable as the set of all possible human thoughts. Therefore, the set of all possible human thoughts does not exist.[14]

When one compares the Luna-Small Russellian diagonal argument with Russell's paradox and its conventional solution—or more generally arguments 1 and 2, on the one hand, with arguments 3 and 4, on the other—a difference becomes apparent. As regards the latter, we have, at least in the conventional narrative, a standard reason to believe that the diagonal object is not well-defined, namely, that its definition fails to distinguish among the different levels of a property (i.e. normality) and this failure makes the definition fail as such. But this is not the case for the former: if algorithms or brainstates make up a definite totality (as we have good reasons to believe), then they can be all given at once and a definite aboutness* relation together with a one-level normality property must exist for them.

So, the criterion for distinguishing a Russellian construction that leads to a paradoxical object (in the conventional account) or fails to produce the desired contradiction (in the alternative one) from a genuine diagonal argument is this: do the objects for which normality is defined form a definite totality that can be given once for all or are they members of an indefinitely extensible hierarchy and come in levels?

[14] Recall that the analysis of argument 1 accomplished from footnote 10 to this can be rendered an analysis of argument 2 by making the substitutions described in footnote 10.

If they come in levels, normality without a level index may be ill-defined for them (which would cast just as much doubt on the existence of the diagonal object as there is about the existence of the old Russell set) or it may fail to be inclusive enough to bring about a contradiction. Otherwise, the case is essentially other than Russell's paradox.

As a consequence, unless we can substantiate the claim that algorithms or brainstates are unable to form definite totalities, we should not dismiss arguments 1 and 2 on the grounds that they are but avatars of the old Russellian paradox. *It is noteworthy that neither computationalists nor physicalists have seriously addressed the problem that human thoughts appear to spread along a hierarchy of levels that extend beyond sethood whereas algorithms and brainstates seem to be given or available once for all, so as to make up definite totalities.* Upon analysis, this is ultimately the state of affairs that renders arguments 1 and 2 plausible.

5. Assessing a Russellian Argument against Platonic Forms

Let us finally consider a Russellian argument proposed by Michael Loux (1998: 34-35)[15] though forms of the argument appear already in Russell (1903, par. 78: 80-81) and in Mally (1914: 225). It is ultimately an intensional version of Russell's set theoretical paradox. Loux' purpose is to deny that a famous Platonic thesis—to be found in *The Republic*, Book X, 596a-b[16]—is tenable in full generality. The thesis contends that whenever many different things are of a same kind, so that there is a same name convening to all of them, a corresponding Form exists (this is sometimes called the One Over Many argument for the existence of Platonic Forms). So, for example, a beautiful poem and a beautiful melody have beauty in common and we say that they are both beautiful; hence, according to the thesis, beauty exists as a Form. Thus, the argument turns the property P common to all entities in some collection c into a Platonic Form F_P. Thus, the following argument assumes, for *reductio,* that, for any collection c_P containing all the objects that have some property P in common, there is a Platonic Form F_P corresponding to P.

ARGUMENT 5

Assume the One Over Many argument.

As the One Over Many argument is right, there is a function f that takes a collection c_P of all objects sharing some property P and returns the corresponding Platonic Form F_P.

Call a Platonic Form F_P normal iff $F_P \notin c_P$, that is, iff it does not exemplify itself. For instance, the Platonic Form T corresponding to the property of being a table is normal because T is not a table but a Platonic Form.

Being a normal Platonic Form is a property which T has in common with other Platonic Forms (e.g. being a chair). Let c^* be the collection of them all. Then $f(c^*) = F^*$ exists.

F^* is the Form corresponding to the property of being a normal Form.

But F* is normal iff it isn't. Contradiction.

Therefore, f does not exist and the One Over Many argument is unsound. □

[15] I am indebted to James Grindeland for bringing this argument to my attention.

[16] See for instance Plato 1992: 265.

We have good reasons to believe that, even if Platonic Forms exist, there is no set of them all. We can argue, for instance, that there is at least one Platonic Form for each set s, namely, the Platonic Form corresponding to the property of being a member of s.[17] If so, Platonic Forms come in levels and normality for them is ill-defined in the sense in which normality is ill-defined for sets in argument 4 and for thoughts in argument 3. Therefore, either F^* is ill-defined and does not exist (as the conventional approach would have it) or, if it does exist, then it can be normal without contradiction at a higher level than all Forms exemplifying it, as the alternative approach to paradoxes would contend.

6. Mathematical Tools in Metaphysical Argumentation

Arguments 1 and 2 belong in a family of anti-reductionistic arguments attempting to show that thoughts are too different from other kinds of objects to be ontologically reduced to them or to be put in some (too narrow) dependence relations with them. For instance, it has been argued that thoughts have an intrinsic semantic nature while algorithms do not or that brainstates are spatial but thoughts are not. The novelty is that the difference here invoked is ultimately of mathematical nature: adding up to a definite totality (so algorithms and brainstates) vs. being spread along an indefinitely extensible hierarchy of levels.

