The paradox of predictability

Gijsbers, V.A.

Erkenntnis
https://doi.org/10.1007/s10670-020-00369-3

ORIGINAL RESEARCH

The Paradox of Predictability

Victor Gijsbers[1]

Received: 20 March 2019 / Accepted: 22 December 2020

Abstract
Scriven's paradox of predictability arises from the combination of two ideas: first, that everything in a deterministic universe is, in principle, predictable; second, that it is possible to create a system that falsifies any prediction that is made of it. Recently, the paradox has been used by Rummens and Cuypers to argue that there is a fundamental difference between embedded and external predictors; and by Ismael to argue against a governing conception of laws. The present paper defends a new diagnosis of the roots of the paradox. First, it is argued that the unpredictability has to be understood in the light of Turing's famous results about computability, in particular his proof that there is no solution to the 'halting problem.' This allows us to see that previous analyses of the paradox were either mistaken or not fully adequate. Second, the sense of paradox that nevertheless remains is traced to the idea that rational behaviour is not dependent on contingent environmental circumstances: that it is always *up to us* to engage in activities such as rational prediction or rational belief. The paradox of predictability teaches us that this idea, natural though it may be, is mistaken.

1 Introduction

In his 1965 paper "An Essential Unpredictability in Human Behaviour", Michael Scriven argued that it is possible to create deterministic systems whose behaviour cannot be predicted, not even in principle. Roughly, his idea is that we can create a system that is either informed of or repeats the prediction that is made about what it itself will do, and that then performs a different action than the one predicted. You predict that I will go left, and therefore I go right, invalidating your prediction. In the years immediately following his paper, a debate sprang up about free will and speed of computation—with, for instance, Lewis and Richardson (1966) arguing that Scriven's scenario makes contradictory assumptions about computation speed, and Good (1971) claiming that free will in the strong sense can only be had if "you

✉ Victor Gijsbers
V.Gijsbers@hum.leidenuniv.nl

[1] Institute for Philosophy, Leiden University, Leiden, The Netherlands

Published online: 19 March 2021

can think or compute fast enough" to outperform potential predictors. This literature was then all but forgotten.

More recently, however, the so-called 'paradox of predictability' has been revived, in particular by the excellent analysis of Rummens and Cuypers (2010), but also by Holton (2013) and Ismael (2016). All these authors believe that the paradox shows that determinism *does not* imply predictability. But how do we square this with our intuitive idea that it *does*? Rummens and Cuypers answer this by making a distinction between *embedded* and *external* predictors, and arguing that determinism only implies predictability for external predictors. Ismael, on the other hand, claims that determinism only implies predictability if we accept a governing conception of laws of nature. She concludes that Scriven's scenario shows that we must abandon this conception.

In the present article, I argue that the sources of the paradox have not been correctly diagnosed by these authors; and I will propose what I take to be the correct diagnosis. The impossibility of predicting deterministic systems follows from two things: (1) *formal* considerations about prediction—roughly, Alan Turing's proof that the halting problem is undecidable—and (2) the *embedability* of the process of prediction in the target world. I will show that Turing's impossibility proof is structurally identical to Scriven's scenario under the assumption of embedability, and I will use this to show that Ismael's attack on the governing conception of laws is misguided and that the embedded/external distinction of Rummens and Cuypers must be fundamentally reinterpreted.

I will go on to argue that the paradox of predictability fits in a larger class of philosophical thought experiments, all of which teach us the lesson of the *contingency of rationality*: whether we, even if we remain endowed with all the mental powers we have at the moment, will be able to engage in rational activities is dependent on contingent matters of circumstance. The seeming unnaturalness of this idea—and the deep anchorage of the opposite idea in the philosophical tradition—is, I will suggest, in no small measure responsible for the sense of paradox that Scriven's scenario creates. Throughout I hope to prove at least to some extent the thesis of Blanchard (2017) that "this understudied puzzle has much to teach us about freedom and determinism."

2 The Paradox

The paradox of predictability is generated from two ingredients: a deterministic universe and a counterpredictive device. Suppose that we live in a deterministic universe. This means that the initial state of the universe and the laws of nature together determine everything that will happen. In principle, then, everything that will happen can be predicted. (This phrase 'in principle' serves to cover up many difficulties, some of which are explored in Popper's (1950, 1982). Among other things, we ignore imprecisions of measurement and calculation, causal effects of measuring on the object of measurement, the impossibility of obtaining information about events outside of one's past light cone, and the question of whether it is possible for

a system to contain a detailed enough representation of the state of that system's universe.)

Let us imagine a *predictor*, a being or machine that has been provided with the initial conditions and the laws, and which we will assume to have arbitrarily mighty powers of computation. This predictor should be able to successfully predict everything that will happen. But now we also imagine a 'counterpredictive' device, the *counter*, which operates on a simple principle: it waits to hear what the predictor predicts that it, the counter, will do, and then it does the exact opposite. Evidently, the predictor cannot successfully predict what the counter will do. But the predictor can predict everything. Contradiction.

The idea of a predictor and a counter may sound far-fetched, but it is in fact rather mundane. If I claim to have deep insight into your psyche, deep enough that I can predict what you will do, you are likely to turn yourself into a counter, simply to prove your own independence. If I predict that you will order a glass of Merlot, you will order a gin tonic instead, thereby disproving my claim to be a good predictor of your behaviour. But we need not even be thinking about intelligent agents. We can ask a computer to predict whether a given device will emit a beep or not. Assuming that the device has been built in such a way that it emits a beep if and only if the computer predicts that it will not, then the computer cannot possibly make a correct prediction. Building a counter can be entirely trivial.

