Logical underdetermination in physics

By Arne Hole, January 10th 2017

Abstract

This paper discusses physical implications of a result concerning underdetermination of truth in mathematics. If the time development of a complex physical system on the macro level for logical reasons may be underdetermined by the physical laws which determine its behavior on the micro level, then there arises a free maneuvering space concerning that physical system, giving room for a form of "free will". This logical mechanism also offers a possible explanation of the frequently discussed contrast between the second law of thermodynamics, implying increase in total entropy, and the fact that the universe is developing structure and structured organisms.

1. Introduction

Traditionally, one assumes that the principle of induction on size or complexity is applicable when it comes to defining or deducing properties of mathematical or physical objects. In mathematics, the principle of induction is, in one form or another, an important ingredient in every axiomatization of number theory, that is, the study of the set N of natural numbers 0, 1, 2, 3, ... . This principle expresses the idea that if a property can be shown to hold for some natural number n, and it can also be shown that if the property holds for a certain number, then it also holds for the next, then the property will hold for all natural numbers.

There is, however, a snag which may go unnoticed here: When we use the term "show", we refer to a proof of finite length. However, as is well known, the concept of finiteness cannot be formalized. See any textbook on mathematical logic, for example Shoenfield [1967]. This implies that we cannot assume from the concept of "proof" a degree of absoluteness which is not shared by mathematical structures such as N. The set of admissible proofs is infinite, and as such it admits non-standard models in every formalized description, just like N itself does.

Thus when we speak about the concept of truth in the mathematical structure N of natural numbers, the set N plays a double role. On one hand, it is a set which we assume determines the truth of each formula in languages such as the language L(PA) of Peano Arithmetic (PA). On the other hand, we may prove by induction that proofs in PA preserves truth in N, that is, that if the premises of the proof are all true in N, then the conclusion of the proof is also true in N. But the admissible length of the proof is again determined by the set N, through the non-formalizable concept of finiteness.

Hence, the definition of truth in N has a kind of circularity in it. This has not been perceived as a problem traditionally, since it has not been shown to lead to inconsistencies of any kind. Indeed, no such inconsistencies will be claimed to exist in this paper either. But while inconsistencies correspond to overdetermination, we will investigate the opposite, namely underdetermination. The circularity mentioned above raises the question of whether the concept of truth in N is completely determined by the inductive definition of truth commonly used, a definition relating to Tarski [1956]. There is a theoretical opening for the possibility that the concept of truth in N might be underdetermined. Up until now, there has not been any evidence of such a thing. However, the main result in Hole [2017]
indicates that an underdetermination effect of this kind does indeed arise naturally for certain complex formulas. In this paper, I will describe the logical result from Hole [2017] and discuss the implications it has for physical theories.

2. The logical result: Underdetermination of truth in number theory

The result in Hole [2017] is concerned with formal theories in which results about the set N of natural numbers may be inferred. The classical example of such a theory is Peano Arithmetic (PA). This theory is a so-called first order logical theory, and it has axioms which are translations into the language L(PA) of a set of formulas generally considered as evidently true when interpreted in N. The set N corresponds to what is called the standard model of PA (Shoenfield [1967]). The class of theories considered in Hole [2017] is labeled extensions of PA in that paper. Roughly speaking, an extension of PA is a theory T with a language L(T) containing the language L(PA) of PA, and with the property that all theorems of PA are also theorems of T. In Hole [2017], all extensions of PA by definition are assumed to be axiomatized and to have the property that they do not prove their own inconsistency. ‘Axiomatized’ means that there exists an algorithm which can distinguish an axiom of T from formulas of L(T) which are not axioms of T. Hence we do not include, say, theories where we take all N-true formulas of L(PA) as axioms. An extension T of PA is called a reachable theory if we know in advance that all the axioms of T are true when interpreted in N. The class of reachable theories then corresponds to theories we may actually use for inferring new knowledge about N.

