Research Article

Free will and the paradox of predictability

Alexandros Syrakos[1]

1. Department of Mechanical and Manufacturing Engineering, University of Cyprus, Cyprus

In recent literature there has been increased interest in the so-called "paradox of predictability" (PoP) which purportedly shows that a deterministic universe is fundamentally unpredictable, even if its initial state and the laws that govern it are known perfectly. This ostensible conclusion has been used to support compatibilism, the thesis that determinism is compatible with free will: supposedly, the PoP shows that determinism is misunderstood and actually allows freedom, hence also free will. The present paper aims to disprove this conclusion and show that the PoP has absolutely no implication concerning the predictability of deterministic systems and the nature of determinism itself. Hence the PoP is irrelevant to the free will debate. Its paradoxy arises from a confusion between mental and physical notions in its formulation (the PoP tacitly premises a mental arbiter with respect to whom notions such as prediction and signification have meaning) and disappears once it is expressed in purely physical language. Ultimately, the PoP demonstrates not that prediction is impossible under determinism, but merely the obvious fact that it is impossible to predict while simultaneously acting so as to disprove your prediction. The impossibility of self-prediction is also discussed.

Correspondence: papers@team.qeios.com — Qeios will forward to the authors

1. Introduction

According to Spinoza, an early hard determinist,

> "[H]uman beings are mistaken in thinking they are free. This belief consists simply of their being conscious of their actions but ignorant of the causes by which they are determined. Their idea of their freedom therefore is not knowing any cause for their actions[1]". (Spinoza 1677, **Ethics** 2P35S [[1] p. 73]).

Science has made significant progress since the time of Spinoza, and the features of the physical world that modern like-minded philosophers believe to be the determining causes of our behaviour are now known and understoood to a significant degree: they are the fundamental laws of physics, which govern the behaviour of the countless fundamental particles that compose our bodies. But then the following questions arise naturally: if we know the causes that determine our actions, and we know the mechanics of these causes, can we predict our own actions? (Even if quantum-mechanical indeterminism does not allow us to make precise predictions, it still determines the probabilities of our potential actions, and these probabilities are predictable). After all, nowadays numerical simulations (predictions) of the behaviour of physical systems are routinely performed in many, if not most, scientific and engineering disciplines, exploiting the deterministic nature of the (macroscopic) physical world and its law-governed behaviour. And if indeed we can predict our own actions, which seems reasonable if we are just physical systems, what will happen if we choose to act contrary to those predictions, i.e. to act differently than what Spinoza's causes dictate? Our power to do just that seems equally reasonable, and having this power would disprove epiphenomenalism and prove that we have free will[2].

Let us consider a specific example in order to make things more concrete. A molecular dynamics simulation of our whole bodies will reveal what the physical laws have determined for our future (overlooking the facts that the required computational effort is, by current standards, so immense that such a simulation is practically impossible, and that there are epistemic limitations in knowing precisely the exact initial state of every atom in our bodies). So, suppose I had the appropriate equipment (a molecular scanner and a powerful computer equipped with molecular dynamics simulation software) and I used it to predict, via such a simulation of my whole body, that after two minutes I will get up from my chair, go to the fridge, open it and get something to eat. Having acquired this knowledge, I then deliberately decide to instead stay seated at my desk for a whole hour and watch YouTube videos on my laptop, despite my hunger. On first glance, there doesn't seem to be anything inconsistent with this scenario. Does this possibility refute physical determinism when it comes to humans[3]? Is this a hard problem for epiphenomenalism, and consequently for physicalism? This prospect would be highly welcome to people who, like me, desire that Cartesian dualism and free will libertarianism be true, i.e. that persons are not physical systems: if physical systems behave deterministically but persons do not, then persons are not physical.

Unfortunately, the above scenario is not as consistent as it may seem at first sight, and hence such a conclusion cannot be drawn so readily. That there is something wrong with our scenario will become

apparent if we consider the "paradox of predictability" (PoP), which is a paradox that highlights an impossibility to make predictions even about deterministic physical systems governed by laws, under special circumstances. The concept of the PoP is the following. Physical determinism entails that the state of the universe at any time instant in the past together with the physical laws determine all the future states of the universe from that instant forward. Laplace famously noted that a super-powerful intelligence (it later became customary to refer to it as "Laplace's demon") who knows the current state of the unviverse in all detail and the precise form of the physical laws can deduce the future down to the last detail [[2], p. 4]4. So, suppose the following scenario: the demon does predict the future of the universe, and somewhere in the universe there is a device, referred to as a "frustrator" or "counterpredictive device", which is such that it acts counter to any prediction of its behaviour that is revealed to it. Suppose then that the demon is somehow forced to reveal his prediction to the device; the device then acts counter to the demon's prediction, and hence the prediction is proved wrong. But how could this be if the demon knew precisely the initial state of the universe and the laws, and these determine the future states of the universe, including those of the frustrator? This is the paradox of predictability. An intelligent reader may have already noticed that the logic of the paradox is not completely sound but there are some weak points, in which the explanation lies. I will analyse the paradox shortly, but for now notice that the same weak points may be a latent part of our own argument that purports to show that we have free will because we can decide to act differently from a revealed prediction of our behaviour. Indeed, the paradox of predictability does not suppose that the frustrator is a person; it could be a physical object with no free will whatsoever. For example, it could be a device with two light bulbs, one green and one red, and two buttons, again one green and one red; the demon is instructed to predict now which bulb is going to be lit sometime in the future, and he has to indicate his prediction by pressing the corresponding button. But the device is wired such that if the green button is pressed then at the designated time only the red bulb is lit, and if the red button is pressed then at the designated time only the green bulb is lit. So, it seems that despite knowing everything about the universe, including precisely how the light bulb device is designed and constructed, the demon cannot predict correctly which bulb will light up. If he predicts that the green bulb will light up and presses the green button, then it will actually be the red bulb that lights up; and similarly if he predicts that the red bulb will light up, it will be the green bulb that actually does so. The frustrator device is purely physical and functions mechanistically. Hence it could be that our own ability to frustrate predictions made about our behaviour is physicaly explainable — it comes down to the way our brains are wired, and has nothing to do with free will and agent-causation[5].