It is so trivial that the paradox may seem to disappear immediately. *Of course* the computer cannot successfully predict the behaviour of a machine that will always do what the computer does not predict. To restore the sense of paradox, we must keep the opposing intuition clearly in mind: if this universe is deterministic, then it simply *follows* from the laws and the initial conditions that the machine will emit a beep (or not). But then why is the predictor unable to predict this event? How can the counter, which does not even causally influence the predictor, prevent the predictor from arriving at the truth? Surely it cannot! Hence the paradox.

In the next section, we will canvas several important ways of responding to the paradox. But there is one objection that is so obvious that we need to discuss it immediately. As stated, the paradox depends on the prediction being *revealed* to the counter. If the predictor keeps its prediction a secret, the counter cannot ensure that it acts contrary to the prediction. If I write 'Merlot' on a piece of paper which you cannot see until after you've ordered your drink, you cannot ensure my defeat. But there is of course no contradiction between the claim that everything can be predicted and the claim that some predictions cannot be revealed; as Scriven writes, "[s]ecret predictions are still predictions" (1965, p. 414). So the paradox disappears.

There are three possible responses to this. First, one can embrace this as the correct way out of the paradox. This might seem to drain it of all its interest—it turned out to be nothing more than a trivial mistake of reasoning—but it can then still be used for one non-trivial purpose: defending the idea that even in a deterministic universe, we have what Rummens and Cuypers call 'take-it-or-leave-it' control. This would mean that "a human being who is confronted with a prediction about its own future behaviour might consider several other conditions and, depending on whether or not these hold, decide to go against or, alternatively, decide to act in accordance with the prediction made. (Rummens and Cuypers

2010, p. 247)" Ismael similarly concludes that we have "the ability to thumb our noses at anyone that thinks they can tell us what we will do" (2016, p. 182).

But this way of responding to the objection leaves much to be desired. Does the paradox really show that there can be take-it-or-leave-it control in a deterministic universe? Rather, it seems that we have simply *assumed* that such control is possible by assuming that a counter is possible. The argument just given comes down to:

1. We have take-it-or-leave-it control.
2. Possibly, we live in a deterministic universe.
3. Hence, determinism is compatible with take-it-or-leave-it control.

But nothing we have said so far shows that this argument is better than the incompatibilist argument that goes like this:

1. We have take-it-or-leave-it control.
2. Take-it-or-leave-it control is incompatible with determinism.
3. Hence, it cannot be the case that we live in a deterministic universe.

If one wants to get any clarity or insight from the paradox, one has to keep it intact for a little longer and delve more deeply into *why* we have opposing intuitions and *what* is going wrong. And let me stress that neither Rummens and Cuypers nor Ismael are satisfied with just giving the above argument—as we will see in the next section, they do attempt to explain the paradox at a deeper level.

The second response is to claim that the revelation condition does not truly diminish the paradoxical nature of the situation. Suppose that I have deduced your choice of drink from the initial state of the universe and the laws of nature. Then why can't I reveal it to you? If I *literally* can't reveal it to you—e.g., whenever I try to say it out loud I choke up, or the noise of a passing lorry drowns out the sound of my voice—then it seems that this bizarre set of coincidences cries out for explanation. Is the universe conspiring to keep my predictions secret? How could that be? If I *metaphorically* can't reveal it to you—when I reveal it to you, I thereby invalidate it—then my prediction turns out to be false, which is inconsistent with our assumptions. In either case, there remains a paradox.

The third response is to claim that the revelation of the prediction is *not* an essential aspect of the paradox. For assume that the predictor uses data D and an algorithm L to derive its prediction. Then we can build a counter that *also* applies algorithm L to data D, thus deriving the same prediction as the predictor, and then doing the opposite of whatever this duplicate prediction claims it will do. This effectively circumvents the need for revelation, though at the price of introducing a second arbitrarily powerful predictor—a price that, according to Lewis and Richardson (1966), cannot be paid, since each of the two predictors is thereby supposed to be more powerful than the other. Rummens and Cuypers (2010, pp. 238–239) present several ways of avoiding the Lewis and Richardson result. But the easiest solution is perhaps simply to remove the predictor as an

independent entity and *identify* it with the predicting part of the counter. Nothing of the paradox gets lost if we see the counter simply as falsifying its own 'internal' prediction. Why can't it predict itself, given its knowledge of the laws and the initial conditions? This remains unclear.

I conclude that we need not worry too much about the revelation or secrecy of the prediction. Let us now look at ways of dissolving the paradox.

3 Proposed Solutions

The contradiction arises when we claim that every event in a deterministic universe is in principle predictable *and* that such a deterministic universe may contain a counter for any predictor. There are thus three high-level strategies for dissolving the paradox, defined by adopting one of the following theses:

1. Not every event in a deterministic universe is predictable, even in principle.
2. It is not possible to build effective counters in a deterministic universe.
3. There are failures of either prediction or countering on a case-by-case basis.

The third strategy may sound strange, but it is in effect adopted by Blanchard (2017) in a review of Ismael's book:

> Consistency can be retained by maintaining that anytime the envisioned situation occurs, either the predictor somehow fails to derive or reveal the correct prediction, or the counterpredictive device somehow malfunctions. […] Arguably, there is nothing more mysterious here than the fact that time travelers slip on banana peels every time they try to kill their younger selves. (Blanchard 2017, p. 163.)

Blanchard goes on to suggest that what counterprediction shares with autoinfanticide is that in setting up a scenario, we build into it information about the future that makes it "epistemically certain that the *present* must unfold in a certain way" (p. 163). Perhaps this defence works for the case of time-travel. But do we really build information about the future into the counterprediction scenario? Only to the extent that we postulate how the counter will respond to specific inputs. But that is just the kind of if–then hypothetical information we use for *every* prediction, including the most mundane ones; so it can't be *this* that is generating the sense of paradox. The paradoxical element of time travel scenarios is precisely that they postulate as certain some non-hypothetical claims about the future; and that is not happening here.[1]

[1] Ismael (2019) argues that in a Minkowski space–time, the kind of information needed for prediction—information about the total state of the world, or at least about a total time slice of the past light cone of the event that is to be predicted—should indeed be considered *information from the future*. I find her argument fascinating, but it would take us too far afield here to consider her interpretation of relativity theory. Since we are discussing the relation between determinism and predictability in general, one may, if one wishes, assume that we are from now on speaking about Newtonian worlds.

