STUDIA SEMIOTYCZNE (SEMIOTIC STUDIES), t. XXXIV, nr 1 (2020), s. 203–232
ISSN 0137-6608, e-ISSN 2544-073X
DOI: 10.26333/sts.xxxiv1.09
Article

DAVID KASHTAN *

DIAGONAL ANTI-MECHANIST ARGUMENTS

SUMMARY: Gödel's first incompleteness theorem is sometimes said to refute mechanism about the mind. §1 contains a discussion of mechanism. We look into its origins, motivations and commitments, both in general and with regard to the human mind, and ask about the place of modern computers and modern cognitive science within the general mechanistic paradigm. In §2 we give a sharp formulation of a mechanistic thesis about the mind in terms of the mathematical notion of computability. We present the argument from Gödel's theorem against mechanism in terms of this formulation and raise two objections, one of which is known but is here given a more precise formulation, and the other is new and based on the discussion in §1.

KEYWORDS: mechanism, mind, computability, incompleteness theorems, computational theory of mind, the cogito, diagonal arguments, Gödel, Descartes, Tarski, Turing, Chomsky.

Mechanism about *X*, roughly, is the view according to which *X* can be understood in terms of a machine. Descartes famously held the doctrine of mechanism with regard to everything but the human mind. More recently, some writers have argued that Gödel's famous theorem about the impossibility of a complete computable axiomatization of arithmetic shows that the human mind is not amenable to a mechanistic explanation.[1] Gödel's theorem is a likely candidate for the job of combatting mechanism, first, because it is very famous; second, because the notion of computability is the modern approach to mechanism about the mind,

[*] Hebrew University of Jerusalem, Edelstein Center for the Philosophy of Science. E-mail: david.kashtan@mail.huji.ac.il. ORCID: 0000-0002-4237-1809.

[1] Most notably (Lucas, 1961; Penrose, 1989; 1994). I refer the reader to (Krajewski, 2020) for a brief recap of the dialectic.

and Gödel's theorem is, or at least entails, a limitative result on it; and third, because it is a diagonal argument, and diagonal arguments seem to be exactly the right tool to wield against a thesis such a mechanism.

However, Gödel's theorem is a precisely formulated and decisively proven mathematical theorem, and mechanism about the mind is a vague and messy philosophical question. Any attempt to bring a mathematical theorem to bear on a philosophical question should be viewed with suspicion. Shapiro, for example, asserts that "there is no plausible mechanistic thesis on offer that is sufficiently precise to be undermined by the incompleteness theorems" (1998, p. 275). The problem, according to Shapiro, lies in the many idealizations that are involved in applying the theorems to humans and to machines. But if Shapiro's condemnation is correct, this is hardly comforting to the mechanist, who should want to resist the Gödelian argument by virtue of being right, not by the vice of being vague. The goal of this paper is to formulate a mechanistic thesis sharply enough that it stands a chance both of being refuted by and of resisting the Gödelian argument.

The purpose of §1 is to get a handle on mechanism about the mind. Starting with Descartes, we review the central epistemological motivations for mechanism, distinguish between programmatic and metaphysical mechanism, and inspect Descartes' reason for denying mechanism about the human mind. Then, we ask whether the advent of computability theory and modern computing machines would have made Descartes change his mind. §2 is about the Gödelian anti-mechanistic argument. Based on the discussion of §1, a sharp criterion is offered for deciding the metaphysical mechanistic thesis, in terms of the computability of sets of mental representations. The classic Gödelian anti-mechanist argument is formulated with reference to this criterion, and two objections to it are raised, one of which is new. In the final subsection a sketch of an alternative diagonal anti-mechanist argument, based on Tarski's indefinability theorem, is given.[2]

1. Mechanism and Anti-Mechanism

The bare term "mechanism", or mechanism about some thing *X*, can be glossed as the claim that *X* is, often despite appearances, essentially a machine. We are interested specifically in mechanism about humans—the question whether humans are essentially machines. Gödel's theorem is thought to have bearing on this question because of its relation to computability theory. The machines in question are, therefore, the special class of computing machines. However, the origins of mechanism, and of mechanism about the mind, lie long before the modern theory of computing machines, in the philosophy of science of several

[2] Michael Goldboim and Balthasar Grabmayr read an earlier draft and helped reduce the number of errors. Those that remain are mine alone. I have also benefited greatly from conversations on the themes of the paper with the aforementioned, as well as with Eli Dresner, Philip Papagiannopoulos and Carl Posy.

important early modern thinkers, most notably Descartes. Descartes' is an interesting case because he both endorsed mechanism about animals and rejected mechanism about humans. By examining his reasons, we can form an idea of why mechanism is attractive, as well as of why it is not compelling. In addition, we may ask whether Descartes' anti-mechanism was not tied to the particular kinds of machines that he knew, and whether the advent of computing machines would have caused him to change his mind.

1.1. Cartesian Mechanism

A classical machine, or mechanical system, roughly, is a finite collection of basic corporeal objects that can move in space and interact through contact, i.e. collision or pressure, and together achieve some desired effect. The properties of the set of basic parts, the sizes and shapes of the objects, and their spatial configuration, we call the basis of the system. From the basis, using the mechanical laws of motion and force, one can calculate the effect; and conversely, if one is interested in a certain effect, one can set up a basis that will achieve it, in other words one can engineer an effect. What allows engineering is the fact that mechanical systems are "bottom-up": that the basis is describable, perceivable and manipulable independently of the effect, and that the effect is "generated" from the basis according to determinate laws.

Mechanism in science or natural philosophy is, first of all, an empirical research program according to which natural phenomena should be studied as though they are effects of mechanical systems. A mechanistic explanation of a phenomenon consists in hypothesizing a basis, and showing that the phenomenon is indeed generated from it by the laws of mechanics. The epistemological virtue of mechanistic explanations is that, since the basis of a mechanical system is describable independently of the effect, they make a positive, self-standing assertion about reality. Such an assertion is straightforward (though not always technically possible) to test, and, in principle, allows nature to be manipulated with design. They, therefore, yield a kind of engineer's, or maker's, knowledge. The contrast is with explanations in terms that are abstract, or that are describable only top-down, in terms of their explanandum, like Molière's *virtus dormitiva*. The basic claim of mechanism is that only positive theories can count as explanations, whereas abstract or top-down theories are explanatorily vacuous.[3]

An explanation is positive rather than vacuous, I propose, when it postulates a basis that has an independent criterion of existence and identity. It should be possible to determine, at least in principle, whether the basis of a hypothesized mechanical system exists in reality without appealing to the properties of the

[3] For comprehensive and detailed accounts of mechanism, especially in its Cartesian brand, see (Gaukroger, 2002; 2007; 2010), also the papers in (Gaukroger, Schuster, & Sutton, 2000). For mechanistic explanation as maker's knowledge, see (Funkenstein, 1986, p. 290).

effect; otherwise the explanation is circular. For Descartes, the basic existent is inert matter, where "inert" means that the spatial extension of material objects, including their motion in space and their collisions with one another, is all there is. Inert matter is a basic existent because spatial shapes, in principle at least, are vividly, distinctly and publicly perceived; and they are bottom-up in the sense that a spatial extension is the "sum" of its parts; the parts are independent of the whole, but not the other way around. Consequently, a mechanical system can be exhaustively described in purely geometrical terms. Mechanistic explanation becomes a kind of geometrical construction.[4]

Mechanism as a research program is thus the call to explain natural phenomena in terms of mechanical systems, and ultimately in terms of geometrical constructions. But aside from being a research program, mechanism is sometimes asserted or denied as a metaphysical thesis. Roughly, metaphysical mechanism about a natural phenomenon *X* is the thesis that *X* is a classical machine, or that the true theory of *X* is a mechanistic theory. Spatially extended matter, on this view, is not only epistemologically virtuous in being vividly perceived or imagined, it is also metaphysically substantial. Metaphysical mechanism about a phenomenon *X* is the claim that *X*, metaphysically, is extended substance, or *res extensa*.[5]

Descartes famously held that animals were, metaphysically, mere machines.[6] This thesis may sound banal to us, but in Descartes' time it was paradoxical and even revolutionary. Supposedly, what made it so unlikely in the eyes of Descartes' predecessors was the seemingly unbridgeable disparity between the behaviors of machines and of animals. In particular, there was the fact that animals and their physiology exhibit spontaneous and organized movement, whereas machines typically do no more than transform external force that is applied to them, and, therefore, cannot move "on their own". This led pre-Cartesian natural philosophy to postulate an intangible life force animating the bodies of animals. The problem with this kind of theory, in the eyes of a mechanist, is that it gives no independent information about the nature of this life force, no way to con-

[4] See (Sepper, 2000; McLaughlin, 2000) for some elaboration. Descartes' notion of the geometrically (as opposed to mechanically) constructible is wider than that of the ancients, but does not cover arbitrary curves. I am unsure whether Descartes' "geometrical" conception of matter is restricted to his geometrically constructible curves or whether it extends to arbitrary curves.

[5] Descartes doesn't, as far as I know, distinguish explicitly between mechanism as a research program and mechanism as a metaphysical thesis.