The paradox of predictability has, in recent times, puzzled philosophers. It attracted attention in the 60′s and 70′s, when an intense debate about it arose in British philosophical circles, e.g. [3][4][5][6][7][8][9]. Subsequenty, this debate was largely forgoten until recently, when interest in the paradox rekindled [10][11][12][13][14][15][16][17][18]. In my opinion, the most accurate and detailed analysis of the paradox to date is that by Landsberg and Evans [6]. It is therefore unfortunate that that analysis has gone largely unnoticed, and that recent discussion exhibits a regression and deterioration in the current understanding of this paradox. For example, the PoP has been attributed to epistemic limitations of predictors (that they cannot know exactly the initial state or the laws) [14], or to a purported flexibility of the physical laws [[13], Chapter 7], [18], with the latter depending on our future decisions in what essentially amounts to agent-causal indeterminism despite the authors′ referring to it as "determinism". The explanation by Rummens and Cuypers [10] is neither entirely correct nor completely free of confusion (something that is admitted in [17]), but it is in the right direction. A couple of authors have noticed a similarity between the paradox of predictablility and Tuning′s halting problem [12][16], but this does not shed light on the nature of the PoP, and hence these authors were misled to incorrect conclusions[6].

In contrast to our original motivation of using this paradox as a means of disproving determinism and establishing the truth of libertarian free will, most of the above cited works attempted to use the paradox as a means of establishing compatibilism. Determinism is the thesis that all aspects of the future are determined by the past according to laws. To put it more precisely, if t_p and t_f are two times such that $t_p < t_f$, then the state of the universe at time t_f is completely determined by its state at time t_p according to a set of laws that govern it. The state of the universe at time t_p is itself, in exactly the same manner, determined by the state at any prior time $t_{p'} < t_p$, according to the same laws. So, essentially, the state of the universe at all times is determined by its initial state, at the beginning of time, and the form of the governing laws. Obviously determinism, by definition, means that any compatible definition of "free will" can bear only a superficial semblance to what we normally mean by these words, since our will is determined entirely by factors beyond our ultimate control. Yet many of the philosophers cited above thought or hoped that the paradox of predictability offers a way to reconcile determinism with free will, through the "loophole" of predictability. In particular, they thought that the paradox of predictability separates determinism from predictability in a strong, non-epistemic sense, making a deterministic universe unpredictable even in principle, even if the initial state and the laws are perfectly known. This would mean that determinism is ontologically different that what is commonly perceived, in particular that it actually allows some freedom and hence can be compatible with free will.

However, in my opinion such a view is entirely fallacious. It will be shown in what follows (as has already been shown in [6][10]) that the "paradox of predictability" actually does not have any implications concerning deterministic predictability. Its seemingly paradoxical character is due to a hidden inconsistency in the formulation of the problem: the "paradox" arises when the demon is asked to manipulate the system whose evolution he/she is tasked with predicting in a manner that, according to the deterministic rules that govern the system, necessarily invalidates the prediction. So, the "paradox" arises not because determinism allows some freedom, but because it does not allow any. We will examine both the case that the demon is external to the predicted deterministic system and the case that it is part of it. The latter case will further be shown to be impossible even if the demon is not asked to invalidate their prediction (self-prediction by a physical system is impossible).

It should be noted that the above compatibilistic line of argumentation is in fact an indirect acknowledgement that determinism is incompatible with free will, since it is acknowledged that for them to be compatible requires that the nature of determinism as currently perceived is false and that "determinism" actually be indeterministic. Unfortunately, just as the paradox of predictablility turns out to be of no help to the compatibilist cause, it also does not help the cause of dualism/ libertarianism — it does not give rise to a new "hard problem" for physicalism. Nevertheless it is noteworthy that the notion of prediction and the assignment of meaning to the actions of the demon that signify his/her prediction (e.g. pressing the green or red button) are not grounded in the physical realm, but in the mental one — they require minds. Confusion between the two realms is part of why the paradox of predictability seems paradoxical. If one analyses it from only a physical perspective then much of the paradoxy disappears, as discussed in the sections that follow.

2. Determinism, simulations and predictions

Predictability is a corollary of determinism, at least in principle: if states are determined by previous states according to logically comprehensible laws, then knowledge of the present (or past) state of affairs and of the structure of the laws enables us to predict the future states. It should be noted that physical predictability is a meta-physical notion: it is not itself something physical but it is *about* the physical world, how it will be in the future according to how it was in the past and the laws of its evolution. It implies a metaphysical mental understanding of the physical reality, it belongs in the world of meanings. Prediction is something that only a mind can do. However, we know by experience that our cognitive capacities have limitations pertaining to our memory capacity, our attention span etc. that do not allow

us to predict wholly mentally the evolution of complex physical systems, but we can resort to the use of physical aids ranging from simple stuff such as pen and paper to complex computers. In this case we try to replicate the evolution of the system under study with processes occurring in another system (the computer) whose components we relate with those of the original system by a signifier-signified convention (e.g. patterns of bits in RAM memory may denote particle masses, positions and velocities, and similarly screen pixels coloured appropriately may denote the same things, when we want to visualise the results). We program the computer so that the processes occurring therein reflect those occurring in the original system so that the evolution of the signifiers will reflect that of the signifieds. The processes occurring in both the original system and the simulator are governed by the same physical laws, but since these systems are very different we must employ considerable intelligence, ingenuity and scientific accumen to ensure that the computer processes indeed reflect those of the original system. Hence computational science is a very complex and demanding field. It must be reiterated that the processes occurring in the computer constitute a prediction *only with respect to a mind*, in which exists the conceptual link between signifiers and signifieds (such a link does not exist inherently in the computer, it is something mental); otherwise, from a purely physical perspective, what the computer does (moving arround electrical signals, changing the charge of capacitors etc.) has nothing to do with what the original, simulated system does (e.g., in the case of numerical weather prediction, the motion, temperature and humidity content of air masses).