The second strategy also fails, but dealing with it requires us to be somewhat more clear about the exact reach of our claims. Suppose that there is a universe *U* in which there are no counters. Then of course the paradox of predictability does not imply that there are events in this universe that cannot be predicted. Are we allowed to say that even though there are no counters in *U*, there *could have been* counters in *U*? This would certainly involve us in troublesome questions about modality. For is it not the case that any world has all its properties essentially? And if we move to another possible world, have we not moved to a world that the prediction was not supposed to be about?

However, the paradox requires only a far weaker claim: that counters are *possible*, that is, that there are *some* deterministic universes in which there are counters. When we consider predictors that exist within the world-to-be-predicted, it seems quite clear that there is no general problem with building counters. Building a counter to a predictor is merely a matter of attaching a device that does the opposite from what it is predicted to do. This is as simple as writing a computer program that displays a 0 just when it is sent a 1 and a 1 when it is sent a 0; or building a logical not-gate from electrical circuitry. It is hard to see how attempts to build such devices could fail in general. (We will consider the question of predictors *outside* the universe in a moment, when we come to the proposals of Rummens and Cuypers.) And this closes off the second way of dealing with the paradox.

So that leaves strategy one: the lesson of the paradox is that *not* every event in a deterministic universe is predictable, even in principle. This is something that most writers on the paradox agree on, including Scriven (1965), Rummens and Cuypers (2010), Holton (2013) and Ismael (2016, 2019). But of course we need to say more in order to dispel the paradox. We need to explain *why* or *in what sense* determinism fails to deliver predictability. The most substantive proposals in this direction come from Rummens and Cuypers, who hold that predictability fails for physical systems that are part of the universe they attempt to predict but not for 'external' predictors; and from Ismael, who holds that the failure of predictability can be explained by rejecting the governing conception of laws. Let us look at these two proposals in turn; in the next section, I will go on to give a different account of the origin of unpredictability.

Rummens and Cuypers distinguish between two kinds of predictability: external and embedded. *External predictability* is the possibility of a "(God-like) external observer, not part of the universe U, to make predictions of all the future events in U on the basis of its perfect knowledge of the initial conditions [and the laws of nature]" (p. 234). *Embedded predictability* "holds in a universe, U, if there exists a subsystem, S, embedded in U […] that is able to predict all the future events in U" (p. 235). According to Rummens and Cuypers, the paradox is resolved once we see that determinism implies external predictability, whereas the counterpredictive scenarios prove the impossibility of embedded predictability. There is of course no contradiction between the two. (This solution is highly reminiscent of Popper 1982s distinction between internal and external predictors, although Popper sets up the argumentative context somewhat differently.)

The problem for the embedded predictor, Rummens and Cuypers explain, is that it is constrained by a set of three equations that overdetermine their variables. Let

P be the predictor's prediction and A the counter's action. Then P and A are both determined by an equation that describes their dependence on the initial state of the universe; and in addition, the predictor has to satisfy P = A. But there is no guarantee that this satisfaction is possible, since the values of P and A are already determined. An external predictor, on the other hand, making prediction P* that is *not* a physical event in the universe and thus *not* determined by the initial conditions, does not run into the same problem. Its prediction can therefore always satisfy the constraint P* = A, that is, be correct. The difference between the two prediction events P and P*, Rummens and Cuypers say, is that P* is *not physical,* i.e., not part of the "law-like causal chain of events in the deterministic universe" (p. 233). Ismael (2019) seems to be in rough agreement with this analysis, as she tells us that the problem arises when there are causal interference effects: "If your predictions create disturbances in the domain you are trying to predict, it is not surprising that they limit predictability" (p. 491).

There's something right about this analysis of the paradox, but it cannot be correct as it stands. Let us assume that the external predictor makes its prediction through some well-defined process of reasoning that takes as input the initial state of the universe and the laws of nature. Then its prediction is in fact determined by an equation that describes its dependence on the initial state of the universe. And this means that it is in exactly the same situation as the embedded predictor.

Here is another way to make the same point. If the external predictor EX goes through some process of reasoning, then it is in principle possible to have an embedded predictor EM that goes through exactly the same process. Now suppose that EX is attempting to predict what will happen when EM encounters a counter. EX will then get into the exact predicament that EM is in: whatever it predicts, EM will predict the same, and hence the counter will do something else. EX may even realise that this is happening; but if so, EM is also realising what is happening. And in just the same way that I am unable to outsmart a machine that will always do the opposite of what I predict, so EX will be unable to outsmart the counter—whatever EX tries, his physical counterpart EM will by hypothesis try the very same thing, but to no avail.

That said, I do believe there is a way to sharpen the ideas of Rummens and Cuypers so that they *do* shed important light on the paradox. I will return to this in the next section, after presenting my own analysis.

The final proposal to be discussed is that by Ismael (2016, chapter 7). As we have seen, the paradox of predictability is based on the idea that already at the first moment of the universe (or at least at some time before the predictor makes its prediction) there are facts in place that determine the choice of the counter. But Ismael points out that we can read this statement in two ways. Reading one: the facts in place are the initial conditions of the universe, which *nomologically* determine the event to be predicted. Reading two: the facts in place are the initial conditions of the universe and the laws of nature, which together *logically* imply the event to be predicted.