[6] This formulation is a little misleading. Descartes thought that animal bodies, including human bodies, are machines. As we will see, he did not think the same about minds. Animals did not, and humans did, have minds, so it is in this sense that *non*-human animals are *mere* machines. See (Cottingham, 1978) for more about this.

struct it in the imagination or calculate its properties. In other words, it is a vacuous theory.[7]

Descartes' endorsement of mechanism about animals was motivated by two circumstances. First, around Descartes' time, mechanistic theories of animal physiology were being developed and were achieving remarkable empirical success. Descartes himself proposed extensive theories of this kind, ranging from an account of the heart and blood circulation system, to theories of feelings and the imagination, which Descartes considered part of physiology. The second circumstance was some recent advances in technology, which allowed the construction of self-moving machines, or clockwork automata, operated by a spring or a hydraulic mechanism. Such machines were often used for recreational purposes, and given the shape of a human or an animal. Their "capacity" for self-movement would make them startlingly life-like in the eyes of Descartes' contemporaries, a fact which served to dull the edge of the perceived disparity between animals and machines.[8]

Descartes' metaphysical mechanism about animals was motivated by the science and technology of his time, but neither the empirical success of mechanistic theories nor the advent of new machines can establish a metaphysical thesis. Descartes' own physiological theories turned out to be largely incorrect. For example, though he enthusiastically accepted Harvey's momentous discovery of the circulation of the blood, Descartes rejected the attendant theory of the movement of the heart in terms of muscular expansion and contraction, and favored an account, incorrect as we now know, in terms of "ebullition" (Anstey, 2000, p. 421f). Surely, we don't want to say that an incorrect theory can establish a metaphysical thesis. Likewise, as lifelike as moving statues can get, we know perfectly well that the mechanism behind their movement has nothing in common with the mechanisms behind animal movement. It will be false, then, to say that Descartes' metaphysical mechanism about animals is in any way proven, or even strictly speaking confirmed, by the science and engineering of his time, though it was certainly suggested or motivated by them.

Still, it is not unreasonable to say that the relevance of the two motivating circumstances does spill over a little from the context of discovery to the context of justification. The fact that Descartes' theories were incorrect is less important than the fact that they were mechanistic, which is to say positive, and that they were plausible. It showed that, in principle, mechanistic physiology had a chance to succeed, even if Descartes' own theory happened to be incorrect. Likewise, the existence of moving machines, though they do not simulate the true mechanism of animal movement, shows that self-movement is mechanically possible,

[7] See (Ben-Yami, 2015, Chapter 4) for a less ahistorical discussion of the claim that pre-Cartesian science denied mechanism because the machines it knew did not move on their own.

[8] See Part Five of the *Discourse* (Descartes, 2006) for a summary of Descartes' mechanistic theory of the blood circulation system and his statement of mechanism about animals. Ben-Yami (2015) gives an extended discussion of Descartes' physiology.

and this opens the way to positive speculations about the actual mechanism. We can say, then, that although the science and engineering available to Descartes were a long way off from *proving* metaphysical mechanism about animals, they did provide *positive grounds* for it. Arguably, that's the best metaphysics can hope for anyway.

1.2. Universal Mechanism?

At least as famously, or infamously, as he endorsed mechanism about animals, Descartes rejected mechanism about humans, specifically about the human mind. In the next subsection, we will review his reasons. First, let's see what goes wrong with a seemingly quick and easy argument for mechanism about humans: the argument from universal mechanism.

In the previous subsection, we distinguished between two ways in which mechanism about a phenomenon *X* can be maintained: (a) Metaphysical mechanism is the theoretical claim that the *true* explanation of *X* is mechanistic; (b) Scientific mechanism is the programmatic call to *seek* mechanistic explanations for *X*. Now given the characterization of mechanism sketched above, one may argue that the phrase "mechanistic explanation" is redundant, since an explanation that is not positive, in the required sense, and therefore mechanistic, is no explanation at all. Let's agree to assume this, that is, that all adequate explanations are mechanistic. In addition, one may wish to deny the possibility that some natural phenomena are not amenable to explanation at all. Let's assume this as well, without discussion. From these two assumptions it is tempting to conclude a kind of *universal mechanism*—the claim that every phenomenon has a mechanistic explanation. From this, metaphysical mechanism about the human mind seems to follow immediately.

There are two main problems with this line of reasoning which it will be instructive to uncover. The first concerns the logical form of the inference. On the face of it, we have here a run-of-the-mill universal instantiation: From mechanism about all phenomena we infer mechanism about the particular phenomenon of the human mind. However, such an inference holds only if the instance is in the range of the quantifier, in this case if the human mind is a natural phenomenon that stands to be explained. The problem is that what counts as a natural phenomenon is not a simple and non-negotiable matter. To state the issue clearly, let's distinguish between the *data*, which is immediately given (by the senses, say), and the *phenomenon*, which is that which we need to explain. The phenomenon is extracted, or constructed, from the data, by various conceptual operations we can call *idealizations*, which consist primarily of extending the scope of the phenomenon beyond what has actually been perceived in the data, and of cleaning it up of factors that supposedly belong to the measuring procedure or to external factors, and not to the phenomenon itself. Which idealizations are to be applied is an issue that can be negotiated, and different decisions affect the domain of the quantifier in the statement of universal mechanism.[9] Thus, whether we are prepared to accept the inference from universal mechanism to mechanism about the human mind ultimately depends on whether or not we accept the mind as a phenomenon to be explained.

One way the explanatory burden of mechanism can be reduced is by "eliminating" some would-be phenomenon. For example, pre-Cartesian natural philosophy considered vital processes in animals a phenomenon to be explained. What the data contained, however, was not the vital processes themselves, but observations of seemingly organized spontaneous movement in animal bodies. Mechanistic physiology does not explain vital processes in mechanistic terms ("reduce" them to mechanics), it rather rearranges the data so that vital processes cease to count as a phenomenon (they are "eliminated"). The data is kept the same, but it is idealized differently, into a phenomenon of spontaneous movement, which yields more easily to explanation in terms of inert matter.[10] In a similar fashion, the inference from universal mechanism to mechanism about the mind can be avoided if we take the mind out of the domain of the quantifier. Then, there is simply nothing there to explain. Now, certainly the exclusion of recalcitrant data from the domain of the explanandum flirts dangerously with question-begging. However, since some idealization of the data is anyway unavoidable, ultimately the legitimacy of elimination turns on whether it can be motivated independently of the recalcitrance of the data, and on whether what is left to explain is interesting enough.

The second problem with the argument from universal mechanism is that it is too cheap. Scientific mechanism is predicated on the distinction between positive and vacuous explanations, and on the rejection of the latter from science. Metaphysical mechanism is a metaphysics guided and supported by scientific mechanism. Descartes' metaphysical mechanism about animals, for example, was grounded in positive (though incorrect) physiological theories and actual engineering techniques. But the inference we are now considering proposes that we accept mechanism about the mind on the basis of a general principle, in complete absence of any positive theory of the mind, or of any machine that can simulate it. Such an inference goes against the very grain of mechanism. This doesn't make it a logical fallacy, but it should make us uneasy about accepting its conclusion. The mechanism we end up with is a vacuous doctrine, and we should not be satisfied with it.

We have cited two reasons why mechanism about the mind should not be accepted on the basis of the argument from universal mechanism. Descartes, however, goes farther and rejects universal mechanism altogether.

[9] See (Bogen & Woodward, 1988) for the classical modern statement of the distinction between data and phenomenon, and (Woodward, 2011) for a summary of the ensuing discussion.

[10] This is, at least, how Gaukroger presents things in (Gaukroger, 2007, p. 323ff).

1.3. Cartesian Anti-Mechanism

In the *Discourse*, after having stated his thesis that animals are mere machines, Descartes goes on to say:

> [I]f any such machines resembled us in body and imitated our actions insofar as this was practically possible, we should still have two very certain means of recognizing that they were not, for all that, real human beings […]. The first is that they would never be able to use words or other signs by composing them as we do to declare our thoughts to others. For we can well conceive of a machine made in such a way that it emits words, and even utters them about bodily actions which bring about some corresponding change in its organs […] but it is not conceivable that it should put these words in different orders to correspond to the meaning of things said in its presence, as even the most dull-witted of men can do. (Descartes, 2006, p. 56)[11]

In this passage Descartes puts forth a test for deciding that a humanoid machine is not a genuine human. The claim is that language is a reliable indicator of the presence of mind, and that no machine can simulate human linguistic competence. Note, that Descartes is comfortable with machines voicing sentences, even as a response to stimulation; but that would stop short of genuine linguistic capacity, which consists, first, in the ability to form indefinitely many sentences, and second, in the fact that these sentences are used in accordance with their meaning, hence, that they are meaningful. In other words, we should not consider a mechanistic theory to be a theory of the mind, if it does not account for the syntactic and semantic aspects of language use.[12]

Unfortunately, Descartes doesn't explicitly say why he thinks linguistic capacity resists a mechanistic explanation. Here's a conjecture. Although there is no difficulty in imagining a mechanistic theory that accounts for sound and voice,[13] there is no way to calculate the syntactic and semantic properties of an utterance from its acoustic or even phonological properties alone. Semantic phenomena cannot even be described, let alone explained, in phonological terms. Consequently, there is a basic incongruity between the explanatory resources of mechanism and the phenomenon of linguistic competence, an incongruity that makes genuinely linguistic machines unimaginable for Descartes. Nor does Descartes think that the option of eliminating the mind is open to us (below we'll see why). Linguistic capacity, and with it the mind, presents an ineliminable phenomenon which the explanatory resources of mechanism simply have no chance of accounting for.