The phenomenologist Ernst Mally (1914) used an argument very similar to argument 3 to support the claim that 'thought D is directed toward thought D' is meaningless because a duly typed language—akin to the proposed by Russell and Whitehead in the *Principia*—would not allow for it. If that sentence were not meaningless—Mally argues—it would make sense as well to construct a thought G directed exactly to all thoughts that are not directed to themselves; and G would be paradoxical.

There is a well-known argument against physicalism due to Kripke (Kripke 1980). Kripke relies on the following proof that true identities are necessary:

1. $\forall P \forall xy\ (x=y \rightarrow (Px \rightarrow Py))$	Premise (indistinguishability of identicals)
2. $\forall x\ (\Box(x=x))$	Premise (necessity of self-identity)
3. $\forall xy\ (x=y \rightarrow (\Box(x=x) \rightarrow \Box(x=y)))$	1, Universal Instantiation for 'P'
4. $\forall xy\ (x=y \rightarrow \Box(x=y))$	2, 3, Propositional logic.

Kripke adds a conceivability argument: the identity of mental and physical states is at most contingent since its falsity is conceivable; and concludes that the identity is false. No matter how controversial, it is an example of a formal argument in metaphysics.

A set theoretical argument against *global supervenience materialism* (the thesis that the whole sphere of the mental supervenes on the whole physical state of the world) has been proposed by Franz von Kutschera (1994). The following is a version of that argument.

If we represent each proposition as the set of all possible worlds at which it is true, there are more possibly true propositions (i.e. propositions that are true at some possible world) than possible worlds, because of Cantor's theorem. This suggests that not all possibly true propositions can be believed (the original argument elaborates on this point). Let us divide the class of all possible proposi-

[17] Except the empty set, if we reject necessarily uninstantiated forms.

tions into two (not *a priori* exclusive) classes: *doxastic propositions,* which are the propositions about states of belief such as 'that 1+1=2 is believed', and *objective propositions,* which are the propositions made true or false by the state of the physical world. It seems that all possible objective propositions can be believed (unfortunately, the original argument does not elaborate on this). As a consequence, not all doxastic propositions are objective. But if states of belief supervened upon physical states, all doxastic propositions would be objective propositions. Therefore, the targeted kind of supervenience must fail. There are a number of difficulties with this argument but, at the very least, it must be credited the audacity of suggesting that mental states and physical states may make up multiplicities with different mathematical properties.

Michael Detlefsen has published a paper (Detlefsen 2002) in which he utilizes Löb's theorem (Löb 1954) to reveal some difficulties of Computationalism (called *mechanism* by the author). Löb's theorem states that for any consistent arithmetical system Σ and any formula φ, if Σ proves 'if φ is Σ-provable, then φ', then φ is Σ-provable.[18] Detlefsen's most significant conclusion in the referenced paper is that, on plausible assumptions, our proof resources cannot regard themselves both as reliable and as mechanizable. Assume they consider themselves both things, reliable and mechanizable. As they believe they are mechanizable, they regard themselves as subject to Löb's theorem. As they believe they are reliable, they prove for any sentence *s* 'if *s* is provable, then *s*' but they cannot consider themselves able to prove all sentences, because they could hardly consider themselves reliable if they believed they prove a sentence and its negation; but this is obviously incompatible with being subject to Löb's theorem; so they cannot consider themselves subject to such theorem; and we have reached a contradiction: they regard themselves as subject to Löb's theorem and they do not. As far as I can see, the argument has no evident flaw.

Arguments such as Mally's, Kripke's, Kutschera's, Detlefsen's, the original Luna-Small argument or its extension in this paper are likely to look suspect to a number of readers who may distrust metaphysical arguments based on logico-mathematical phenomena, regarding them as extremely likely to contain some fallacious sleight of hand. In some cases these suspicions may rest on a prejudicial belief in the existence of some iron curtain isolating the realm of the metaphysical from the realm of the logico-mathematical. This prejudice may be one in a bundle of inherited beliefs evidencing just inertial resistance to disappear. In the last decades, the analytical tradition has passed from a plain rejection of metaphysics (inspired by neo-positivist empiricism) to a cautious approach to some metaphysical issues, and in this transition it has brought with itself the use of logico-mathematical tools for the treatment of philosophical topics, hence also of metaphysical ones. This can be seen as one of the most significant contributions of the analytical tradition to contemporary metaphysics.

As far as I know, argumentation from logical phenomena to properly metaphysical topics was inaugurated by Mally in 1914 with the argument sketched above. Gödel's famous Gibbs lecture some sixty years ago (Gödel 1951) is a perspicuous case of this type of argumentation. Lucas' and Penrose's Gödelian arguments against computationalism (Lucas 1961, Penrose 1989, 1994), even if not equally esteemed by all, have spurred prolific discussion over decades. Alvin

[18] Löb's theorem relies on a standard form of representing the provability in Σ in Σ's language.