These two readings may seem to be identical, since A nomologically implies B just in case (A + the laws of nature) logically imply B. But there is a subtle difference. On the second reading, but not on the first, we assume that the laws of nature are *already in place* at the initial moments of the universe. When we claim that the

predictor can use perfect knowledge of all the facts to deduce the choice of the counter, we are evidently adopting this stronger second reading, since we are assuming that *the laws of nature do not depend on the future choice of the counter*. This leads Ismael to the following argument:

1. If the initial conditions and the laws of nature were already in place before the predictor makes its prediction, it could make a correct prediction (Assumption.).
2. The predictor cannot make a correct prediction (Assumption.).
3. Either the initial conditions or the laws of nature are not already in place before the predictor makes its prediction (From 1 and 2.).
4. The initial conditions are in place before the predictor makes its prediction (Assumption).
5. Hence, the laws of nature are not already in place before the predictor makes its prediction (From 3 and 4.).

What does it mean to either affirm or deny that the laws of nature are already in place at the initial state of the universe? Roughly, one affirms this claim when one adopts a *governing conception of laws*, according to which later states of the universe are generated from earlier states by laws of nature that govern temporal evolution. And one denies this claim when one adopts a *supervenience* account of laws, according to which the laws of nature are some sort of summary of all the events in the (four-dimensional, timelessly regarded) universe. Thus, Ismael concludes that the paradox of predictability shows that we must adopt the second kind of theory.

One possible objection to this argument is that it doesn't matter whether the laws are already *in place* when the predictor makes its prediction, as long as it *uses* the laws in making its prediction. Even if the laws are not in place, the predictor might still guess the laws correctly—nothing in our scenario requires knowledge in any strong sense of the word—and then the paradox reappears. Perhaps Ismael can argue that whatever guess the predictor makes about the laws, the counter can then falsify this guess. But doesn't this require a dynamic, A-series theory of time that sits uneasily with a supervenience account of laws? This is certainly suggested by Ismael's frequent use of phrases like "already" and "beforehand". Since she herself rejects the notion that there is a fundamental asymmetry to time, the status of these phrases remains mysterious. However, pursuing such questions would take us too far afield in the current paper.

Other puzzles can also be raised for Ismael's account. It just doesn't seem to be the case that the counter requires the predictor to be ignorant of the laws of nature for its counterpredictive strategy to be a success. It seems that the predictor can know absolutely everything about the laws of nature—for instance, that we live in a perfectly Newtonian world—and *still* be unable to predict what the counter will do. Nor is there any reason to suppose that depending on what the predictor predicts, the universe will end up with different laws. (Will the laws of nature end up being different when you do not order the Merlot?)

But I will leave these puzzles aside too, for the most fundamental reason to reject Ismael's proposal is this: her argument is based on the assumption I have labelled 1, namely the assumption that *if* the initial conditions and the laws are in place, *then* the predictor can make a correct prediction. Without this assumption, no attack on the

being-in-place of laws can follow from the paradox of predictability. But assumption 1 is false. Predictability does not follow from determinism, not even when both the initial conditions *and* the laws are given and arbitrary computational power is assumed. To prove this is to defend my own analysis of the paradox, and I will do so in the next section.

4 Undecidability and Unpredictability

It is natural to assume that determinism implies predictability, at least when the laws and initial conditions are given. But however natural, the assumption is hard to hold on to in the light of Turing's (1936) paper *On Computable Numbers*, which contains the famous proof that the so-called 'halting problem' is undecidable. It pays for us to look at a non-formal version of this proof, since its resemblance to the paradox of predictability is striking.

Suppose that we have what has become known as a Turing machine, a deterministic general purpose computer that can be programmed to execute any algorithm on any input. Then the halting problem is this: decide, given a program and input, whether a Turing machine given this program and input will finish its calculations in a finite number of steps—that is, whether it will ever halt. Thus, we want a decision procedure that tells us correctly, for all programs and inputs, whether they will ever stop. This is essentially a predictive task: we are being asked to predict whether some machine will or will not halt.

Sometimes this is easy. Take the following program A(x):

```
if x is a number:
        while x is not 13:
                increase x by 2
stop
```

When the input of A is not a number, it will immediately stop. When it is a number, A will keep adding 2 to x until x is 13. This means that A will stop if x is 13, or 11, or 9, or any other number of the form (13—2n). But for all even numbers and all numbers larger than 13, A will go into an infinite loop and it will never halt.

Now we ask: can the halting problem be solved *in general*? Is there an algorithm H(x,y) which outputs 'true' if and only if program x would stop with input y, and which outputs 'false' otherwise? Given what we said about program A, we know that H(A, 7) should output 'true', while H(A, 6) should output 'false'. No doubt we can create an algorithm that gets these two cases correct. But could any H get *all* cases correct? Turing gives a very elegant proof that no such H can exist.

The proof is a proof by contradiction. We assume that H exists, and then we prove that H sometimes gives the wrong answer—which of course means that it is not really H. Assuming, then, that H exists, we construct an algorithm G(x) that works like this:

```
if H(x,x) outputs 'true':
        go into an infinite loop
stop
```

So if we put some input x in G, G first runs algorithm H to see whether program x would stop when given itself as input. (The fact that not all inputs might be valid programs is immaterial. We can just stipulate that invalid code halts immediately.) If H says yes, G goes into an infinite loop. If H says no, G stops.

What does H say when we ask whether G will stop with input G? That is, what is H(G,G)? Suppose, first, that H(G,G) outputs 'true'. Then when we actually execute G with input G, G will go into an infinite loop—we can just look at the way the program is set up to see this. Thus, H has given us a false answer. It claimed that G will stop when given input G, but in fact G will loop forever. Now suppose that H(G,G) outputs 'false'. Then when we actually execute G with input G, G will stop. Again, H has given us the wrong answer. So whatever answer H gives us, it will always be false. Thus, we must conclude that H cannot exist. There is no decision procedure that always tells us correctly whether a given program will halt on a given input. The halting problem is *undecidable*.