[11] In the text, Descartes mentions another test for humanity, which (Gunderson, 1964, p. 199) calls "the action test". I shall not address it here.

[12] See (Gunderson, 1964) for an extensive discussion.

[13] This was, in fact, one of the earliest mechanistic theories, by Beeckman, see (Cohen, 1984, Chapter 4).

It has been suggested that the problem here has more to do with what Descartes can and cannot imagine than with any objective incongruity between machines and the mind. Descartes was familiar with a certain type of machine, and his imagination, remarkable though it was, was inevitably limited to that type. Recall how the pre-Cartesian anti-mechanists, according to the story as we've told it, had difficulty imagining that animals were machines because the machines they knew could not move about on their own. This limitation to the imagination was removed by the appearance of self-moving statues. Similarly, the development of new machines with previously unforeseen features, namely modern digital computers, might provide positive ground for mechanism about the mind. This issue will be taken up presently.[14]

But apart from the perceived incongruity between the linguistic phenomenon and the explanatory resources of mechanism, Descartes also had a more properly philosophical argument for his anti-mechanism, the *cogito*. Although the *cogito* is not expressly presented as an anti-mechanistic argument, its anti-mechanistic import is easy to establish. Briefly, since mechanism for Descartes is metaphysically limited to *res extensa*, it suffices to find one thing which is not *res extensa* in order to refute universal mechanism, even in its vacuous form. The *cogito*'s twin conclusions are, first, that the self[15] exists, and second, that it is a kind of thinking substance, or *res cogitans*, and not *res extensa*. This self is identified with the mind, and it follows that the mind cannot be a machine.

The *cogito* is an interesting case because it resembles a diagonal argument, but on closer inspection it isn't.[16] It resembles a diagonal argument (a) in the form of its conclusion, and (b) in the structure of the argument, as follows. (a) Like many diagonal arguments, the *cogito* (on its anti-mechanistic reading) purports to refute a completeness claim by producing an "outsider" element. For example, Cantor's diagonal proof of the indenumerability of the real numbers refutes the claim that there is a complete enumeration of the reals by producing, for each enumeration, an outsider. In Descartes, the completeness claim is that all things are *res extensa*, and the outsider element is the human mind, or the thinking self. (b) Diagonal arguments typically construct an outsider element by applying a procedure involving self-reference and negation to all members of the putatively complete class. Cantor shows how to construct, for a given enumeration, a real number based on the negation, in the relevant sense, of all members of the enumeration. Descartes' procedure is to doubt the reality of all extended substances; but when he arrives at his own self, he finds that the procedure fails:

[14] The claim that technological advancements might affect our assessment of Descartes' anti-mechanism is suggested by the discussion in (Ben-Yami, 2015, p. 126f).

[15] Or the *I*, or whatever. The *cogito* is awkward to report in the third person.

[16] I defer a more detailed treatment of the *cogito* for another occasion. See (Slezak, 1983; 1988) for a different take on the *cogito*'s being a diagonal argument, (Sorensen, 1986) for a critique.

> But I have convinced myself that there is absolutely nothing in the world, no sky, no earth, no minds, no bodies. Does it now follow that I too do not exist? No: if I convinced myself of something [or thought anything at all] then I certainly existed. (Descartes, 1996, p. 25)[17]

The thinking self is discovered when we try to include it in the domain of our skeptical procedure and fail. However, only the thinking aspect is immune from doubt in this way. It follows that a non-extended object exists.

However, on closer inspection neither the form of the argument in the *cogito*, nor the form of its conclusion, are those typical of diagonal arguments. We show this, again, (a) for the form of the conclusion, and (b) for the structure of the argument. (a) Diagonal arguments typically show the existence, not of an absolute outsider, but of a method to generate an outsider given a particular completeness claim. For example, Cantor does not show that there is a real number which absolutely cannot be enumerated. That's absurd. Rather, he shows how to find, for every enumeration, a real that's outside of it, even if it does belong to some other enumeration. By contrast, Descartes' *res cogitans* is meant to be an absolute outsider, not being captured by any mechanistic system.[18] (b) Diagonal arguments construct the outsider using a negative procedure on the members of the putatively complete system. The identity of Cantor's outsider element for an enumeration *E* is a function of all elements of *E*, negating, as it were, every one of them, and thereby establishing its distinctness from them all. By contrast, in the *cogito*, no diagonal element is constructed. The stage in the *cogito* that resembles the diagonal procedure, quoted above, is the one in which doubt is applied to all things (step A: "I convinced myself that there was nothing at all in the world, no sky, no earth, no minds, no bodies"), including, tentatively, the self (step B: "did I therefore not also convince myself that I did not exist either?"), but unsuccessfully in the latter case (step C: "certainly I did exist, if I convinced myself of something"). Here, what sanctions the inference from self-doubt (B) to self-affirmation (C) is the fact that doubting in general implies the existence of the self, regardless of the object of doubt. But doubting in general was performed already in step (A). Therefore (C) could have been inferred directly from (A). The act of self-doubt (B), ostensibly the diagonal heart of the *cogito*, doesn't play any logical role in the argument. It is primarily an expository device, serving to highlight, but not to establish, the existence of the self. In diagonal arguments, by contrast, the diagonal construction is an essential step of the inference.

[17] The interpolated part is from the French version.

[18] The difference is rooted in the respective claims that diagonal arguments and the *cogito* purport to refute. Diagonal arguments typically refute existential claims. In Cantor's case, the claim is not of the form "every real number is thus and so", but "*there is an enumeration such that* every real number is thus and so". By contrast, Descartes' *cogito* purports to refute the claim "every existent is extended", or something along these lines. This is why a diagonal argument yields only a relative existence claim, and the *cogito* purports to yield an absolute existence claim.

The conclusions from this brief discussion are as follows. First, we see that the *cogito* resembles a diagonal argument, but turns out upon scrutiny not to be one. Second, the aspect in which it fails to be a diagonal argument is exactly the point at which it loses much of its force, since the thinking *I* has not been given a definite enough constitution in order to count as a genuine existent.[19] If this is correct, then we get the impression, wildly anachronistic though it may sound, that Descartes is here groping for a diagonal argument, in a hunch that this is the kind of argument that can refute mechanism about the mind.

1.4. Turing's Computing Machines

With Descartes' pseudo-diagonal anti-mechanist argument out of the way, we can come back to the question, raised in the middle of the previous subsection, whether computing machines can provide positive support for mechanism about the mind in a way that classical machines could not. In order to begin to answer this, we have to state clearly what distinguishes the two kinds of machines.

Today, the term "computing" already implies "machine", but originally the notions were only indirectly related. Computation was just another name for calculation. We get an intuition about what calculation is by looking at a simple case, the grade-school algorithm for addition. The fundamental way to add two numbers, e.g. 13 and 28, is to produce collections, of fingers, say, with the corresponding cardinalities, and then count the members of their union. But counting is not always a good option, and for practical purposes we usually turn to methods that exploit properties of the numerical notation. The positional notation system for numbers, for example, allows us to perform sums of arbitrarily large numbers in terms of the iteration of the operation of summing up two single-digit numbers, in our example case first 3 and 8, and then 1 and 2. Since the possible sums of two single-digit numbers are few, they can be memorized or written down in a small instruction table. When we appeal to such a memorized or written table, reference to the numbers themselves effectively drops out. The table simply instructs us, when we see the digits "3" and "8", to write "1" below them and mark a carry above the next column. What is in play here are the digits, not the numbers 3 and 8, since these latter were not part of the original problem at all. The procedure is iterated for every position of the numerals, resulting in a string of digits, in our example case "41", which we then interpret as referring to a number.

[19] This point deserves elaboration, which, for reasons of space I shall not provide. Briefly, the problem is that the *I* that is discovered in step A is abstract, or vacuous, in something like the sense of the previous subsection, and, therefore, cannot be the basis of a genuine existence claim of the kind that the *cogito* aims to establish. This difficulty is, I believe, recognized by Kant in his discussion of the "*I think* that accompanies all my representations" in the *Critique of Pure Reason*. See esp. the discussion in §16 of the B-deduction (B131ff), where the necessity of the *I think* is affirmed, and in the Paralogisms of Pure Reason (A341/B399ff), where the substantiality of the *I* is denied.

Strictly speaking, therefore, the numbers themselves are only present at the entry and exit points of the calculation, but are completely absent during the calculation itself. The process of calculation is mechanical, in a sense quite close to the one used in relation to machines and mechanism. What it applies to are (usually) marks on paper, and it is sensitive only to their visible geometrical properties. The operations that we perform (writing further symbols) are also definable in terms of geometrical properties. In other words, calculation is performed almost purely within the bounds of *res extensa*. There are two exceptions: the interpretation of the symbols at the input and output points. The meaning of symbols cannot be counted part of their *res extensa* aspect, and yet without acknowledging their meaningfulness, we can't think of our procedure as calculation over and above the mere producing of marks on paper.[20]

What the mechanical nature of calculation provides is epistemic security. Since each step can be written down, surveyed in a glance and compared with the table of instructions, any mismatch will be evident and public. Many philosophers, most notably Leibniz, have entertained the hope of enjoying the epistemic virtues of calculation in domains beyond mathematics. Such a project requires a notation system in which the properties of the subject matter are reflected in the form of the symbols. The formalized languages of modern logic, for example Frege's *Begriffschrift* and Hilbert's deductive systems, can be seen as systems of generalized calculation, designed to be applied to any subject matter. However, these systems implement one particular method of calculation, and the question of a general analysis of the concept of calculation is left open.