Now by the well-known incompleteness theorems of Kurt Gödel (see Shoenfield [1967]), it follows that if T is an extension of PA which does not prove its own inconsistency, there exists a sentence B in L(PA) such that (i) T does not prove B, and (ii) T does not prove that consistency of T implies nonprovability of B in T. In Hole [2017], the Gödel number of a certain sentence of this kind is singled out and labeled the naturally inaccessible number (NIN) of T. On the basis of this NIN, for each natural number \( n > 0 \), in Hole [2017] there is defined a length \( n \) contribution sequence of T. This sequence consists of \( n \) binary bits taken from chosen, scattered bit positions close to the middle of a binary number obtained by performing some fixed arithmetical operations on the NIN. The idea is that "substantial" knowledge about the NIN should be required for obtaining nontrivial information concerning the bits in the contribution sequences of T. On the other hand, due to the requirements (i) and (ii) above, the theory T itself cannot even determine a possible candidate for its own NIN. The next step in Hole [2017] is to combine certain parts of the contribution sequences of all extensions of PA into an infinite bit sequence \( \tau \). The length \( n \) of the contribution sequence chosen for each extension T is determined by its place in an ordering of a certain PA-decidable set of theories which contains all extensions of PA. Then an initial segment of that contribution sequence of length equal to the place of T in an ordering of all extensions of PA, is included in \( \tau \). The infinite sequence \( \tau \) consists of 1 bit taken from the chosen contribution sequence of the first PA extension, followed by two bits from the contribution sequence of the second extension, followed by three bits from the contribution sequence of the third, and so on. We call a natural number \( n \) a winner with respect to \( \tau \) if \( n = 0 \) or all the \( n \) bits from the contribution sequence of the \( nth \) PA extension in \( \tau \) are 1.

More generally, if \( s \) is any infinite bit sequence, divide the bits of \( s \) into groups such that the first group consists of the first bit in \( s \), the second group consists of the following two bits, the third group
contains the next three bits, and so on. We call an integer $n$ a winner with respect to $s$ if $n = 0$ or the $n$th group of bits in $s$ consists of bits 1 only.

The final concept we need for describing the central result of Hole [2017], is the concept of a probability $p$ property of infinite bit sequences. The set of all infinite bit sequences is often called Cantor space. In Cantor space there is a probability measure which is often referred to as the fair coin measure. This measure so defined that if $s$ is a bit sequence of finite length $L$, then the probability measure of the set of all infinite sequences which begin with $s$, is $(0.5)^L$. If $P$ is a property of infinite bit sequences such that the set of all infinite bit sequences with the property $P$ has probability measure $p$, then $P$ is called a probability $p$ property of infinite bit sequences.

The main technical result of Hole [2017] may now be stated as follows:

**Theorem (Proof: Hole [2017])**

There is a sentence $D$ in $L(PA)$ which is true in $N$, with the following property: If $T$ is an extension of PA proving $D$, then either

(i) $T$ proves that there are an infinite number of winners with respect to the sequence $\tau$, or

(ii) There is a natural number $n$ such that $T$ proves the existence of a zero in its own $n$-length contribution sequence.

Having an infinite number of winners is a probability zero property of infinite bit sequences.

This result may be generalized in a number of directions, but from our physical perspective that is not essential.

It is argued in Hole [2016] that on basis of the theorem above, one may conclude that the formula $D$ cannot be neither proved nor disproved in any reachable extension of PA. This latter argument is a form of reduction ad absurdum, but corresponding to the fact that we are dealing with underdetermination instead of overdetermination, the absurd situations involved are not contradictions. Instead, they represent "infinitely unlikely coincidences" or "magical" results. As a result of this, the undecidability argument is not formalizable in, say, first order logical theories. But still, it is arguably equally convincing as a standard mathematical proof establishing a logical consequence.

To briefly go through the undecidability argument, note first that since $D$ is true in $N$, the formula $D$ cannot be disproved in a reachable theory. Further, the argument is that given the way the contribution sequences of $T$ are defined on basis of its NIN, a reachable theory has no possibility of proving a property such as the existence of a zero in concrete contribution sequence based on its own NIN. For any given $n$, the probability that a randomly chosen length $n$ bit sequence contains a zero, is strictly less than 1. Hence, some nontrivial information concerning the length $n$ contribution sequence of $T$ must be derivable in the theory $T$ itself. On the other hand, as was mentioned above, the theory $T$ cannot even produce a candidate for its own NIN, on which the contribution sequence in question is based.

A similar argument can be used for the impossibility of condition (i) in the theorem in the case where $T$ is reachable. Condition (i) says that $T$ can prove the existence of infinitely many winners with respect to the sequence $\tau$, a sequence which combines information about the NINs across all extensions of PA. Just like assuming that we can find a reachable theory satisfying (ii), this is also
arguably absurd. One can easily verify that for any given natural number $N$, the probability that there are no winners $n > N$ with respect to a randomly chosen infinite bit sequence, is nonzero and increasing with $N$. It follows from the law of large numbers that the property of having an infinite number of winners is a zero probability property of infinite bit sequences. Also this property is unrelated to the definition of $\tau$. Given the way $\tau$ is defined, combining information about the NINs of all PA extensions, it would be absurd if one particular reachable theory could prove this property of $\tau$.