What does algorithm G remind us of? It looks structurally identical to the counter from the paradox of predictability. G takes the prediction that H makes about G itself, and then it does the exact opposite of whatever H predicted that G would do, just as the counter does when it meets the predictor. But this means that Turing's proof may shed light on the paradox.

Indeed, Turing's argument is a strictly *formal* proof that a deterministic system S cannot be predicted by a predictor P, given certain constraints. For our purposes, the one truly substantial constraint is that the method of prediction used by P must itself be able to occur in S. If we want to apply Turing's result to the paradox of predictability, this constraint becomes a *material* condition on S: S must be able to incorporate the process that generates P's prediction. If that condition is met, the formal proof allows us to show that P will not, in general, be able to predict the behaviour of S.

Before looking at how this insight compares to other treatments of the paradox, some remarks about the proof and its generality are in order.

1. One might worry that Turing's proof is essentially connected to the property of 'halting', which is not a property that a real predictor is necessarily interested in. In fact, however, the proof does not depend on the exact nature of this property. It can be shown—the result is known as Rice's Theorem (1953)—that all non-trivial semantic properties of programs/algorithms are undecidable. To be non-trivial merely means that some programs have it and some do not. To be semantic means to have to do with the *meaning* of the program, with what it *does* rather than with its nature as an uninterpreted string of symbols. Thus, we can of course give an algorithm for deciding the syntactical question of whether a given program starts with the letter 'f'. Just look at the first letter. But we cannot give an algorithm for deciding semantic questions such as whether a given program outputs the sign `0′, or calculates the first seventeen primes, or is semantically equivalent to some other program.

2. One might worry that Turing's proof is essentially connected to the idea of algorithms, and that perhaps some predictors—including humans?—do not use algorithms to make their predictions. But be the human mind as it may; as soon as one sits down with a set of initial conditions and some mathematical laws, and starts calculating what happens when the laws are applied to the conditions, one is presumably engaged in a task that could also be done by a computer.
3. But suppose that it *could not* be done by a computer; that the predictor is engaged in something that transcends all algorithmic computation. Then we are in effect asked to contemplate not just the set of all algorithms, but some larger set, a set that does not only include the algorithms but also certain algorithm-transcendent decision procedures. Let us call it the set of hyper-algorithms. Now it is true that Turing's proof could never show that a hyper-algorithm cannot predict all algorithms, since we can never set up the self-referential defeater needed for the proof by contradiction. But if our universe contains a predictor that uses a hyper-algorithm, then that predictor must do *more* than just predict algorithms if it wants to predict all events in our universe; it must also predict at least this specific hyper-algorithm. And so an equivalent of Turing's proof would be possible.
4. One might still have a residual worry about the application of a proof in mathematics to the real world. But what makes the idea of a Turing machine so useful is that it can be seen as a type of physical device *and* as a type of mathematical construction: all the properties that are relevant to the proof are shared by the machine and the mathematics. Hence, we can make the step from the *undecidability* of a mathematical problem to the *unpredictability* of a certain type of physical device. This is no more mysterious than the fact that we can use the truths of arithmetic ($1+2=3$) when dealing with physical objects that obey the same rules (if you bring one bottle of wine and I bring two, we'll have plenty).

We thus find that there is a strong analogy between the paradox of predictability as applied to the physical world and Turing's proof of undecidability as applied to mathematics.[2] What philosophical conclusions can we draw from this? First, we

[2] We are not, of course, the first to apply this kind of reasoning to problems of prediction. One interesting precursor is the diagonalisation argument of Putnam (1963), which is explained and discussed in detail by Kelly et al. (1994) and, more recently, Sterkenburg (2019)—see also the references therein. Putnam is interested in a certain type of *induction*, where we think of the data as an ordered (and algorithmically compressible) string of symbols that is revealed sign by sign. Now we try to find an algorithm that will *always*, *eventually*, become a perfect predictor of the remaining part of the string. What Putnam proves is that no such algorithm exists. This is of course a different kind of prediction than what we're interested in in the present article. In one sense, the task for our algorithm is *easier*, since it is given the laws of nature and initial conditions; in another sense it is *harder*, since we want perfect predictions not in the indefinite long run, but immediately. Another, much more recent, precursor is Seth Lloyd's (2012) argument that the halting problem explains free will. Lloyd argues that it takes longer to predict what your decision will be than to make that decision, and that this is why we feel we have free will. I must admit that there seems to be something of the non sequitur about that argument. It may take longer to simulate the flight of an airplane through a thunderstorm than to actually fly it through one, but that surely doesn't mean we cannot use simulations to determine in advance whether the plane will crash in thunderstorms? Certainly Ismael (2019, p. 494) is right to claim that it is if anything the possibility of *countering* your predictions, rather than any practical limits on computation time, that explains freedom here. (I doubt that this possibility of countering *exhausts* human freedom, but no matter.).

can conclude that any feature of the real world that does not have an analogon in Turing's proof must be inessential to the derivation of unpredictability. In particular, this shows our criticism of Ismael's proof was well-founded. According to Ismael, the only—or at the least the best—explanation for the unpredictability of the deterministic universe was the fact that the predictor cannot possess the laws of nature. But in the case of Turing machines, the laws of temporal evolution are perfectly known and indeed most intuitively thought of as *governing* the machine. Turing's proof shows that perfect knowledge of perfectly deterministic and perfectly determinate laws does not imply predictability. And so Ismael's argument does not go through.