Historically, the need for such a general analysis became pressing with the appearance of Gödel's incompleteness theorems. Gödel showed that any formal system, the syntax of which can be captured by recursive functions, and which can represent recursive functions in an appropriate sense, is incomplete. Hilbert's arithmetical system fulfilled these conditions, and was thereby proven incomplete.[21] However, since recursive functions were only one particular form of calculation, there was doubt (at least, Gödel doubted) whether the theorems would apply to any formal calculus whatsoever. This doubt was resolved by Turing's general analysis of calculability, in terms of imaginary computing machines. Turing's machines operate on written symbols, and their simple and highly scalable design enables a mathematical definition of various classes of relations on strings, most importantly the class of computable and computably enumerable (c.e.) relations. It was proven that the class of recursive functions on strings is exactly the class of Turing-computable functions on strings, which showed that Gödel's results hold for all systems with computable syntax. Tu-

[20] This "formal symbol manipulation" account of computing has come under attack in recent decades, e.g. in (Smith, 2002; Fresco, 2014, who provide many further references). However, this literature is mostly concerned with the concept of physical computation, especially in the context of computational cognitive theory, so it is irrelevant here. (It will become more relevant in the next subsection).

[21] Frege's system had, of course, other problems.

ring's account was accepted (in particular by Gödel) as definitive of the concept of computability.[22]

With respect to this history we ask: (a) What role does the notion of a machine play in Turing's account? (b) What convinced Gödel of the account's validity and generality? Regarding question (a), we note that calculation by itself, as exemplified above in the long addition algorithm, makes no reference to machines, though it does depend on the notion of the mechanical. We note also that of the general accounts of calculation that preceded Turing's or were independent of it—Church's lambda calculus, Gödel's and Kleene's general recursive functions, Post's production systems—none mentioned machines in any way. In fact, if we look closely at Turing's account, we see that the reference it makes to machines is, strictly speaking, superfluous. The actual Turing machine construction is nothing but a straightforward generalization of the long addition algorithm, arrived at by simplifying the instruction table to very basic operations. There is no compulsion to describe it as a machine, and we can just as well describe it in psychological terms (as Post did). On the face of it, then, the answer to question (a) is that the notion of a machine plays no role at all in the content of Turing's analysis of computability.

However, when we ask about the perceived conceptual superiority of Turing's account over the other accounts (question (b)), it is hard to avoid the impression that it is due precisely to Turing's appeal to the notion of machine. The point of this appeal is to convince the reader that the procedure described remains squarely within the bounds of the mechanical, or *res extensa*, and, therefore, guaranteed to have the epistemic virtues associated with paradigmatic calculation. The fact that a machine can "perform" Turing's generalized algorithms promises that only geometrical properties of the objects of calculation are appealed to, and in particular that no semantic properties are exploited. This is, however, ascertainable even without the machine trope, for example if we interpret Turing's algorithms as instructions for human computers. The machine trope ultimately has just an auxiliary expository role in Turing's argument.[23]

In order for the machine trope to fulfill its expository role, the machines in question have to be exactly the classical mechanical machines, the ones that Descartes was thoroughly acquainted with. There is no impediment to realizing

[22] This story is told in many places, for example in (Sieg, 2013).

[23] For Gödel's well-known endorsement of Turing's account, see for example the 1951 essay in (Gödel, 1986): "The most satisfactory way [of arriving at a definition of computable function of integers] is that of reducing the concept [...] to that of a machine with a finite number of parts, as has been done by the British mathematician Turing".

See (Sieg, 2013, §1) for more quotes by Gödel to a similar effect. See also (Sieg, 2001) for a more general discussion. Gödel does mention machines already before the appearance of Turing's work. In his 1933 paper (Gödel, 1986), inference rules in formalized languages are characterized as: "[P]urely formal, i.e. refer only to the outward structure of the formulas, not to their meaning, so that they could be applied by someone who knew nothing about mathematics, or by a machine".

Turing's machines with gears, chains and a crank rather than scanners, printers and tapes. Indeed, but for their scalability, Turing's machines would be much simpler to engineer than the average moving statue. But if computing machines are not essentially different from mechanical machines, why do we have the impression that they stand a better chance than the latter of simulating human cognition? Conversely, why did Descartes fail to see that classical machines are capable of such simulation?

The answer to this puzzle has been given in our discussion of the long addition algorithm above. There, we mentioned two points in the procedure of calculation that went beyond the merely mechanical: these were the input and output points, at which the meanings of the symbols were appealed to. If we ignore these semantic limit points, the computer, whether human or machine, cannot properly be said to compute, but only to be moving bits of *res extensa* about (or rather, to be bits of *res extensa* moving about). It is the person writing the input on, and reading the output off, the tape who is performing the computation, using the machine as they would a slide rule or a sophisticated abacus.[24] In order to treat machines as genuinely computing, we have to revise our notion of a machine. We now include the semantic limit points of the computation procedure within our notion of computation, and accordingly, we include the interpretation of the symbols on the machine's tape as belonging to the machine. Thus, we distinguish between, on the one hand, Turing's machines, which are strictly mechanical and used by Turing as an expository trope in his classic account of (human) calculation; and on the other, Turing machines which are to Turing's machines as a closed interval is to an open one—they include the limit points. Turing machines are not just bits of *res extensa* moving about, for their symbols are not mere geometrical shapes—they are symbols properly so-called, i.e. bits of *res extensa* that designate things. Only in this sense can we say that the machine returning, say, the string "41" upon receiving the strings "13" and "28", computes addition. It is the concept of a Turing machine, and not a Turing's machine, that stands a chance at simulating human cognition.

Recall, Descartes' worry was that the semantic aspect of linguistic competence was incongruent with the explanatory resources of mechanism. On our new understanding of computing machine, the semantic layer is built-in. Should Descartes be satisfied? Probably not. The new machines might be said to compute, but they also explicitly go beyond the purely mechanical. It is therefore not correct to say that we have shown, by appealing to computing machines, that mechanism about the mind as Descartes understood it is tenable. Rather, we have changed the subject.

[24] See the first three sections of (Papayannopoulos, 2020) for a more detailed statement of roughly this view.

1.5. The Cognitive Inversion

The main empirical enterprise that makes use of the theory of computability as the basis for a mechanistic outlook is the computational theory of mind. This theory emerges from the concept of computation through two conceptual twists.

First, as we recounted above, the notion of computation originally referred to a species of conscious human activity, not something specifically related to machines. Admittedly, due to the mechanical or rote character of computation, no emphasis needed to be placed on "conscious". But calculating is something a person does intentionally, not something done in the background. In Turing, I claimed, the appeal to the idea of a machine was just an expository trope. However, once it was shown in theory what such machines could do, it was a (relatively) small step to building them in practice. The success of this project was so overwhelming that the word "computing" and its cognates came to be associated exclusively with machines. This is a conceptual twist. Turing exploited the fact that both human computation and machines are mechanical, in order to argue that his mathematical model captures human computation. Actual computing machines do the conceptually opposite—they exploit the mechanical character of human computation in order to take over the rote part, leaving humans the sole job of interpreting the results.

Once actual physical computers became available and familiar, it was a (relatively) small step to using them as a model for human unconscious cognitive activity. In digital computers we distinguish between the hardware, which is the machine itself, and the software, which is, roughly, a representation of the design of the machine which abstracts from any physical implementation. This distinction allows us to implement several abstract machines on a single physical machine.[1] Analogously, one is tempted to view the brain as a piece of naturally developed computer hardware, on which various software programs are implemented. The software corresponds to the human mind. Cognitive phenomena are then explained in terms of software implemented in the brain. In this way the notion of computation, which started out as pertaining to the human mind, made its way back to the mind after having been appropriated by machines. And it came back transformed, being now postulated to underlie the whole of human unconscious cognitive makeup, rather than being one type of conscious human activity.[2]

The idea of a computing machine thus provides a basis for a methodlogically sound and empirically fruitful science of the human mind. Does this provide support for metaphysical mechanism about the mind in the same way that mechanistic physiological theories provided support for metaphysical mechanism about animals?

[1] For complications regarding the hardware/software distinction, see (Duncan, 2017).

[2] See (Gardner, 1987) for a detailed history of cognitive science (with relation to the present paragraph, see pp. 16ff, 40f, 138ff, 384ff).