Nowadays computational predictions and simulations are very common[7]. For example, meteorologists perform weather predictions that we hear about in the news and atmospheric scientists predict the dispersion of pollutants in the atmosphere; aerodynamicists perform predictions of the lift and drag forces that will be exerted on cars or airplanes of certain shapes if they travel at certain speeds, which enable them to optimise these shapes; engineers simulate the deformation of solid objects under specific loads, such as bridges, buildings, shafts and gears, which enables them to select their dimensions so that they can bear the loads witout breaking; chemical engineers and materials scientists use molecular dynamics simulations to understand and predict the emergence of macroscopic properties of materials from their microscopic structure and to design better materials; biochemists use molecular dynamics simulations to explore the functionality of proteins or the interaction of virus particles with our cells, to design new drugs etc. The evolution of numerical simulation into a mature, obiquitous and indispensable part of scientific and engineering practice during the last decades is due to the large advances in computer technology, which enabled computation at a scale previously unimaginable. From a

computational perspective, there is no difference between simulation and prediction; they are performed in exactly the same way. A prediction pertains to an actual, existing physical system, whereas simulation more generally can refer to a hypothetical scenario. For a simulation to be useful as a prediction, the computer and algorithms must be powerful enough such that the simulation progresses faster than the evolution of the actual simulated phenomena (it wouldn't make sense to attempt to predict something that will occur within an hour using a computer that needs 2 hours to compute the prediction). Nevertheless, oftentimes "slow" simulations provide extremely useful insight and understanding into complex phenomena that are unattainable by direct observation, such as molecular dynamics simulations that require many days of computational time to simulate the evolution of the conformation of protein molecules over the span of just a few microseconds [19]. In both simulations and predictions the ingredients are the same: the initial state of the system (initial conditions), the external influences it receives (boundary conditions), and the physical laws, usually expressed in the language of mathematics (differential equations) determine its temporal evolution. These ingredients are essentially the constituents of determinism, and they fully determine the evolution of the system in a logically comprehensible way, allowing us to predict or simulate it.

3. Determinism, predictability, and free will

Despite the impressive progress in computational science in terms of both hardware and software (algorithms), there are still significant limitations concerning our ability to simulate and predict. For example, there are many physical systems of interest that are simply too complex to be practically simulatable. If the simulation of the conformational evolution of a single protein over a few microseconds requires days of computational effort on powerful computers, then any meaningful molecular dynamics simulation of a whole human body is out of the question for the forseeable future, and perhaps will never be achievable.

There are also limitations of epistemic nature [20]. Simulations require knowledge of the initial state, the external influences, and the laws. In the case of simulations of hypothetical scenarios the first two (initial and boundary conditions) are hypothetical as well, and so can be considered error-free. But if the simulation concerns a real scenario, such as in the case of predictions, then errors or insufficient data concerning the initial and boundary conditions can have a significant impact on the accuracy of the results. Indeed, measurements cannot be perfectly accurate, and cannot be performed at every single point of the domain or its boundary, and hence we have to interpolate in between. In general, these errors

grow as the simulation progresses and eventually become as large as the variables we solve for, at which point the simulation results become useless. The rate of growth of the errors depends on the equations solved; some equations contain non-linear terms that amplify the errors at an exponential rate, making it very difficult to obtain accurate predictions beyond a certain point in the future. For example, meteorological predictions cannot be meaningfully advanced beyond two weeks into the future [21][22].

Furthermore, the equations themselves (expressing the physical laws) usually contain errors. Even molecular dynamics simulations employ simplified equations, for efficiency. As mentioned, molecular dynamics simulations are only applicable to very small systems, due to the immense computational power required. For larger systems we do not apply the fundamental laws of physics at the level of fundamental particles but macroscopic laws that derive from the former by a statistical averaging procedure (or even that are empirically constructed), whereby some microscopic information is unavoidably lost. Finally, when actually solving these equations our computational algorithms also involve discretisation errors, roundoff errors, and solution (iteration) errors — i.e. we can only obtain approximate solutions of these equations, even if the equations themselves were exact representations of reality.

Do these facts have any implications concerning the relationship between determinism and free will? To me, it is obvious that they have absolutely no implication. The laws of physics are independent of us and of our epistemic abilities, and exist since long before we existed ourselves — since the beginning of time. On the contrary, we ourselves are obviously not independent of the laws of physics. However, the question is whether or not we are *completely* dependent on them, whether physics determines everything about us. If so, then free will is merely an illusion and it is the physical processes occurring within us that give rise to feelings of will, desire, decision, thought, etc. all of which are entirely determined by the laws of physics. On the other hand, free will requires that we are substances that have freedom of choice that transcends (when it comes to the mind) and overrides (when it comes to the body) these laws, in fact any laws[8]. Therefore, the crucial point is whether our will is ultimately determined entirely by factors beyond us[9]. If so then free will is just an illusion and whether or not the deterministic basis of our behaviour is tractable enough so that we can predict our future will and actions is irrelevant.

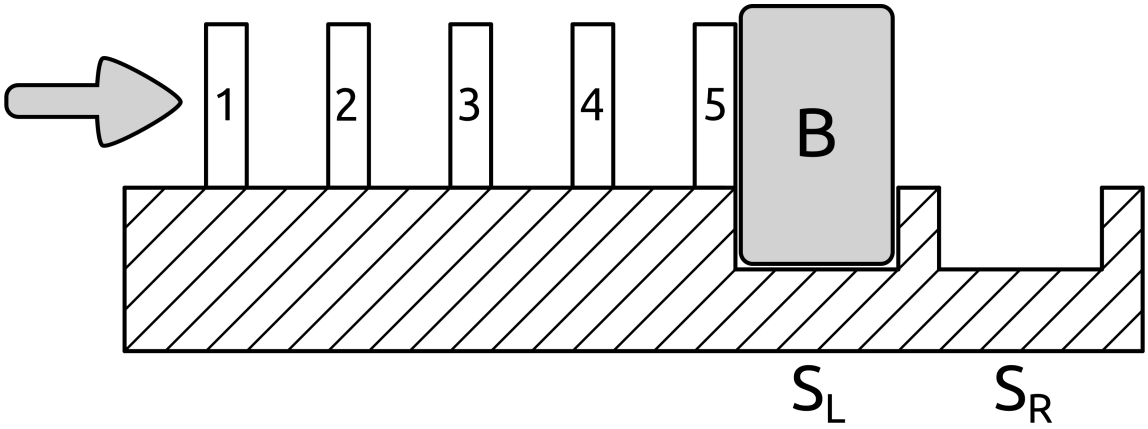
In summary, determinism means that the past and the laws completely determine the future; it is this that precludes free will. Predictability, on the other hand, is a subjective, first-person perspective notion that means that an intelligent agent, a mind, can deduce the future from knowledge of the past and the laws. In principle, predictability is a corollary of determinism. However, if the agent does not have perfect