But isn't a governing conception of laws still in trouble? For if we cannot algorithmically predict the future from the initial conditions, how do the laws *generate* the future from the past? This worry would be based on the idea that a governing conception of laws is committed to a computational view of how the laws work, to what Sterkenburg (2019) calls "some kind of physical variant of the Church-Turing thesis […] that what nature can *do* must be Turing-computable." But a governing conception of laws is not committed to any view of how the laws 'do' what they 'do'; not to any *particular* view, but not even to *having any view at all*. Perhaps the computational interpretation of the paradox of predictability can furnish an argument against the 'simulation hypothesis', the idea, well-known from science fiction, that we are all living in a computer simulation. But it cannot tell us anything more about the ontological status of laws.

Second, linking the paradox of predictability to Turing's proof allows us to state more clearly in what sense Rummens and Cuypers were right to make a distinction between external and embedded predictors. Remember that I argued in the previous section that there seemed to be no principled difference between an embedded predictor EM and an external predictor EX. Assuming that EM goes through the same process of reasoning as EX, it is a logical necessity that the defeat of the one implies the defeat of the other. And so for any EX, it seems that there will be a situation in which it faces the paradox of predictability.

But of course this assumes that EM *can* go through the same process of reasoning as EX. Now the mere idea that the external predictor is non-physical, i.e., not part of the causal structure of the universe, is not strong enough to undermine this assumption. But what if EX is 'God-like' not merely in the sense of being non-physical, but in the sense that its processes of reasoning do not allow of physical implementation in any of the universes U that it is supposed to predict? *Then* my counterargument does not get off the ground, since no process in U mirrors the external predictor's reasoning. In other words, the *material condition* that allows us to apply Turing's result to the system doesn't hold. So if the universes in question are so poor in structure that they cannot contain Turing-machines, then there might be an algorithm that perfectly predicts all events in all of them. Even if they can contain Turing-machines, there might be a non-algorithmic process of reasoning that allows for perfect prediction of all universes in U, *provided* that these processes do not appear in U itself.

There's another way in which external predictors can trump embedded predictors. Let EX use algorithm A and initial data D to answer question Q. And assume that

the entire process of applying A to D to answer Q is physically present in some but not all of the universes in U. Then EX will run into the paradox of predictability, but *only* in the universes in which this process is present. For all other universes, it may achieve success. But an embedded predictor EM going through this process of reasoning will always, by logical necessity, be in a universe that physically contains this process—since it contains EM itself. So there is a sense in which running into the paradox is more 'likely' for an embedded predictor.

There are, then, at least two ways in which external predictors can outperform embedded predictors: by transcending the computational possibilities of the universes they are meant to predict, and by limiting themselves to universes in which the paradoxical computations do not appear. Once interpreted in this computational way, the analysis of Rummens and Cuypers becomes essentially correct. So why did they themselves feel the need to stress the idea of the physical versus the non-physical? I will give a tentative explanation for that in the next section, where we will see that this distinction is *also* important for understanding the paradox, but only once we take a different perspective on it.

In conclusion: we must reject the idea that determinism implies predictability, and reject this on a combination of formal and material grounds: Turing's proof on the formal side, the possibility of embedding prediction processes into our universe on the material side. The laws of nature may determine that B will follow from A, without knowledge of those laws giving us an effective algorithm for computing or deducing B from A. Depending on what precisely transpires in a universe, correctly predicting it may require us to transcend its embedded computational possibilities.

5 Rationality and Constitutive Norms

The previous section dissolves the paradox of predictability. Yet a sense of mystery may remain. There is still something *strange* about the fact that when I am the predictor, my predictions are *bound to fail*; especially given that I seem to have total insight into the situation. In this section, I want to look at what it's like for us human beings to be a predictor. This will allow us to explain more fully the sense of paradox and to draw an additional lesson from the paradox—the lesson of the contingency of our rational activities.

Let us first take an extremely simple example of me being a predictor faced with a counter. I sit at a desk, facing two buttons. One of them is labelled "I predict that the light will turn on" and it does absolutely nothing. The other is labelled "I predict that the light will not turn on" and it acts as a light switch that turns on the light. (This is roughly the example of Good 1971.) I know what the buttons do; in fact, I can simply see all of the electrical connections. There is, then, no uncertainty in my mind: when I press a button, whatever prediction I make thereby will be false.

In this example, we are assuming that it is the pressing of the button that counts as the making of a prediction. Is this an innocuous assumption? One might, for instance, hold that a prediction is not a physical act of expression, but a mental act of judgement. If so, the example could be changed to involve not buttons, but a brain scanner, or some other device that is causally linked to my mental act of predicting.

The one thing that would undermine the example, and all other examples in this section, is if our predictions and other rational actions are *essentially private*, in the sense that they could not in principle be detected. But this would mean that they are in an important sense outside the causal order of the world, that there is at least one sense in which we are 'external' rather than 'embedded' predictors. I will assume that we are, if not material beings, then at least endowed with a physical manifestation, and hence not external in this sense.

What follows from the example? Does it follow that, like the hypothetical H in Turing's proof, I necessarily make a mistake when I predict what the light will do? No. Clearly, I cannot make a *correct* prediction. But, I want to argue, I also cannot make an *incorrect* prediction. Why not? Because, surely, no act counts as making a prediction unless it involves at least a minimally positive belief that the predicted event will actually happen. Anyone who says: "I predict that *p* will happen, but I do not think it more likely than not that *p* will happen" is either confused or abusing the logical grammar of prediction. In the absence of belief one might *guess* that *p* will happen; in the presence of a contrary belief one might *pretend to predict* that *p* will happen; but only with at least some positive belief can one *predict* that *p* will happen. To return to the example, pressing the button that states that the light won't turn on cannot count as predicting that the light won't turn on, given that I know perfectly well that my very pressing will turn on the light. So in these circumstances I just cannot make a prediction at all. Not making a prediction is the only rational attitude that a predictor can take towards a recognised counter; what's more, it is also the *only* attitude she can take. Known counters do not falsify predictions; rather, they make prediction impossible.