Ideally, a full mechanistic theory of the mind would show how cognitive phenomena follow from the physical description of the brain in the same way that the behavior of a digital computer follows from its physical makeup. However, currently at least, we are nowhere near such a full derivation. In practice, cognitive science makes progress by abstracting from the physical implementation, the hardware, of the mind, and studying just the software, the system of conscious and unconscious mental processes and representations that underlie human capacities and behavior. For example, Chomsky explains the cognitive phenomenon of linguistic competence by postulating a specialized language faculty, described as an "abstract linguistic computational system" which is "an internal component of the mind/brain" (Hauser, Chomsky, & Fitch, 2002, p. 1570f). By "abstract" what is meant, supposedly, is that the physical implementation of the computational system is abstracted from.

This strategy, of abstracting from the physical implementation, has implications for our question. Computation, as we have understood it in the previous subsection, is the mechanical manipulation of strings of symbols, such that the symbols are visually, or anyway sensibly, individuated, in other words that they are *res extensa*. This bottom-up character is what endows computation with its epistemic virtue, and also what connects it with the doctrine of mechanism. With computing machines, things become a little more complicated because we add a non *res extensa* layer, the level of interpretation; but the bottom-up character of computation is kept, because the symbols are still manipulated strictly mechanically. However, in the mind there is nothing that corresponds to the visible strings of symbols of conscious computation, and the abstraction from hardware eliminates all reference to the physical substrate of the computing machine. The representations that the computations of cognitive science operate on are individuated top-down, by their systematic contribution to the computation of the phenomena, or in other words, functionally. Such functionalism does not, perhaps, invalidate cognitive science as a science; but it seems to waive the mechanistic demand for positivity.[3]

Another way to state the problem is this. As far as I know, Chomsky never fully specifies what is meant by "/" in "mind/brain" in the quote two paragraphs above. The intention is probably to flag the fact that, although abstract representations are immaterial, no metaphysical dualism is thereby implied, since we expect them to be reduced to neural terms at some time in the future. On this understanding, the "brain" in "mind/brain" is a promissory note to the effect that cognitive science (in this case linguistics) will, at some point, be reconciled with mechanism proper. Read in this way, the phrase "mind/brain" is an implicit endorsement of mechanism about the mind. However, the promissory note is a rain check, not motivated by any existing positive theory of the relation between

[3] See (Miłkowski, 2013) for a relatively nuanced discussion of mechanism about the mind in the context of cognitive theories and the so-called new mechanism. Miłkowski's view of computation is not the one presented in the previous subsection.

linguistic capacity and anything in the brain; it is motivated solely by the widespread conviction that everything cognitive has its seat in the brain. This conviction is too widespread to doubt, at least in certain prominent circles, but this is not the same as being a positively discovered empirical fact. To the extent that it purports to be more than just a methodological injunction to seek explanations of cognitive facts in the brain, in other words, to the extent that it presumes to be a metaphysical thesis, it is a vacuous mechanism, in the sense of §1.2.[4]

In this subsection and the previous we considered, all too briefly, the notion of a computing machine and its application to the empirical study of the mind. The question was whether computing machines can provide positive ground to mechanism about the mind in the same way that moving statues and mechanistic physiological theories grounded mechanism about animals for Descartes. On the one hand, the new conception of computing machine, and the cognitive science built upon it, give up on many of the features that made mechanism attractive—the inertness of matter, the bottom-up derivation of phenomena, etc. On the other hand, the empirical success of cognitive theory, as well as the engineering achievements in the field of computing machines, show that the new conception is stable and fruitful. I leave the issue undecided. I turn now to consider whether we cannot find a principled argument against mechanism in Gödel's incompleteness theorem.

2. The Gödelian Argument

Our foregoing exposition of computational mechanism about the human mind was not sufficiently precise for Gödel's theorem to be applied to it. In the present section, our tasks are, first, to provide a sharp(ish) formulation of mechanism; second, to give a correspondingly sharp rendering of Lucas's famous Gödelian anti-mechanist argument; and finally, to topple this argument from several angles. In closing, I shall sketch a diagonal argument that I think stands a better chance.

2.1. Mechanism About the Mind

Mechanism about the mind, on the construal sketched in §1.5 above, is the claim that every natural aspect of human cognition, or in other words every cognitive phenomenon, is the result of a computational system that is part of the mind. Cognitive science studies many phenomena that don't manifest themselves through language, but they are arguably not relevant to the issue at hand, and

[4] This is not to say that no connection has been made between linguistic theory and the brain sciences. See in particular the findings reported in (Grodzinsky & Santi, 2008) and papers cited therein. However, it is clear that these findings are very far from sufficient to metaphysically ground Chomskyan theory in the brain sciences. They should rather be considered the fruit of Chomskyan mechanism, where the latter is viewed as a research program, not a metaphysical thesis.

I put them aside. We therefore think of a cognitive phenomenon as a set of sentences uttered or assented to by speakers, and of a mechanistic explanation as an algorithm that enumerates the set.

We assume that natural languages are fully intertranslatable, and identify them all with a single language *L*, which we assume for convenience is a formalized first-order predicate language. In addition, we assimilate to *L* the language in which the mental computations are carried out, the Language of Thought, as it were. Such assumptions seem to be implicit in much of the practice of (linguistically oriented) cognitive science. Let S_L be the set of sentences of *L*. Cognitive phenomena are associated with certain subsets of S_L, which I shall call their yield. Scientific mechanism is the call, given a cognitive phenomenon *A*, to look for an algorithm that enumerates its yield S_A. Metaphysical mechanism is the claim that the mind really is computational, which we express as follows:

Metaphysical Mechanism: Let *L* be the language of the mind. If $S_A \subseteq S_L$ is the yield of a natural human cognitive phenomena *A*, then S_A is computably enumerable (c.e.).

By our assumptions, S_L itself is infinite and c.e. (in fact, computable). Being infinite, it will have non-c.e. subsets. On pain of trivializing the question, mechanism therefore cannot be equated with the claim that all subsets of S_L are c.e. We need some means of characterizing the class of subsets that are of interest, i.e. that correspond to cognitive phenomena.

Cognitive science is an empirical discipline, one that studies phenomena as they are given in observation and experiment. Let's further restrict the experimental paradigm to that in which sets of sentences (or judgments about sentences) are collected from subjects, whether they occur naturally or through elicitation. In practice, experiment and observation can only give rise to finite sets of perceived sentences. Again, on pain of trivializing the question, we cannot assume that the yields of phenomena are finite, since finite sets are trivially computable. To bridge the gap between the finitude of the sets actually given, and the infinitude of the sets yielded by cognitive phenomena, we call on our previous distinction (§1.2) between data and phenomena, and on our notion of idealization whereby the latter is constructed from the former. This allows us to consider infinite subsets of S_L as cognitive phenomena. Clearly, much hangs on which idealizations are allowed.

In order to have an actual case before our eyes, let's think again about Chomskyan linguistics, and in particular the cognitive phenomenon of linguistic competence. The premise of linguistics is that there is a language faculty which computationally enumerates, or "generates", the set of sentences that speakers accept as grammatical in their language. It is the job of the linguist to find the generating algorithm.[5] The phenomenon of linguistic competence is associated with an

[5] See, e.g., (Chomsky, 1957; 1975).

infinite set S of sentences. The data available to the linguist has to be finite.[6] S is therefore the product of idealizations performed on a finite data set D. The operations involved are roughly two: extrapolation to an infinite set, and cleaning up the data into a well-behaved collection that exhibits enough regularity to be studied scientifically.

Let's look closely at an (unrealistic) example of idealization. Let s_J be the sentence:

(1) John is very, very tall.

Let $s_J\{n\}$ be the result of replacing "very, very" in s_J with a string of n times "very". Since data sets are finite, no data set will contain $s_J\{n\}$ for every n. However, clearly we could procure a data set D_J such that $s_J\{n\} \in D_J$ for every n less than some integer k, i.e. with no gaps below k. It seems reasonable to extrapolate D_J to a set S_J, such that $s_J\{n\} \in S_J$ for every n. Note that, clearly for some integer l, sentences $s_J\{n\}$ longer than l will not appear in any data set. They will simply be too long. But this shouldn't discourage us from accepting S_J, since we can reasonably ascribe the absence to factors that lie outside the language faculty proper, for example to constraints on memory or on patience.

In addition, and especially if D_J is drawn from a corpus of naturally occurring speech rather than elicited speaker judgments, there may be sentences in D_J that we refrain from carrying over to S_J. Naturally occurring speech is the product of many heterogeneous factors, the language faculty being just one. The subject may be distracted midsentence, or interrupted, or there might be another reason for us to decide that an observed sentence does not reflect a genuine product of our language faculty. In this way not only will there be many sentences in S_J that were not directly given in D_J, but also sentences that were given can be filtered out of the phenomenon to be explained. S_J is, therefore, both extrapolated and pruned, relative to the data set D_J.

This doesn't mean that anything goes. There have to be constraints on which extrapolations and which prunings are legitimate, constraints which I shan't attempt to specify precisely. Instead, let me give an example of an illegitimate idealization. Consider the following experiment. The subject is presented with a natural number n, and is asked to form a grammatical sentence with n words. This is a task that linguistically competent subjects can perform with ease, for example by giving $s_J\{n - 3\}$ as an answer (for $n > 2$, of course). Let D_B be the set of sentences actually collected in the experiment, a finite set. Now let B be some non-c.e. set of natural numbers with an initial segment identical with the set of lengths of sentences in D_B. Finally, let S_B be a set of sentences such that $D_B \subset S_B$, and such that if $s \in S_B$ and n is the length of s, then $n \in B$. In other words, B is the set of lengths of sentences of S_B, and S_B is an extrapolation of D_B, conditioned by

[6] See (Pullum & Scholz, 2010) for a critique of the assumption that an infinite set has to be assumed.