knowledge of the past or the laws then he/she cannot predict the future perfectly. Nevertheless, this inability has no implication concerning determinism, which most of the philosophers studying the PoP seem to have acknowledged[10] (with a possible exception being [14]). However, what if unpredictability was due to non-epistemic reasons? What if, in a deterministic world, even if one knew precisely the past and the laws, and had unlimited intelligence, still it was impossible to predict the future? This would seem to indicate that determinism has been misconceived, which may raise hope for it being compatible with (genuine) free will, especially since unpredictability seems intuitively to be a characteristic of freedom. It seems that the philosophers who studied the PoP were motivated along these lines. They mistakenly perceived the PoP, perhaps driven by hope and enthusiasm, as demonstrating that there exist fully deterministic systems whose evolution is impossible to predict even if their initial state and the laws that govern this evolution are precisely known. The hope seems to be that such a discovery would show that determinism does not fix everything but allows some freedom. However, such enthusiasm is not warranted as drawing such a conclusion of unpredictability from the PoP is a fallacy, as will be shown next. Furthermore, the case of not everything being fixed has a name and it is not "determinism" but "indeterminism", and does not necessarily imply free will (e.g. quantum mechanical indeterminism).

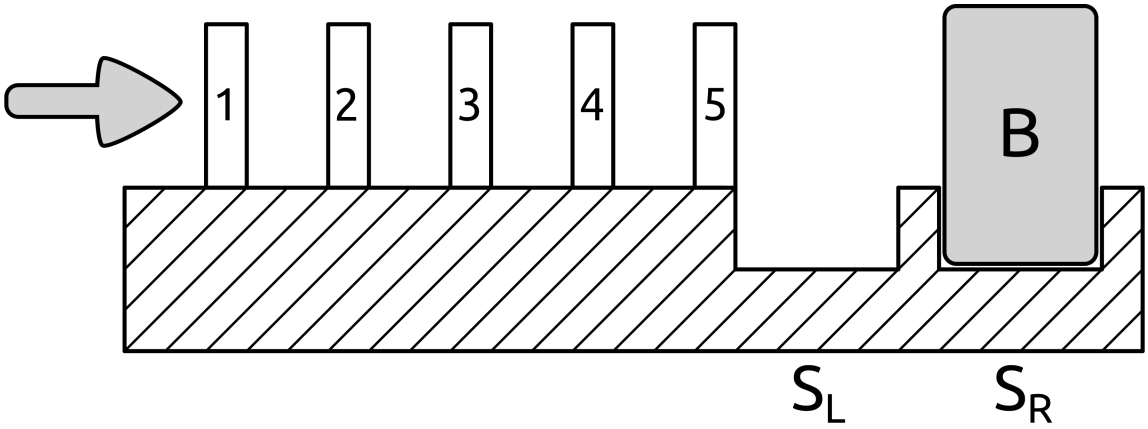
4. Analysis of the PoP when the demon is external

To make things clear and simple, consider a scenario illustrative of the paradox of predictability which takes place in a universe with simple deterministic rules where it is easy for us to play the role of the demon ourselves. So, let our universe be that depicted in Figure 1, consisting of a number of dominoes, D_i, for $i = 1, 2, \ldots, N$ ($N = 5$ in the setup of figure 1), a table with a flat surface on which the dominoes are placed and with two slots S_L and S_R beyond that surface, and a solid block B which can fit into any of the two slots. Initially, the dominoes are placed standing in succession, at a fixed distance to each other, close enough such that if one falls it will fall onto the next one and topple it as well. So, a law of this universe is that if D_i falls then D_{i+1} will fall as well, for $i = 1, 2, \ldots, N-2$. But for $i = N-1$, the law is somewhat more complex: if domino $N-1$ falls then so will domino N provided that block B is not located in slot S_L; otherwise, if block B is located in slot S_L then domino D_N will remain standing, buttressed by the block. These laws and the initial state of this universe determine its future states. Of course, in real life these laws would not be fundamental but would derive from more fundamental physical laws such as the law of gravity. But for our purposes let us regard them as fundamental. As part of the initial/boundary conditions, suppose that at time $t = 0$ a force (the arrow on the left in figure 1)

topples domino D_1; this is the "big bang" event that sets our universe into motion. We have not specified completely the initial conditions, in particular with regards to the location of block B, but we will consider both the case that B is initially located in S_L and the case that it is initially located in S_R. There are no internal factors in this universe that could cause the block B to move from its initial potition; it is too heavy, and locked in place by the slots, to be moved by domino D_N falling onto it.



(a) Block in the left slot; signifies that domino 5 is predicted to fall.



(b) Block in the right slot; signifies that domino 5 is predicted to remain standing.

Figure 1. The simple domino universe for demonstrating the paradox of predictability.

Suppose then that you are asked to play the role of the "demon", by predicting the final state of domino D_N. Obviously, the initial conditions and the laws determine that if B is in S_L then D_N will remain

standing, and if B is in S_R then D_N will fall. The paradox of predictability can be made to arise by imposing the rule that you have to indicate your prediction through placement of the block B thus: if you predict D_N to fall, place B in S_L; and if you predict D_N to remain standing, place B in S_R. With this rule, obviously it becomes impossible to make a correct prediction, because that would require that either B is in S_L and the last domino falls, or that B is in S_R and the last domino remains standing. Both of these scenarios are precluded by the laws of this universe.

This simple example sheds significant light on the origin of the paradox, and in fact the situation does not seem to be that paradoxical after all, but the paradox seems to arise artificially. Before we analyse the paradox, it is useful to point out that some of the explanations proposed in the literature are clearly irrelevant. In particular, the paradox is not due to any epistemic limitations on the demon's part concerning the laws or the initial state, as both are perfectly known. Furthermore, the laws are not "bent" or altered anywhere in the process; they are fixed, and in fact it is these specific laws that preclude a successful prediction under the given requirements; if the laws could be bent, then a successful prediction may have been possible (e.g. if the domino unexpectedly fell through the block, or remained standing without being buttressed). So, what is happening can be explained as follows:

1. In both cases (B in S_L or in S_R) the initial state and the laws completely determine the future states of this universe *as long as it is allowed to evolve on its own, without outside interference*. And this is perfectly reasonable — the initial state and the laws cannot be expected to determine external influences as well. The demon can unambiguously predict how this universe will evolve on its own, by advancing the initial state in time according to the laws.
2. But then the demon is asked to act on this universe, to interfere with its evolution. Tampering with the universe automatically nullifies any previous prediction that was based on the assumption that there are no external influences. If the demon interferes with the universe, then it is not just the initial state and the laws that determine the future, but the initial state, the laws, *and the external interference*. If the demon were free to interfere however he/she pleased, then the evolution of the universe would be *underdetermined*. The demon would not only be able to predict the future, but also to shape the future. There would not be a single possible future but many, with the demon being able to choose among them by interfering appropriately.
3. However, the demon is not free to interfere however he/she pleases, but must act according to a set of rules whereby his/her action is assigned a meaning concerning a prediction about the future. The correspondence of meanings to actions is deliberately set so that any prediction will be invalidated

by the corresponding action. The demon is not asked to do something that he/she does not know how to do, but something that is impossible. In particular, what he/she is effectively asked to do is to act in such a way that either B is in S_L and D_N falls, or B is in S_R and D_N remains standing. Although he/she is given a choice between two scenarios, neither of these scenarios is possible and therefore the demon's freedom to choose amounts to nothing.