Plantinga (Plantinga 1974) is the most notorious of several philosophers who have used some features of the Kripkean accessibility relation among possible worlds (Kripke 1963) to argue for the existence of God. Patrick Grim (Grim 1988, 1991) has harnessed some topics in set and model theory for theological purposes, namely, to set up arguments against the possibility of divine omniscience. Arguments of this kind have compelled some theists to espouse a sort of *process theology,* in which God is not immutable, so modifying the traditional definition of God.[19]

One of the goals of this paper is to make apparent that this type of argumentation could be a promising line of metaphysical research even if to date it seems suspect to many and is for the most part absent from mainstream metaphysical discussion.

References

Brendel, E. 2001, "Allwissenheit und 'offenes Philosophieren'", *Erkenntnis*, 54 (1), 7-16.

Butler, R.J. (ed.) 1962, *Analytical Philosophy*, I, New York: Barnes and Noble.

Detlefsen, M. 2002, "Löb's Theorem as a Limitation on Mechanism", *Minds and Machines*, 12 (3), 353-81.

Gödel, K. 1951, "Some Basic Theorems on the Foundations of Mathematics and their Philosophical Implications", in *Collected Works*, III, *Unpublished Essays and Lectures*, Feferman, S. *et al.* (eds.), Oxford: Oxford University Press, 1995, 304-23.

Goldstein, L. 2009, "A Consistent Way with Paradox", *Philosophical Studies*, 144 (3), 377-89.

Grim, P. 1988, "Logic and Limits of Knowledge and Truth", *Nous*, 22, 341-67; reprinted in Martin, M. and Monnier, R. (eds.), *The Impossibility of God*, Amherst, New York: Prometheus Books, 2003, 381-407.

Grim, P. 1991, The Incomplete Universe: Totality, Knowledge, and Truth, Cambridge, MA: The MIT Press.

Kripke, S. 1963, "Semantical Considerations on Modal Logic", *Acta Philosophica Fennica*, 16, 83-94.

Kripke, S.A. 1980, Naming and Necessity, Cambridge, MA: Harvard University Press.

Kutschera, F. von 1994, "Global Supervenience and Belief", *Journal of Philosophical Logic*, 23, 103-10.

Loux, M. 1998, A Contemporary Introduction to Metaphysics, New York: Routledge.

Löb, M.H. 1955, "Solution of a Problem of Leon Henkin", *The Journal of Symbolic Logic*, 20, 115-18.

Lucas, J.R. 1961, "Minds, Machines, and Gödel", *Philosophy*, XXXVI, 112-27.

Luna, L. and Small, C. 2009, "Intentionality and Computationalism. A Diagonal Argument", *Mind & Matter*, 7 (1), 81-90.

[19] This seems to be the case for Brendel 2001, though her reference is not just to Grim's but also, and fundamentally, to W.K. Essler's akin argumentation.

Luna, L. and Taylor, W. 2010, "Cantor's Proof in the Full Definable Universe", *Australasian Journal of Logic*, 9, 10-25.

Luna, L. 2013, "Indefinite Extensibility in Natural Language", *The Monist*, 96 (2), 295-308.

Mally, E. 1914, "On the Objects' Independence from Thought", *Zeitschrift für Philosophie und philosophische Kritik*, CLV, 1, 37-52; transl. by D. Jacquette in *Man and World*, 22, 215-31, 1989.

Parsons, C. 1983, *Mathematics in Philosophy*, Ithaca, New York: Cornell University Press.

Penrose, R. 1989, *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*, Oxford: Oxford University Press.

Penrose, R. 1994, *Shadows of the Mind. A Search for the Missing Science of Consciousness*, Oxford: Oxford University Press.

Plantinga, A. 1974, *The Nature of Necessity*, Oxford: Oxford University Press.

Plato, 1992, *The Republic*, translated by G.M.A. Grube, 2nd edition, Indianapolis, IN: Hackett Publishing Co.

Priest, G. 2002, *Beyond the Limits of Thought*, New York: Oxford University Press.

Rayo, A. and Uzquiano, G. (eds.) 2006, *Absolute Generality*, New York: Oxford University Press.

Russell, B. 1903, *Principles of Mathematics*, Cambridge: Cambridge University Press; reedited by Routledge, London-New York, 2010.

Russell, B. 1905, "On Some Difficulties in the Theory of Transfinite Numbers and Order Types", *Proceedings of the London Mathematical Society*, 2 (4), 29-53.

Thomson, J.F. 1962, "On some Paradoxes", in Butler 1962, 104-19.

Shapiro, S. and Wright, C. 2006, "All Things Indefinitely Extensible", in Rayo and Uzquiano 2006, 255-304.

Zermelo, E. 1930, "Über Grenzzahlen und Mengenbereiche. Neue Untersuchungen über die Grundlagen der Mengenlehre", *Fundamenta Mathematicae*, 16 (1), 29-47.