B. It follows that B is Turing-reducible to S_B—all we need to do is count the lengths of sentences in S_B—and therefore that S_B is non-c.e. But S_B was extrapolated from the (imagined but) plausible data set D_B. If this extrapolation is a legitimate idealization, then S_B is the yield of some cognitive phenomenon, and thus a counterexample to computational mechanism.

Clearly, S_B is not a legitimate idealization of D_B. I will not attempt a statement of general conditions on legitimate idealization, but the condition that this example suggests is that extrapolation has to preserve and continue trends existing in the original set. The concept of a trend and its continuations is not a very precise or determinate notion, but since B was chosen arbitrarily, it obviously does not fit the bill. In the case of D_B, the idealization is conditioned by a set that is completely external to the mind, so the fact that it is not c.e. is not a counterexample to mechanism after all. More generally, when we idealize a data set into a phenomenon, the principle that guides the idealization must somehow reflect the situation in the mind.

2.2. Gödel's Theorem and Its Proof

With this sharpened statement of mechanism in hand, let's turn our attention to the Gödelian argument. First, let's review Gödel's theorem and proof.

Let L_T be a formalized language.[7] A set $T \subseteq S_{L_T}$ is a theory if it contains the logical axioms and is closed under the logical inference rules; it is a formalized theory if it is, in addition, c.e.; and it is said to contain arithmetic if it contains the Peano axioms (L_T is therefore implied to contain the language of arithmetic).[8] T is consistent if $T \neq S_{L_T}$, equivalently if for no L_T sentence α: $\alpha, \ulcorner\neg\alpha\urcorner \in T$.[9]

Theorem G1: *If T is a consistent formalized theory that contains arithmetic, then we can compute from (the algorithm that enumerates) T a sentence $g_T \in S_{L_T}$ such that g_T, $\ulcorner\neg g_T\urcorner \notin T$.*

Proof: Let $c(x)$ be a mapping from numbers to L_T sentences, called the coding scheme. $\ulcorner\overline{n}\urcorner$ is the numeral in L_T that refers to the number n.

[7] As before, I limit consideration to standard first-order languages.

[8] It is possible to state the theorem also for weaker conditions, but this will not affect the argument.

[9] The corner quotes are used in order to form names of expressions by concatenating symbols with other names of expressions. "α" in the text is a variable over sentences; "$\ulcorner\neg\alpha\urcorner$" is a function taking a sentence and returning its negation. The source is (Quine, 1940, p. 33ff), though I also allow constant names (e.g., "g_T" below) to occur in corner quotes. This is not exactly the same as the (more common) use of corner quotes to signify Gödel codes.

Lemma 1 (Reflection): *There is a unary formula $PRV(x)$ of L_T, such that for every n:*

$c(n) \in T$ if and only $\ulcorner PRV(\bar{n})\urcorner \in T$.

Lemma 2 (Diagonalization): *There is a number k (which can be computed from T) such that:*

$\ulcorner \neg c(k)\urcorner \in T$ if and only if $\ulcorner PRV(\bar{k})\urcorner \in T$.

From the two lemmas it immediately follows that:

(2) *$c(k) \in T$ if and only if $\ulcorner \neg c(k)\urcorner \in T$.*

We put g_T for $c(k)$. By consistency of T, the theorem follows. □

The sentence g_T effectively says of itself that it is not in T. Consequently:

Corollary: *Under the hypothesis of the theorem, g_T is true.*

2.3. The Gödelian Anti-Mechanist Argument

The Gödelian argument applies theorem G1 to the sharp statement of mechanism given in §2.1. The language in question will be the general language of cognition L. Call a set $T \subseteq S_L$ Gödelian if it is a consistent formalized theory that contains arithmetic. For each Gödelian set T, by G1 and its corollary, we have a true sentence $g_T \notin T$. Let $S_G = \{g_T: T$ is Gödelian$\}$. Clearly no superset of S_G is Gödelian. Otherwise put:

Fact: A consistent superset of S_G that contains arithmetic is not c.e.

For the anti-mechanist it therefore suffices to find a cognitive phenomenon A such that:

(a) S_A contains arithmetic,
(b) S_A is consistent,
(c) $S_G \subseteq S_A$.

Condition (a) points to an immediate suspect. On the model of Chomsky's approach to language, we posit a human cognitive faculty C, which accounts for our arithmetical competence—our ability to recognize the truth of arithmetical sentences. Clearly, S_C contains arithmetic in the appropriate sense.

Why assume, as per condition (b), that S_C is consistent? On the face of it this seems false. After all, individuals often make mistakes and change their minds about arithmetical statements, resulting in inconsistencies in the accumulated set of accepted sentences. However, it is clear the arithmetical judgments that people actually make, the data set, do not fully reflect their cognitive arithmetical faculty, if such there is. Following the Chomskyan practice outlined in §2.1, we allow S_C to be an extrapolated and pruned extension of the set of arithmetical sentences that are actually asserted. First, the finite data set (the set of actually uttered arithmetical judgments) is extrapolated into an infinite set (this was already implicitly assumed for condition (a)). Second, it is cleaned up by pruning it of inconsistencies and, perhaps, of falsehoods generally. The Chomskyan procedure for arithmetical competence is therefore assumed to result in a sound, or at least consistent, set S_C.[10]

In order to show condition (c), that $S_G \subseteq S_C$, the anti-mechanist appeals to the corollary to G1, in which g_T is proven for arbitrary T. The reasoning here is that g_T is mathematically proven (though not, of course, in a formal system), which is to say recognized as true. Since C was characterized as the ability to recognize as true arithmetical sentences, we have $g_T \in S_C$, and since T was arbitrary, we have $S_G \subseteq S_C$.[11]

Since conditions (a, b, c) hold, by our Fact above it follows that S_C is not c.e. By our characterization of Metaphysical Mechanism (§2.1), it follows that mechanism is false. This is the Gödelian anti-mechanist argument. We now turn to its refutation.

The arguments for all three conditions contain serious problems. First, in proving that condition (c) holds, we assumed that all g_T's are proven true. For this we have relied on the corollary to G1. However, the corollary carries over the hypothesis of the theorem, namely that T is Gödelian, and in particular, consistent. The sentence g_T for a particular T is only proven by the corollary if T is consistent. But it is no claim of the anti-mechanist argument that we can see whether a given arithmetical theory is consistent or not. Nor is this a plausible premise to add to the argument. But without it, it is not the case that we have proved the g_T's, not even informally, and, therefore, it is not the case that $S_G \subseteq S_C$. What we have proven is the set of conditionals $S_G^* = \{\ulcorner\text{if } T \text{ is consistent, then } g_T\urcorner\}$, for T c.e. and containing arithmetic. But S_G^* is certainly c.e., and so are many of its supersets. It was, therefore, not shown that S_C is non-c.e.[12]

Though this objection seems conclusive, the other problems I shall mention are arguably more illuminating philosophically. The first problem is with the appeal to arithmetic in condition (a). The second is with the idealization performed in the appeal to competence in condition (b).

[10] Compare (Shapiro, 1998, p. 275).

[11] Compare this with the moves in (Lucas, 1961) and (Penrose, 1989).

[12] An early statement of this objection is in (Putnam, 1960). See also (Bowie, 1982) and (Krajewski, 2020).

2.4. Against Arithmetic

The first point to become aware of is that, strictly speaking, it is not arithmetical competence that is doing the work in the Gödelian argument. By a specifically arithmetical competence we mean, I assume, the kinds of specifically arithmetical reasoning that we perform in order to reach arithmetical conclusions. In formalized theories, "specifically arithmetical reasoning" means logical reasoning from arithmetical axioms. Establishing an arithmetical theorem by means of, say, a set-theoretic proof can hardly count as an exercise of our pure arithmetical ability, but seems clearly to go beyond it and rely on additional resources. From the other direction, it seems that our specifically arithmetical competence is all but idle in our knowledge of the truth of arithmetical sentences such as "$2 = 2 \vee 2 \neq 2$". It is, therefore, not just the character of the theorem proved that determines which cognitive competence is responsible for its knowledge, but also the character of the proof.

Assume, contrary to fact, that theorem G1 does allow us to prove, as a corollary, all members g_T of the set S_G. Formally, the Gödelian argument would then go through, since we would have proven all members of a non-c.e. set. However, it would be wrong to say that we proved them using our arithmetical competence. The reasoning that led us to accept the corollary to G1 has nothing to do with arithmetic. It is justified only by the equivalence of g_T with its own unprovability, which is a metamathematical, not an arithmetical, fact. Granted, g_T itself is, in principle, an arithmetical sentence; but its specific arithmetical content is completely abstracted from in the proof, and is anyway dependent on the coding function $c(x)$, from the precise content of which we have also abstracted. All we rely on in the proof of G1 and its corollary are the metamathematical properties of g_T. Thus, even if the Gödelian argument had been valid formally, it would not show that arithmetical competence is non-computational.