In total, the argument shows that unless we accept that $D$ cannot be proved in any reachable theory, we are forced into making one of these two absurd assumptions.

Since the concept of "reachable extension of PA" includes theories with arbitrary languages extending the language of PA, the argument outlined above shows that the truth value of $D$ is may naturally be called absolutely undecidable. It is in principle impossible to settle the truth value of $D$, even when disregarding limitations in space and time. To avoid confusion at this point, it is important to keep in mind the distinction between direct and indirect reference to formulas. We definitely know the truth value of $D$ in $\mathbb{N}$ when it is referred to indirectly, as in the theorem. But this does not help us decide whether the concrete formula $D$ is true or false, if we are confronted with it. Tracing back definitions given in Hole [2016], we can almost construct $D$ explicitly: We can narrow the set of formulas containing $D$ down to a set of simply related formulas, but then we can go no further.

The way the theorem above is formulated, it assumes that $D$ actually has a truth value in $\mathbb{N}$. But if the truth value can never be determined, not even in principle, then assuming the existence of the truth value is arguably unnatural. What does it mean for the truth value of $D$ in $\mathbb{N}$ to "exist" if it is not observable or inferable? Certainly a theoretical opening for the idea that the truth value of $D$ in $\mathbb{N}$ simply does not exist, is created. However, we do not need to pursue this question further. After all, it boils down to discussing the meaning of "existence" in our context. What is important to us, is simply that the truth value of $D$ is nonexistent in the sense that it cannot be determined, not even in principle. This means that the truth value of $D$ is, in this precise sense, underdetermined by the definition of truth in the structure $\mathbb{N}$ of natural numbers.

### 3. Physical systems with underdetermined dynamics

The implications of the possible existence of the logical underdetermination effect described in the previous section, goes beyond mathematics itself. The structure $\mathbb{N}$ of natural numbers has universal properties which allow mappings of both the structure itself and the associated proof theory into classes of physical systems and questions concerning these systems.

Let us start by looking at how the truth value of $L(\text{PA})$ formulas in $\mathbb{N}$ is traditionally defined. The usual way of doing this starts with defining truth values for atomic formulas. These are simple formulas stating relations between terms. The truth value of more complex formulas are then defined inductively on the complexity of the formula, so that $A \land B$ is true iff $A$ and $B$ are both true, $\exists x A$ is true iff there is a natural number $n$ such that $A$ is true when a numerical term representing $n$ is substituted for the variable $x$ in $A$, and so on. Clearly, one would assume, this procedure fixes a truth value for all formulas of $L(\text{PA})$, since all such formulas may be obtained from atomic formulas by a finite number of such steps. But still, the truth value of $D$ is arguably left underdetermined. So what is happening? There is no escape from the conclusion that the complexity involved in $D$ is of such a
size and kind that the inductive truth definition actually breaks down, despite the fact that it looks perfectly sound when looked at "locally" in the inductive process.

There is an obvious physical analogy to this, namely the relation between the dynamics of physical systems on the micro level and the dynamics of physical objects on the macro level. Suppose that we live, as is generally assumed, in a universe where the dynamics of systems on the micro (Planck) scale is governed by certain physical laws. The fact that these laws may be of a probabilistic nature, does not matter to us here. We do not need to enter discussions of "what" determines quantum events, and so on. The point is simply the basic assumption that on some micro level, the dynamics of the universe is determined by physical laws. Does this mean that these same laws will also determine the dynamics of objects on the macro scale, such as human brains? Yes, one would say. The line of reasoning behind that answer would again be an inductive argument, an induction on the complexity of the physical system in question. But, and here comes the main point of this paper: Given the fact that a similar inductive process in the case of pure mathematical truth values in a sense breaks down, there suddenly arises a very good theoretical reason to doubt the validity of this inductive reasoning in the physical case. The potential logical complexity of physical systems on the scale of, say, human brains, is (at least) comparable to the logical complexity of the formula D of the previous section. Hence, it is certainly possible, in fact I would say completely natural, that an underdetermination effect similar to the mathematical one may exist for such physical systems, for instance human nervous systems. On the input side, the dynamics of a human brain is linked to the physical world in which it exists through its senses. These senses are capable of representing (transmitting) information with complexity on at least the same level as the sequence of theories "touched" by the sentence D. Therefore, in light of the logical result, there is actually no reason to believe that the physical laws determining dynamics on the micro level, will also determine the dynamics of all physical systems on the macro scale. Theoretically, it may be the case that the dynamics of some macro systems for logical reasons is underdetermined by the physical laws governing the micro level. Such underdetermination will then open a maneuvering space which can, for instance, allow for "free will" in human brains. In fact, the natural way to view this would be as follows: When a physical system with the right kind and degree of complexity is formed, a maneuvering space concerning the dynamics of the system is naturally created by logical underdetermination. Then if one assumes that for physical reasons, decisions must be made by "someone" about what is actually going to happen with the system, "free will" arises as a physical necessity. Thus it would not be the case that "free will" arises and then elbows its own physical maneuvering space, it is the other way round: The physical maneuvering space arises for logical reasons, and "free will" is then created as a consequence, to fill out the maneuvering space.