Already the "paradox of predictability" does not seem paradoxical at all. Let us see a couple more instances in other contexts. The first is in the context of mathematics. Suppose that the value of x satisfies the equation:

$$x - 1 = 0 \qquad (1)$$

You are asked to calculate the value of x. This is easy: $x = 1$. However, for revealing your answer you are asked to tamper with the above equation, converting it into the following:

$$x - 1 = 0 + y \qquad (2)$$

where y is a variable to which you can assign a value. Obviously, the problem (2) is not equivalent to the original problem (1). The new problem (2) has infinite solutions of the form $x = 1 + y$, one for each value of y that you choose. Nevertheless, the way that your task is communicated to you makes it appear as though it concerns the solution of (1), although the instructions imply that you are in fact required to solve (2). This misleading formulation is a common theme in variants of the paradox of predictability. But the rules require something further: they require of you to communicate your solution in a way that affects the solution itself. Suppose that you are required to communicate your solution by assigning it to the variable y; that is, you must choose the value of y such that

$$y = x \qquad (3)$$

Comparing equations (2) and (3) we can see that it is impossible to satisfy them both, whatever the values of x and y (if, according to (3), we substitute x for y in (2) then the resulting equation can only be satisfied if $-1 = 0$, which does not hold).

Equations (1)-(3) and the associated discussion are equivalent to steps 1-3 of the domino case outlined above. To conclude that the paradox of predictability implies that predictability is not inherent in determinism is equivalent to concluding from the above procedure that equation (1) is not solvable.

As another example consider the following computer program:

```
1  program predict_me 2
3  print “What will I choose (true / false)?”
4  read prediction
5
6  if prediction == true then
7  choice = false
8  else
9  choice = true
10 end if
11
12 print "My choice is: ", choice 13
14 end program predict_me
```

The program asks the user to input a prediction about its output, in the form of a boolean (true/false) value, and then returns the opposite of what the user predicted. Yet the functionality of the program is fully deterministic, and the user who can read the code of the program understands fully how it works.

Many more similar examples can be devised. For instance, suppose that there is an empty square drawn on a sheet of paper and you are aksed to indicate whether after a few seconds it will be filled or empty, but to indicate that it will be empty you must fill it with your pencil, and to indicate that it will be filled you must leave it empty. Or suppose that in a physics exam you are asked to hold a pencil one metre above the ground; then, a timer is started and you are asked to predict, using your knowledge of physics, the reading of the timer (in seconds) at the time instant that the pencil will impact the ground. However, the catch is that you must indicate the number of seconds to impact by holding the pencil for the same number of seconds before letting go (e.g. if you predict the pencil to impact the ground at time $t = 2$ seconds after the timer was started, you must hold it for 2 seconds after the timer was started and then let it go — but then it is impossible for your prediction to be true unless the pencil falls with infinite speed).

In all the previous examples, the prediction/calculation rules that give rise to the PoP consist of heterogeneous parts coupled together in an artificial and unnatural manner. One part is a physical action to be performed on the system under examination, and the other part is a meaning assigned to this action, which refers to a (future) physical state of the system. The action to be taken is the signifier and

the predicted state is the signified, and they are linked together by a mental convention, decided and defined by a mind, without any inherent underlying physical link[11]. Hence, there is nothing precluding that the signifying action being taken and the signified predicted state of the system are incompatible. It is noteworthy that such a mapping can exist only in a mental observer, in a mind that exhibits (meta-physical) understanding of the physical world; it is a meta-physical notion: it concerns the physical world, but itself lies outside of the physical realm and inside the mental realm of meanings and understanding.

The (inevitable) arbitrariness of the mapping between signifier and signified may be concealed by giving the signifier some feature that makes it resemble with the signified or that alludes to a reference to the signified. For example, in the device with the light bulbs and the buttons that was mentioned in Section 1., the significance of each button was highlighted by its colour, which was the same as that of the signified light bulb. Also, in the domino example we could have named the slots S_{falls} and S_{stands} instead of S_L and S_R, and in the mathematical example we could have named our prediction variable $x_{\text{predicted}}$ instead of y. In the computer program example we chose appropriate names for the variables, `prediction` and `choice`. But these have only a psychological effect and play absolutely no role in the actual mechanics of the PoP. Hence in order to have a clear picture of these mechanics one should discard these extraneous elements. Doing so clearly shows that, in the domino example, the paradox merely comes down to finding a possible scenario between the choices (B is in S_L and D_N falls) and (B is in S_R and D_N remains upright), neither of which is possible. In the mathematics example it amounts to finding values for x and y that satisfy both equations (2) and (3); again, no such values exist. And in the computer program example it amounts to finding what value to input to variable `prediction` so that at the end of the program this value is the same as that of variable `choice` — again, impossible by program design. The notion of prediction really has nothing to do with it, it is not inherent in the physical aspect of these problems. It is a meaning that we assign to them.

So, this is quite disappointing. Apparently, there is no paradox after all. There is nothing obscure or mysterious about determinism. It works precisely as expected. Each of the systems considered is entirely predictable, and it is because of this that the demon finds himself in an impasse. It is because whatever action he decides to perform on the system he predicts, the outcome, according to the deterministic rules, is not the desirable.

5. Internalising the demon

Up to this point, e.g. in the cases of the bulbs and buttons, of the dominoes, of the mathematical problem, of the computer function etc., it was implicitly assumed that the demon was an outsider, that he was not part of the deterministic system under study. Now we will consider the demon as part of that system, by augmenting the latter. This has the advantage that there are no external influences on our extended system; it is closed. Hence, if this whole system is deterministic, its evolution will indeed be determined by its initial state and the laws, including the initial state of the demon and the laws that govern him. In the language of mathematics, the evolution of our system will be determined by the initial conditions and the laws, while, contrary to the previous cases we examined, there are no boundary conditions because our system is isolated.