The reason that the Gödelian anti-mechanist appealed to arithmetical competence is that Gödel's theorem is ostensibly about arithmetic. Looked at more carefully, however, we see that the connection between G1 and arithmetic is not that straightforward. Technically, the role that arithmetic plays in G1 is the fact that arithmetical theories represent, in the proof-theoretic sense, all recursive relations between numbers.[13] Given the well-known connections between recursive relations on numbers and computable relations on strings, this means that arithmetical theories can represent computable relations between strings. This is the point at which G1, in its classic formulation, makes contact with the notion of computability. The diagonalization function and the provability predicate are, respectively, a computable and a c.e. relation on strings, whence our Lemmas 1 and 2 in the proof-sketch above.

[13] In what follows I'll be loose and say "represent recursive (computable) relations" as shorthand for "strongly/weakly represent recursive/r.e (computable/c.e.) relations".

Once we highlight this, however, it becomes clear that the coding function and the use of recursive relations are just a detour. Theories are sets of strings, and provability in a theory T (for us, the predicate "$x \in T$") is a property of strings. In order to state Lemma 1 in the above proof (repeated here for convenience), we appealed to the coding function $c(n)$ on one side of the equivalence, and to the predicate PRV on the other:

(3) $c(n) \in T \Leftrightarrow \ulcorner PRV(\overline{n})\urcorner \in T$

Neither the coding function, nor the possibility of mentioning predicates, are part of Gödel's formalized arithmetical object-theory. This is clear from the fact that the language of the object-theory doesn't, in the general case, contain reference to strings. Both the coding function and the term referring to the provability predicate are defined in Gödel's unformalized metalanguage. If we were to formalize the metalanguage, we would have to include strings in its domain, alongside numbers. However, once we can refer to strings, all reference to numbers can be dropped. The syntactic concepts, in particular the provability predicate, are originally defined in terms of strings.[14] The form of Lemma 1 is simpler when stated for theories T that contain string theory instead of number theory:

Lemma 1* (Reflection): *There is a unary formula $PRV^*(x) \in L_T$, such that for every sentence $s \in L_T$,*

$s \in T \Leftrightarrow \ulcorner PRV^*(\overline{s})\urcorner \in T.$[15]

Lemma 1* is a product of the fact that $PRV^*(x)$ is a c.e. relation, and that string theories proof-theoretically represent such relations. Since theories are understood as sets of strings, reference to strings is anyway unavoidable, unlike reference to numbers and, hence, Lemma 1* is a more basic statement of the fact expressed by the original Lemma 1 above. The references to coding and to recursive relations in Lemma 1 are just a detour through arithmetic that allows us to

[14] Gödel himself, in the original paper (Gödel, 1953), refers to the more or less string-theoretic syntactic definitions in Łukasiewicz and Tarski (Tarski, 1956). Łukasiewicz and Tarski define the syntactical notions using set-closure definitions, which are the explicit higher-order counterparts of recursive definitions. Being higher-order, they are not computable. One of Gödel's crucial contributions in (1953) was to show how sequences can be coded into single numbers, allowing him to give, in the case of the syntactic notions, explicit counterparts of recursive definitions without going higher-order. The detour through arithmetic was thus necessary for Gödel at the time, in order to be able to code sequences. However, string sequences too can be coded in terms of single strings, though this technique was probably not available to Gödel. See (Quine, 1936; 1944) for work in string theory (concerned with elementary, not computable, relations). See (Grzegorczyk, 2005) for a development of Gödel's results in a string-theoretic setting.

[15] $\ulcorner\overline{s}\urcorner$ is the name of the string s.

apply Lemma 1^* to a special case. In fact, G1 can be applied to any theory which represents c.e. relations, whether it deals with strings, numbers, sets or what have you. The corresponding form of the theorem is:

Theorem G1*: *If T is a consistent formalized theory that represents computable relations, then we can compute from T a sentence g_T such that g_T, $\ulcorner\neg g_T\urcorner \notin T$.*

The upshot is that arithmetic is not part of the essential subject matter of G1 at all. G1 is a metamathematical theorem about formal proof systems in general, if they capture computable relations.

This suggests that arithmetical competence is the wrong cognitive phenomenon to use in a Gödelian anti-mechanist argument. It is simply not the right setup for an application of G1. We might conjecture some other kind of competence, more in tune with the essential content of G1; but whereas a natural arithmetical competence is somehow easy and smooth to postulate, there is no obvious natural cognitive competence that corresponds to the subject matter of G1 in the way required. Should we say we have a natural metamathematical competence? Or some general epistemological faculty? It is not clear that any reasonable form of mechanism has to be committed to this.

The upshot of the foregoing is that, even if the formal objection of the previous section did not apply (*per impossibile*), still the argument would not achieve its purpose, since it does not show that our arithmetical competence is non-c.e. One comment before we move on. From the foregoing discussion one might get the impression that Gödel's theorem is only accidentally connected with arithmetic, and that it is only a historical accident that the theorem has been discovered through its application to arithmetic as a special case. Inasmuch as Gödel's theorem is ultimately about theories, not about numbers, and theories are made up of strings, this impression is correct. However, the connection between arithmetic and string-theories is not just a historical accident. As we know, both the theories and the structures of strings and of numbers are very similar, and any result about the one can be expressed in terms of the other.[16] Indeed, numbers have a philosophical advantage over strings in that they constitute a single natural domain, whereas when considering strings concretely we have to fix a particular alphabet arbitrarily. We can say that the domain of numbers distills the invariant element in string domains. That would be a sense in which G1 is about arithmetic after all. In any case, it is not immediately about arithmetic, and making the connection clear and explicit will require further work.

2.5. Against Competence

The second philosophical problem with the Gödelian argument is with the notion of competence. In particular, the idealization performed in generating the

[16] See (Corcoran, Frank, & Maloney, 1974; Svejdar, 2008) for more on this.

phenomenon S_C from the set of actually uttered arithmetical sentences (see condition (b) above) is not legitimate.

Recall, following Chomsky, we have allowed the yield of cognitive phenomena, our "competences", to be extrapolated and pruned from data sets. But not every extrapolation and pruning would work. We cannot, for example, idealize away from the phenomenon a pattern of usage found in the data just because it doesn't conform to the theory we are inclined to accept. Nor can we idealize into the phenomenon something that didn't exist in the data set, unless we can be convinced that it reflects something in our cognition, and that some factor related to performance blocks it from being manifested in the data set. For example, we cannot condition our idealization on some decidedly external factor, like the set *B* from §2.1.

With regard to the Gödelian argument, the question is whether the finite and inconsistent set of humanly asserted arithmetical sentences, the arithmetical data set, can legitimately be idealized into an infinite consistent set. The analogy with Chomskyan grammatical competence is misleading. Granted, Chomsky sometimes speaks of grammatical competence in terms of (tacit) grammatical knowledge, and in our case too we speak of arithmetical knowledge. But there is an important sense in which grammar doesn't behave like knowledge at all. Knowledge, as usually understood, describes a belief that could have been false, and happens to be true. For example, if I know that $2 + 2 = 4$, this implies that I could have falsely believed that $2 + 2 > 4$. And my belief counts as knowledge partly because, in point of fact, $2 + 2 = 4$ is the case. Nothing corresponds to this in grammatical competence. There is no external domain of independent facts to which grammatical knowledge needs to correspond. No one is ever mistaken with respect to their tacit beliefs about grammar. It makes no sense to say that the sentence:

(4) John is very, very tall,

is both ungrammatical (for someone) and generated by that person's grammar. There is no external norm against which "knowledge" of language can be assessed.

Compare this now to the case of arithmetic, and to the idealization of arithmetical competence. The kind of knowledge that arithmetical competence is supposed to furnish its bearer with is genuine knowledge, one that refers to an independent reality. Unlike in the case of tacit grammatical knowledge, there is a sense in which we can say that we could have been wrong, that our arithmetical competence could have yielded $2 + 2 > 4$, in contradiction to actual fact. In arithmetic, there is, unlike in grammar, an external standard to which knowledge is compared.

This, I submit, makes the idealization of linguistic competence, relied on in the Gödelian argument above, illegitimate. If, in constructing a phenomenon, we base ourselves on something we know is external to the mind, like the set *B* of

§2.1, then the result cannot be counted a natural human cognitive competence, and mechanism about the mind makes no pronouncement about it. The mind might be thoroughly computational in the sense that the belief system of an individual is c.e., and yet the set of beliefs which constitute knowledge depends on factors external to the computational description.

This problem holds especially clearly if the assumption is that the arithmetical faculty is sound (not just consistent), and arguably this is the assumption that the anti-mechanist needs. But even if we only assume consistency, there is still here a reliance on an external standard, this time the truth of logical sentences. The issue is simply transferred to the question of logical competence, and here it is again soundness that is at stake.

To sum up, the Gödelian argument is mistaken in its assumption that we can treat arithmetical competence as having a consistent yield. This was one of the main premises of the argument, and so the argument fails.