Now suppose that I am not faced with a counter, but with what Rummens and Cuypers call a *fatalist mechanism*. Where a counter is set up in such a way that it will always invalidate a prediction, the fatalist is set up in such a way that it will always make the prediction true. So let us suppose that we switch my two buttons, so that pressing the button that claims that the light will go on now actually makes the light go on. It may seem that I can now easily make a true prediction: if I predict that the light will go on, it will do so; if I predict it will not go on, it will not. But saying this would be a mistake. Given that I understand the situation, given that I know full well that whatever prediction I make, it will always come out true, we cannot possibly interpret my button pressing as a *prediction*. Rather, it is a *choice* or a *command*. I would be like the domineering husband telling his wife that he doesn't believe she wants another Merlot. He is not predicting her wishes; he is phrasing his command in a veiled way.

Does Ismael disagree with this analysis of prediction and choice when she claims that we should think of "decision as a kind of self-fulfilling prediction" (2019, p. 492)? It depends on how we read the phrase 'a kind of'. I would reject the idea that decision is *a species of* self-fulfilling prediction, but embrace the idea that decision is in some ways *analogous to* self-fulfilling prediction. Now it seems to me that we *should* reject the first of these ideas, because it is part of our understanding of prediction that it is an epistemic activity that tries to get something right, that tries to be adequate to theoretical—that is, not purely practical—standards. And this condition is not fulfilled in the case of the fatalist mechanism, or of choice in general. (Even if

one does not share the linguistic intuition that this is how we use the term 'prediction', one may still agree with me that engaging with a fatalist mechanism is a very different kind of activity from that of, say, predicting tomorrow's weather or the next presidential election; and that is really all the agreement we need to make the fundamental points of this section.)

These examples reveal an aspect of prediction that did not become clear from the computational approach of Sect. 4: namely, that certain conditions must be in place before we can *interpret* an event *as a prediction*. Program H will happily chug away at any input you give it; but *we* are interpreting what it does as an act of prediction, and thereby we apply certain normative standards to it that it cannot meet. The paradox of predictability can only be formulated from the point of view of an interpretative, hence normative, being—and this not merely in the trivial sense that only interpretative beings are able to formulate anything, but in the more profound sense that the paradox requires us to interpret some events *as* norm-governed activities, namely, as predictions. There is nothing paradoxical about me pressing a button and a light then not going on; the paradox emerges only when we attempt to understand my pressing as a prediction. But why don't we then simply conclude that this interpretation of the event is wrong? What instils the sense of paradox?

To make a prediction is to stand under certain norms of rationality. Now for some of those norms—call them *regulative norms*—breaking them means *predicting badly*. An example is the norm of taking on board all the evidence. If we fail to do that, if for instance we look only at the evidence that fits our prejudices, then we are predicting badly (even when our prediction turns out to be correct). But for other norms—call them *constitutive norms*—breaking them means *not predicting at all*.[3] An example is the norm of not predicting what you are certain will not happen. Regulative norms are norms you can fail to live up to. Indeed, we are all only too familiar with the phenomenon of failing to live up to such norms. But you can't really fail to live up to constitutive norms. Not obeying those norms simply means not engaging in a particular kind of conduct. If I make a statement about the future that I know to be false, I am not making a prediction at all. (This is true even if I want other people to believe that I *am* making a prediction.)

It is very natural to believe that whether or not we obey constitutive norms is entirely up to us. If I want to usher a prediction, I can. If I want to engage in deception, I can do that too. But the paradox of predictability shows us that this belief is false. There are situations in which certain types of norm-governed behaviour are impossible, even though we are in the full possession of our mental capacities,

[3] My distinction is related, but certainly not identical, to the classic distinction made by Rawls (1955) between the summary conception of rules and the practice conception of rules, where the latter thinks of rules as defining a practice. It approaches more closely to—and is expressed using the same terms as—the distinction made by Searle (1969, 1995). Searle thinks of constitutive rules as rules that define and make possible new social practices, and regulative rules as rules that regulate behaviour within pre-existing social practices. This distinction is perhaps neither perfectly clear nor ontologically significant—see Hindriks (2009) for an overview of some of the problems. But for our present purposes, it is enough that we *do* sometimes distinguish between doing something badly and not doing it at all, and that we do this in at least more or less systematic ways. This much is surely true.

because in those situations it is impossible to satisfy their constitutive norms. Such situations are not very familiar to us. Indeed, they seem to undermine our sense of rational self-determination; and that, I gather, is why the idea of being faced with a counter fills us with a sense of paradox. We assume that it is always possible to engage in any rational activity we would like to, at least when we are not incapacitated; and therefore we believe that it is always possible to make rational predictions. But no, it is not always possible to engage in any rational activity; and while we perhaps cannot be pushed into irrationality, we may well be forced into *non*-rationality. There are situations where we simply cannot take up the rational attitude that we want.

Do those kinds of situations only come up when we think about *predictions*? No. There is for instance Gregory Kavka's (1983) toxin puzzle, in which it is impossible to *intend* an act that is clearly within your power. Staying within the realm of theoretical cognition, consider *beliefs about the present*. Let us assume that there is a device that has direct access to my beliefs; perhaps I am wearing a helmet full of brain scanners. This helmet also comes equipped with a light bulb. Diabolically, the helmet has been programmed to work like this: if I believe that the light bulb is on, the light bulb is off; and if I believe that the light bulb is off, it is on. We can further assume that I know this. So I *know* that whatever belief I have about the state of the light bulb will be false. Since it is impossible to have a belief *and* know that it is false—"do not believe what you know to be false" is surely one of the constitutive norms of believing—it follows that I cannot have any belief about the present state of the light bulb.