Where did the sense of paradox arise from in the first place? Let us return to Laplace's vision, which encompasses the whole universe. If by "universe" we mean everything in existence, including the demon, then there are no external influences (since there exists nothing outside of that universe), and if that universe is deterministic then its initial state together with the laws determine all its future states. Therefore, if the demon is indeed capable of predicting the future based on the past and the laws, he should be able to know everything that will happen in the future, including the existence of the counterpredictive device and how it works, his own existence and what he is bound to do in the future, the fact that he will be asked to predict the frustrator's behaviour, and the outcome of this prediction. So, it may seem strange that, due to the conterpredictive nature of the frustrator, the demon is unable to come up with a clear answer to the question posed, a clear and acccurate prediction. This has led many to the conclusion that there is something wrong with determinism. But in actuality, just like in the case where the demon is external to the universe, the present scenario is also inconsistent as will be shown in the discussion that follows.

5.1. Demon with free will

First of all, by incorporating the demon into the universe, the properties of the demon himself with respect to determinism reflect on the properties of the universe as a whole. We assumed that the universe is deterministic, yet the demon has been regarded up to this point as an intelligent agent, a mind, and whether or not minds have free will is an issue of debate. Therefore, let us first briefly condider the case that the demon is a mental substance, a Cartesian mind with free will, a causal agent. The demon therefore is not governed by deterministic rules but has freedom of choice, and he/she thinks, decides

and acts based on reasons rather than causes[12]; the demon's behaviour is indeterminate and therefore unpredictable. In this case, going back to the domino example (and putting aside the signification convention), he knows that if he places the block in slot S_L then the domino D_N will remain standing and if he places it in slot S_R then it will fall. But in which slot he will place the block, in fact whether or not he will place it somewhere at all, is not determined but is up to him to decide[13]. So, in this case the initial conditions and the laws do not suffice to determine the future evolution of the universe, since the latter contains a source of indeterminism and unpredictability, the demon, and therefore the evolution is governed by the initial conditions, the laws, and the demon's free will (agent causality). Neither the demon himself nor any other demon, even if external to this universe, can predict that universe's future because it is not determined. This is the status of our actual universe if persons are indeed causal agents, with free will — each of us is a small shaper of the universe.

5.2. Deterministic demon

Let us now focus on the case that the demon is a physical (deterministic) system. Then his functionality would be as deterministic as that of the rest of the universe. In that case then, could we not ask the demon to predict, from the initial state and the laws, not only the evolution of the universe outside of him, but of the whole universe, himself incuded? What then, if the universe contains a counterpredictive device? Or, we could even imagine a scenario where the whole physical universe consists of the demon only: the demon is a computing machine that predicts his own future state, but also is programmed so as to act as a counterpredictive device, acting contrary to his own predictions. (Actually, there is no substantial difference between these two cases; if the demon is a complex physical object then whether or not the counterpredictive device is part of him is a matter of convention, as discussed in [[23], comment 7]). Note that actually we do not have to consider the whole universe, but any isolated part of it that includes the demon suffices. For example, consider an isolated room that contains the domino arrangement and a "demon" that is a powerful computer equipped with molecular dynamics simulation software and with a mechanical arm that can move the block B from one slot to the other. The computer is programmed to do a complete simulation of whatever is contained in the room, including its own self, and based on its prediction of whether domino D_N will be fallen or standing at a certain time $t_{\text{after}} > t_{\text{event}}$, where t_{event} is the time instant when the domino wave reaches D_N, to move the block B to the appropriate slot S_L or S_R according to the signifier-signified convention. The moving of the block is to take place at a certain time $t_{\text{before}} < t_{\text{event}}$ (and hence will be included in the simulation as well). There

are no external influences on this system, and therefore the computer can predict the future from only the initial state of the system and the laws. Since everything is deterministic, whether D_N will be standing or fallen and whether B will be in S_L or S_R at t_{after} are determined entirely by the initial state and the laws; there is only a single temporal path to take.

Obviously the computer cannot make a correct prediction because the signifier-signified rule is still contradictory. But this version of the paradox may perhaps indeed seem more paradoxical than the previous one where the demon was an outsider. At first glance, the ingredients to set up this scenario are available: computers and molecular dynamics simulation software already exist and are used; the domino setup is simple; and it is straightforward to program the computer to move the block, using the mechanical arm, to the slot indicated by the simulation results. The initial conditions are just a snapshot of the initial locations and velocities of all the atoms contained in the room, so one just needs to provide the blueprints, in terms of molecular structure, of the various components (the computer, the arm, the dominoes etc.), while the laws are known. It is not obvious where the procedure will stumble, and yet it must necessarily do so.

A closer look reveals that the problem is, essentially, the same as with the scenario where the demon was external to the system. Again, the thought that it is straightforward to program the computer to move the block to the slot indicated by the prediction implies the (false) assumption that there is a one-way dependence of the placement of the block on the prediction: the prediction is performed first in an unambiguous manner, and then the block is moved according to what was predicted. However, actually it is also the prediction that depends on where the block is to be placed. This is essentially the same as the tacit false assumption in the external demon case that it is only the prediction that depends on the future, whereas the future also depends on the prediction through the signifier event. In both cases, it is impossible to satisfly this interdependence both ways because of the self-contradicting signifier-signified rule.

To understand the problem, let us follow the progression of the simulation from its beginning. Let us assume that this PoP problem formulation is well-defined so that there are no ambiguities in the initial conditions; all the components of the system, and therefore their molecular structures, are clearly defined (an assumption that will subsequently be shown to be false). If this is so, then the laws dictate a specific, predictable path towards the future, which the computer can predict. Molecular dynamics simulation software normally use explicit algorithms that advance the prediction in time; this means that when the

computer is predicting what will happen at time t_n, say, then it has already predicted what will happen at all times $t < t_n$ and it has not yet predicted what will happen at any time $t > t_n$.