The main point is that the mere theoretical possibility of logical underdetermination, as described above, shows that a rational explanation of the phenomenon of "free will" is logically possible in a world which is governed by physical laws on a micro level.

It should be remarked that logically, we clearly need not consider the "simple" level to correspond to Planck scale. We could also choose the simple level to correspond to macroscopic objects, and to let our physical theory be classical mechanics. While systems of reasonable complexity will follow Newton’s laws of motion (approximately), extremely complex mechanical systems could again display underdetermination and resulting free physical maneuvering spaces.
4. Entropy and the second law of thermodynamics

As we have discussed, the logical result of section 2 gives us a theoretical reason to doubt the universal applicability of physical laws in situations involving high complexity. It is clear that statistical mechanics is a field of interest in this context, due to the potential $10^{23}$ complexity factor involved when linking this theory to thermodynamics and other macroscopic theories. Entropy may be defined information theoretically, and there are no logical problems with its definition on the micro level, for simple systems. However, when the well-known conclusion concerning increasing total entropy in the universe is proved in the traditional way, one uses thermodynamic postulates concerning macroscopically defined thermodynamic coordinates or physical assumptions equivalent to such postulates. Such thermodynamic postulates are based on empirical evidence, but of course all experiments conducted have been done in situations with a logical complexity far below the levels in which underdetermination of the kind we are discussing here, can occur. Although the potential logical complexity associated with macro scale systems is large when considering the jump from Planck scale, many macro systems are of course structurally much simpler. The logical result of section 2 gives a theoretical reason to doubt the validity of these postulates for some systems of sufficiently high complexity. Of course, an obvious candidate for such a sufficiently complex system would again be the human nervous system.

But, one may then rightfully ask, if there really may occur processes violating one of the commonly accepted thermodynamic postulates, so that there may exist, for instance, irreversible processes in which the total change in entropy is not positive, why has nothing of this sort been observed in connection with the actions of human beings? A first answer would be that in order to actually observe such a thing, one would probably need to tackle the complexity of the processes within the human body on a micro level. It is fair to say that nothing remotely similar to that has ever been done. More importantly, there are good reasons to believe that actually observing a case where one of the thermodynamic postulates breaks down in this way, is impossible even in principle. The reason is simply that we are dealing with underdetermination, not overdetermination. This may naturally mean that the thermodynamics of the real world is consistent with standard thermodynamics theory and statistical physics when it comes to observations. However, given that there still are free physical maneuvering spaces, we might get the impression that in the long run, things do not develop as we would expect when taking into account, say, the principle of increasing total entropy. On the other hand, we would not be able to single out a concrete process which violates the physical laws in question, the reason being that the physical time development of the physical system we consider, is underdetermined by these laws, and therefore compatible with the laws when it comes to observations which can actually be performed. This physical situation would form a perfect analogy to the fact that the absolute undecidability argument for the sentence D of section 2 cannot be formalized.

Reasoning this way we obtain, in particular, a possible theoretical explanation of the fact that physical processes associated with living organisms and other complex systems in the long run does not seem to really confirm with the principle of increasing total entropy. Contrary to this principle, the universe has developed complex, apparently low-entropy structures. Rejecting the general
validity of the increasing entropy principle also means that theories in e.g. cosmology where this principle is used as a premise, may be considered irrelevant.

REFERENCES