But the makers of the helmet have been even more satanically devious. They have made it so that if I have *no* belief about the present state of the light bulb, then the bulb will be on. And I know *this* too. So I know that if I have no belief about the state of the light bulb, then it will be on. In this situation, the norms for believing break down in a dramatic way. I have a perfect reason *not* to have a belief about the light bulb, since any belief I have will be false. Hence, I cannot have a belief about the light bulb. And yet I also know that this means that the light bulb must be on. So I *have* a reason to have a belief about the light bulb, and in some sense I may even be said to have the belief. But knowing what I know about the helmet, that means that I must also not have it… well, clearly, there is no way out of this paradox. In this situation, I simply cannot meet the rational demands of belief. As far as belief about the light bulb is concerned, I am no longer a rational subject.[4]

Note that in this version of the paradox, all references to determinism have disappeared; and so has any idea of computation. *Predicting the behaviour of a deterministic system* is really only an example of a norm-governed activity. The paradox of predictability shows that it can be impossible to engage in that activity, because there are situations in which its constitutive norms cannot be satisfied. The paradox of the helmet shows that *having present beliefs about something* is another norm-governed

[4] Arguably, there is a clear demand of *practical* rationality here: take off that helmet before I go mad! But perhaps the helmet has been designed in such a way that it can only be taken off if the wearer has no wish or intention to do so.

activity for which the same result holds. We are led to speculate whether it might not be possible, with some ingenuity, to take *any* constitutive norm of *any* rational activity and devise a situation in which it cannot be satisfied, through absolutely no fault of the agent's and without the agent losing any of its normal cognitive powers. Whether that is possible or not, the paradox of predictability strongly suggests that whether a potentially rational agent can *be* rational is a contingent matter, dependent on that agent's circumstances. And this, I suspect, is what led Rummens and Cuypers to stress the idea of physical predictors, that is, predictors that are part of the causal structure of the world. For of course only such predictors have *circumstances* and can fall prey to the examples discussed in this section.

"Whatever the demon does to me, it cannot take away my rationality"—that is the fundamental assumption of Descartes' *Meditations*. When this founding idea of modern philosophy turns out to be wrong, it is bound to generate a sense of paradox. But it is wrong nevertheless, and the paradox of predictability helps us see this.

6 Conclusion

The paradox of predictability does not, as earlier authors have argued, depend on the predictor being part of the physical universe or unable to access the full laws of nature. Instead, we have seen that it follows immediately from several well-known results about the mathematics of computation, in particular Turing's proof about undecidability. The sense of paradox it creates can be traced to two sources: first, the mistaken idea that knowing the laws of nature and having an effective algorithm for prediction are more or less the same thing; second, the equally mistaken idea that it ought always be possible to take up a rational attitude. Blanchard (2017) tells us that the paradox "has much to teach us about freedom and determinism." To which I can add: about determinism, certainly; about freedom, yes, if freedom consists in the taking up of certain rational attitudes—and that might well be the case.

References

Blanchard, T. (2017). How physics makes us free. *Journal of Philosophy, 114*(3), 160–164.

Good, I. J. (1971). Free will and speed of computation. *The British Journal for the Philosophy of Science, 22*(1), 48–50.

Hindriks, F. (2009). Constitutive rules, language, and ontology. *Erkenntnis, 71,* 253–275.

Holton, R. (2013). From determinism to resignation, and how to stop it. In A. Clark, J. Kiverstein, & T. Vierkant (Eds.), *Decomposing the will* (pp 87–100). Oxford: Oxford University Press.
Ismael, J. T. (2016). *How physics makes us free*. Oxford: Oxford University Press.
Ismael, J. T. (2019). Determinism, Counterpredictive Devices, and the Impossibility of Laplacean Intelligences. *The Monist, 102,* 478–498.
Kavka, G. S. (1983). The Toxin puzzle. *Analysis, 43,* 33–36.
Kelly, K. T., Juhl, C. F., & Glymour, C. (1994). Reliability, realism, and relativism. In P. Clark & B. Hale (Eds.), *Reading Putnam* (pp. 98–160). New York: Blackwell.
Lewis, D. K., & Richardson, J. S. (1966). Scriven on human unpredictability. *Philosophical Studies, 17*(5), 69–74.
Lloyd, S. (2012). A turing test for free will. *Philosophical Transactions of the Royal Society A, 370,* 3597–3610.
Popper, K. R. (1950). Indeterminism in quantum physics and in classical physics. Part I. *British Journal for the Philosophy of Science, 1*(2), 117–133.
Popper, K. R. (1982). In W. W. Bartley III (Eds.), *The open universe: An argument for indeterminism*. Paris: Hutchinson. https://plato.stanford.edu/entries/popper/#Bib.
Putnam, H. (1963). Degree of confirmation' and inductive logic. In P. A. Schilpp (Ed.), *The philosophy of Rudolf Carnap* (pp. 761–783). New York: Open Court.
Rawls, J. (1955). Two concepts of rules. *The Philosophical Review, 64,* 3–32.
Rice, H. G. (1953). Classes of recursively enumerable sets and their decision problems. *Transactions of the American Mathematical Society, 74,* 358–366.
Rummens, S., & Cuypers, S. E. (2010). Determinism and the Paradox of Predictability. *Erkenntnis, 72*(2), 233–249.
Scriven, M. (1965). An essential unpredictability in human behaviour. In B. B. Wolman & E. Nagel (Eds.), *Scientific psychology: Principles and approaches* (pp. 411–425). New York: Basic Books.
Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge: Cambridge University Press.
Searle, J. R. (1995). *The construction of social reality*. Mumbai: The Free Press.
Sterkenburg, T. F. (2019). Putnam's diagonal argument and the impossibility of a universal learning machine. *Erkenntnis, 84,* 633–656.
Turing, A. (1936). On computable numbers, with an application to the entscheidungs problem. *Proceedings of the London Mathematical Society, 42*(1), 230–265.