So, suppose that the computer has computed the prediction for all times $t < t_{\text{before}}$ and is now about to predict the state of the system at time instant t_{before}. At that time instant, its future self will have to move the block B to the appropriate slot indicated by the signification rule. In order to decide where to move block B, the computer's future self will use its own prediction of the state of the domino D_N at time t_{after}. However, the computer's present self has advanced its own prediction only up to time t_{before} and hence does not know yet whether D_N at time t_{after} will fall or remain standing; therefore, it is unable to deduce whether its future self at time t_{before} will move block B to slot S_L or to slot S_R. As a result, it is unable to advance its prediction beyond time t_{before}.

Therefore, things are not as straightforward as it may have seemed at first glance. In particular, this concerns the initial conditions, which define, among other things, the structure of the computer and its functionality (how it is programmed, which algorithm it is to execute). In order to set the initial conditions for the simulation we need to find an algorithm that can bring the combined tasks of prediction and communication to completion (this algorithm will be embedded in the initial conditions), and the algorithm just described will not do.

One may try to overcome the problem by iteration: program the computer to start its calculations by assuming what the future at t_{after} will be, and then repeat the prediction again and again until successive predictions are indistinguishable (i.e., in technical language, until the calculations have converged). So, since our calculations have got stuck at t_{before} because they do not yet know what will happen at t_{after}, let us assume that the domino will fall in order to proceed; we will correct this assumption later if it turns out to be wrong. With this assumption, the computer will predict that its own self will, at time t_{before}, move the block B to slot S_L, according to the signifier-signified rules. Continuing the calculations from that point on, when time t_{after} is reached, it will be predicted that the domino D_N will actually remain standing, since B is in S_L and obstructing its fall. Hence the assumption that the domino will fall turned out to be incorrect. Therefore, the calculations will be repeated, with the (hopefully better) assumption that the domino will remain standing. But obviously these new calculations will conclude that the domino will fall, and additional calculations based on these will conclude that the domino will remain standing and so on. Our predictions will perpetually oscillate between the domino falling and standing and will never converge. Therefore this strategy fails as well.

The last resort is to program the computer to solve the problem implicitly, computing the states of the system at all times simultaneously. Since the future depends on the past and the past also depends on the future, let us not calculate one before the other, but both at the same time. This is the most expensive method, since the computer must have enough memory to store all the states of the system at all times $t \in [0, t_{\text{after}}]$ (it may have already occured to the reader that this is impossible, since the state of the whole system at any time instant includes the state of the computer's memory, so in order for all these states to fit into the computer memory, the computer memory must be larger than its own self, by multiple times; let us overlook this for the moment, but we will return to it shortly). However, if the problem has a solution, this procedure will find it.

The state $\mathcal{S}$ of the whole system at time t will be related to the initial state $\mathcal{S}_0$ of the system according to the physical laws, expressed through the function $\mathcal{L}$:

$$\mathcal{S} \;=\; \mathcal{L}(\mathcal{S}_0, t) \qquad (4)$$

Part of the initial state $\mathcal{S}_0$ of the system pertains to the computer (which is part of the system), and in particular it defines its design and functionality, how it is built and programmed. Let $\mathcal{S}^c \subset \mathcal{S}$ be the part of $\mathcal{S}$ that describes the computer. Since the computer performs its predictions physically, it is its molecular structural arrangement that enables it to do this. The operation of the computer is governed by equation (4): its structure $\mathcal{S}_0^c$, embedded in $\mathcal{S}_0$, is such that under the effect of the physical laws $\mathcal{L}$ it evolves at time $t > 0$ into $\mathcal{S}^c$, embedded in $\mathcal{S} = \mathcal{L}(\mathcal{S}_0, t)$, which can be interpreted (by a mind) as representing the state $\mathcal{S}' = \mathcal{L}(\mathcal{S}_0, t')$ of the whole system at a later time $t' > t$. At time t_{before}, the state of the system is $\mathcal{S}_{\text{before}} = \mathcal{L}(\mathcal{S}_0, t_{\text{before}})$, and the part of this state that constitutes the computer, $\mathcal{S}^c_{\text{before}}$, represents the state of the whole system at time t_{after}, $\mathcal{S}_{\text{after}} = \mathcal{L}(\mathcal{S}_0, t_{\text{after}})$.

So, let us think how we could design $\mathcal{S}_0^c$ such that the computer would complete the required tasks. First of all, we note that no matter what algorithm the computer is programmed to execute, the domino part of the system (shown in figure 1) remains the same; it is a part of $\mathcal{S}_0$ disjoint from $\mathcal{S}_0^c$. The setup of this part is such that, according to the laws $\mathcal{L}$, at time t_{after} the state of the system, $\mathcal{S}_{\text{after}}$, will necessarily exhibit exactly one of the following:

$$(\ B \text{ is in } S_R \text{ and } D_N \text{ falls}\ ) \quad \text{or} \quad (\ B \text{ is in } S_L \text{ and } D_N \text{ stands}\ ) \qquad (5)$$

The above holds irrespective of how $\mathcal{S}_0^c$ is set up, i.e. of how the computer is built and programmed.

doi.org/10.32388/FHH7ZU

However, in addition to equation (4) which is imposed by the laws of nature, we want $\mathcal{S}_{\text{after}}$ to satisfy the signifier-signified convention, imposed by us, which requires that at time t_{after} the state of the system, $\mathcal{S}_{\text{after}}$, exhibits exactly one of the following:

$$(\ B \text{ is in } S_R \text{ and } D_N \text{ stands}\ ) \quad \text{or} \quad (\ B \text{ is in } S_L \text{ and } D_N \text{ falls}\ ) \qquad (6)$$

We want to impose this convention, but it must be implemented by physical means, i.e. we are seeking a computer/algorithm $\mathcal{S}_0^c$ such that equation (4) leads to one of the outcomes listed in (6). But we saw that no matter what computer we choose and how we program it, i.e. whatever $\mathcal{S}_0^c$ is, the only possible outcomes allowed by equation (4) are those listed in (5), which do not inlude any of the outcomes (6). Hence it is aparent that no matter how we program the computer, or how powerful it is, it cannot perform the task that we want it to.