2.6. The Tarskian Argument

In this section we have reviewed the Gödelian anti-mechanist argument and found, not only that it contains a formal fallacy, but also that its basic premise, the juxtaposition of human arithmetical competence with formalized systems, is deeply misguided. To conclude the paper, let me briefly sketch an anti-mechanist argument that I think has better prospects. This argument is based on Tarski's indefinability theorem, so it is also a kind of diagonal argument. Let's review, first, the theorem and its proof. Say that a language M, expressing the truth predicate for a language L, is L's *metalanguage*. L is then the object-language. A language is *semantically closed* if it is its own metalanguage, $L = M$.

Theorem T1: *No language is semantically closed.*

Premise 1 (Reflection, Convention T): If $TRUE(x)$ is a truth predicate for L in a metalanguage M, then for every $s \in S_L$, the following sentence holds:

$$\ulcorner TRUE(\bar{s}) \leftrightarrow tr(s)\urcorner,$$

where $\bar{s}$ is the M name of s, and $tr(s)$ is the M translation of s.

Premise 2: If $L = M$, then there is a sentence $k \in S_L$, such that the following sentence holds:

$$\ulcorner TRUE(\bar{k}) \leftrightarrow \neg k\urcorner.$$

From the two premises (and the fact that $\ulcorner k \leftrightarrow tr(k)\urcorner$ when $L = M$), we the following for the case that $L = M$:

(5) $\ulcorner k \leftrightarrow \neg k \urcorner$.

By reductio, the theorem follows.

The Tarskian anti-mechanist argument has the following premises. First, though Tarski's own concern was with formalized languages, today, mainly following Davidson, one often uses the general structure of Tarski's theories in the other direction, i.e. assuming truth to be understood and understanding the truth-conditional statements as providing a semantics for the language under consideration. We apply this approach to the semantics of the language of thought *L* (Fodor & Pylyshyn, 2014). The second premise of the argument is that the language of thought *L* can express any scientific theory. This is forthcoming if we accept that scientists cognize the theories that they put forth, and, therefore, that their language of thought should be able to express them. The third and final premise is that a full mechanistic theory of the mind needs to contain a semantic theory for *L*. For if it doesn't, then it can't with justice be seen as giving the content of the mental states of human subjects, and the characteristic property of the mind is its ability to entertain contents. However, together the three assumptions make *L* semantically closed, and this is impossible by theorem T1. The consequence is that no full mechanistic theory of the mind is forthcoming.

REFERENCES

Anstey, P. (2000). Descartes' Cardiology and Its Reception in English Physiology. In J. Schuster, S. Gaukroger, J. Sutton (Eds.), *Descartes' Natural Philosophy* (pp. 420–444). London, New York: Routledge.

Ben-Yami, H. (2015). Descartes' Philosophical Revolution: A Reassessment. London: Palgrave-Macmillan.

Bogen, J., Woodward, J. (1988). Saving the Phenomena. *Philosophical Review*, *97*(3), 303–352.

Bowie, G. L. (1982). Lucas' Number Is Finally Up. *Journal of Philosophical Logic*, *11*(3), 279–285.

Cantor, G. (1890). Ueber Eine Elementare Frage Der Mannigfaltigketislehre. *Jahresbericht Der Deutschen Mathematiker-Vereinigung*, *1*, 72–78.

Chomsky, N. (1957). *Syntactic Structures*. Berlin: Walter de Gruyter.

Chomsky, N. (1975). *The Logical Structure of Linguistic Theory*. Berlin: Springer Science+Business Media.

Cohen, H. F. (1984). *Quantifying Music: The Science of Music at the First Stage of Scientific Revolution 1580–1650*. Berlin: Springer Science & Business Media.

Corcoran, J., Frank, W., Maloney, M. (1974). String Theory. *Journal of Symbolic Logic*, *39*(4), 625–637.

Cottingham, J. (1978). 'A Brute to the Brutes?': Descartes' Treatment of Animals: Discussion. *Philosophy*, *53*(206), 551–559.

Descartes, R. (1996). *René Descartes: Meditations on First Philosophy: With Selections From the Objections and Replies* (2nd ed.). Cambridge University Press.

Descartes, R. (2006). *A Discourse on the Method of Correctly Conducting One's Reason and Seeking Truth in the Sciences*. Oxford University Press.

Duncan, W. D. (2017). Ontological Distinctions between Hardware and Software. *Applied Ontology*, *12*(1), 5–32.

Fodor, J. A., Pylyshyn, Z. W. (2014). *Minds Without Meanings: An Essay on the Content of Concepts*. MIT Press.

Franks, J. (2010). Cantor's Other Proofs That R Is Uncountable. *Mathematics Magazine*, *83*(4), 283–289.

Fresco, N. (2014). *Physical Computation and Cognitive Science*. Springer.

Funkenstein, A. (1986). *Theology and the Scientific Imagination From the Middle Ages to the Seventeenth Century*. Princeton University Press.

Gardner, H. (1987). *The Mind's New Science: A History of the Cognitive Revolution*. New York: Basic Books.

Gaukroger, S. (2002). *Descartes' System of Natural Philosophy*. Cambridge University Press.

Gaukroger, S. (2007). *The Emergence of a Scientific Culture: Science and the Shaping of Modernity 1210–1685*. Oxford University Press UK.

Gaukroger, S. (2010). *The Collapse of Mechanism and the Rise of Sensibility: Science and the Shaping of Modernity, 1680–1760*. Oxford University Press.

Gaukroger, S., Schuster, J., Sutton, J. (2000). *Descartes' Natural Philosophy*. London, New York: Routledge.

Gödel, K. (1953). *Kurt Gödel: Collected Works*. Oxford University Press.

Gödel, K. (1986). *Kurt Gödel: Collected Works: Volume III: Unpublished Essays and Lectures*. Oxford University Press.

Grodzinsky, Y., Santi, A. (2008). The Battle for Broca's Region. *Trends in Cognitive Sciences*, *12*(12), 474–480.

Gunderson, K. (1964). Descartes, La Mettrie, Language, and Machines. *Philosophy*, *39*(149), 193–222.

Hauser, M. D., Chomsky, N., Tecumseh Fitch, W. (2002). The Faculty of Language: What Is It, Who Has It, and How Did It Evolve? *Science*, *298*(5598), 1569–1579.

Lucas, J. R. (1961). Minds, Machines and Gödel. *Philosophy*, *36*(137), 112–127.

McLaughlin, P. (2000). Force, Determination and Impact. In J. Schuster, S. Gaukroger, J. Sutton (Eds.), *Descartes' Natural Philosophy* (pp. 81–112). London, New York: Routledge.

Miłkowski, M. (2013). *Explaining the Computational Mind*. MIT Press.

Papayannopoulos, P. (2020). Computing and Modelling: Analog vs. Analogue. *Studies in History and Philosophy of Science Part A*.

Penrose, R. (1989). *The Emperor's New Mind*. Oxford University Press.

Penrose, R. (1994). *Shadows of the Mind*. Oxford University Press.

Pullum, G. K., Scholz, B. C. (2010). Recursion and the Infinitude Claim. *Recursion in Human Language*, *104*, 113–38.

Putnam, H. (1960). Minds and Machines. In S. Hook (Ed.), *Dimensions of Minds* (pp. 138–164). New York University Press.

Putnam, H. (1963). Degree of Confirmation' and Inductive Logic. In P. A. Schilpp (Ed.), *The Philosophy of Rudolf Carnap* (pp. 761–783). La Salle: Open Court.

Quine, W. V. (1940). *Mathematical Logic*. Harvard University Press.

Sepper, D. L. 2000. Figuring Things Out: Figurate Problem-Solving in the Early Descartes. In J. Schuster, S. Gaukroger, J. Sutton (Eds.), *Descartes' Natural Philosophy* (pp. 228–248). London, New York: Routledge.

Shapiro, S. (1998). Incompleteness, Mechanism, and Optimism. *Bulletin of Symbolic Logic*, *4*(3), 273–302.

Sieg, W. (2001). Calculations by Man and Machine: Conceptual Analysis. *Reflections on the Foundations of Mathematics (Essays in Honor of Solomon Feferman)*, *15*, 387–406.

Sieg, W. (2013). Gödel's Philosophical Challenge (to Turing). In B. J. Copeland, C. J. Posy, O. Shagrir (Eds.), *Computability: Gödel, Church, Turing, and Beyond* (pp. 183–202). MIT Press.

Slezak, P. (1983). Descartes's Diagonal Deduction. *The British Journal for the Philosophy of Science*, *34*(1), 13–36.

Slezak, P. (1988). Was Descartes a Liar? Diagonal Doubt Defended. *The British Journal for the Philosophy of Science*, *39*(3), 379–388.

Smith, B. C. (2002). The Foundations of Computing. In M. Scheutz (Ed.), *Computationalism: New Directions* (pp. 23–58). MIT Press.

Sorensen, R. A. (1986). Was Descartes's Cogito a Diagonal Deduction? *The British Journal for the Philosophy of Science*, *37*(3), 346–351.

Svejdar, V. (2008). Relatives of Robinson Arithmetic. In M. Peliš (Ed.), *The Logica Yearbook* (pp. 253–263). London: College Publications.

Tarski, A. (1956). *Logic, Semantics, Metamathematics*. Oxford: Clarendon Press.

Woodward, J. F. (2011). Data and Phenomena: A Restatement and Defense. *Synthese*, *182*(1), 165–179.