So, the initial intuition that it is straightforward to put together such a system by placing in an isolated room the domino table of figure 1 and a powerful computer equipped with a mechanical arm, installing a molecular dynamics simulation software package on the computer, providing it with the initial locations and velocities of all the molecules in the room (including those of the computer) and programming it to foresee the future and move the block B to whichever slot the prediction indicates according to the signification convention, is false. The easiest way to see this is by considering that when the time comes to predict the motion of the arm the prediction will falter, because this requires knowledge of whether D_N will fall or remain standing which has not been predicted yet, and in fact depends on the arm motion. But we also explored other ways that may at first glance offer some prospect of getting arround this problem and they all fail, due to the self-refuting nature of the signification rule. With a clearer view, it does not seem at all mysterious that the demon who is part of the predicted system cannot satisfy the contradictory signification requirement, and this situation is no more paradoxical than the one where the demon is external to the system. It has no implications with regards to determinism (in fact it is the deterministic laws (5) that do not allow the construction of a machine $\mathcal{S}_0^c$ that can satisfy the contrived rules (6).

6. Self-prediction

But, it turns out that the scenario where the demon (computer) is part of the physical system whose evolution he is to predict is even more problematic than the one where the demon is external. The reason is that the computer is unable to fulfil its task even if there are no contradictory signification rules. In

particular, the computer cannot predict its own future, which is something that was already pointed out by Popper [24][25].

First of all, it should be reminded that since the computer is a physical system, prediction — a metaphysical, not a physical notion — is not, strictly speaking, something that it can do. What it can do is physical stuff: move components, push electrons, emit photons, etc. Rather, the actual prediction must be made by a mind, an observer with metaphysical capacity and understanding of reality, who uses the computer as an aid. For the computer to be useful as a prediction aid, it must (a) have an internal structure that can be construed as *representing* the structure of the predicted system (representation is again a mental notion, it is not inherent in the computer structure but exists in the mind who interprets is as such) and (b) the computer structure must evolve, under the laws of physics, in such a way that it continues to represent, according to the same mental mapping rules, the structure of the predicted system as that itself evolves under the physical laws. In order for such a process to be useful as a prediction, the computer processes must be occurring at a faster pace compared to the actual processess they mimick.

For example, in a weather forecast the density, humidity, temperature, and velocity of volumes of air in the atmosphere are represented by patterns of bits in the computer's physical memory, according to a signification convention defined and perceived by a mind (based on the concept of the binary number system). The electronic operations within the computer's circuitry and processor are set up so that the evolution of these patterns of bits will parallel the evolution of the state of those volumes of air they represent; the patterns of bits in the computer and the motion and thermodynamics of the air are governed by different physical processes, but the programmer of the computer ensures that there are mathematical analogies between them such that the same signifier-signified convention will continue to hold as both the bits and the air independently evolve in time. The mapping between signifier and signified is completely mind-dependent: there is nothing inherent in the computer's structure and processes that physically ties it to atmospheric processes; the tie is entirely mental.

In order for a signifier-signified rule to be establishable between two systems such as the computer and the atmosphere, the two systems must have the same complexity, i.e. the same number of components (also called degrees of freedom). To achieve this, we usually have to coarsen the predicted system's conceptual decomposition into components. For example, the atmosphere consists of molecules but a one-to-one mapping between these molecules and the computer memory components is impossible, due to the exponential number of molecules. Therefore, in practice we split the atmosphere into a number of

volumes, each of which contains innumerable molecules, and it is these volumes that are mapped onto the computer memory. Coarsening results in some loss of information and therefore of accuracy in the predictions, but this is acceptable as long as the coarse structure is still fine enough to capture / give rise to all the aspects of the mechanics / dynamics of the predicted system that we are interested in within acceptable error bounds.

Now let us consider a computer that must be programmed to predict its own self in the future. In order to do that, a map must be established between the components of its future self and the components of its present self. The computer memory (both in the present and in the future) must store the representation of its future self (the data) and the instructions of how to act on this data to advance the prediction further (the algorithm). Now, a full molecular dynamics simulation of its own self is out of the question for the computer, since the smallest storage units it has available are bits (and usually these must in fact be grouped together in larger units of, say, 32 or 64 bits, corresponding to single and double precision arithmetic, respectively, to be able to store any useful information such as particle mass or velocity), each of which consists of a large number of atoms. Hence the computer contains many more atoms than representation units, and cannot represent itself atomistically. But we do not need to represent every single atom; in order to represent the functionality of the computer it suffices to represent its individual memory bits. Whatever individual atoms are doing inside a bit is of no consequence, only the state of that bit as a whole matters (whether it is in the ″0″ or ″1″ state). But further discounts cannot be made, since a single bit may play a crucial role for the progression of the executed algorithm. For example, in a conditional statement (″`if ... then ... else` …″) the value of a single bit can determine the flow of the algorithm.

Therefore, an obvious map is to let each memory bit at present represent its own self in the future. But this map will not do, because it requires that each bit has now the exact same state as it will have in the future, so that the computer state as a whole must now be exactly the same as what it will be in the future. This map is applicable only in the trivial case that the computer state does not evolve in time (the future state is the same as the present state, i.e. the computer is idle), or in the other trivial case where the pace of prediction is the same as that of the actual flow of time, and what is ″predicted″ is actually the present rather than the future (i.e. the ″future self″ that the computer predicts at time t is actually just its present self, at time t also, hence their bit patterns are identical). In other words, if the computer at time t represents its future state at time at then for this to be a prediction requires that $a > 1$ whereas in this trivial case we have $a = 1$. Clearly, therefore, this mapping is not satisfactory.

But is there any other better mapping that will allow for prediction in every case in general? It seems not. First of all, at time $t = 0$, when the prediction begins, the computer must represent itself as it is now, i.e. respresentation and represented are necessarily one and the same: the current bit pattern construed as a representation represents the same bit pattern. Therefore, each bit must represent its own self by necessity, according to the aforementioned trivial mapping, if the mapping is to be applicable no matter what pattern of bits the computer initially has. Then, since the task of prediction requires that the predictor can calculate faster than the predicted system operates, the computer must advance its internal representation of its future self faster than it itself can calculate, with its current bit pattern representing, at the same time, both its current state and an evolving and diverging future state. This is not possible, no matter what fixed mapping we choose between the current and future bit patterns. In self-prediction, predictor and predicted system are one and the same; they have the same complexity, run the same algorithm, and at the same speed. Hence the predictor engages in a futile race to outpace his own self. It therefore seems that the best that can be done is that both the predicted state (internal representation) and the actual state progress at the same speed, hand-in-hand — the aforementioned trivial case.

Consider the repercussions if indeed self-prediction were possible. A prediction is, by nature, performed faster than the predicted system evolves. So, suppose that the computer predicts its own behaviour at a rate twice that of the actual flow of time. This means that after it has spent, say, 5 seconds calculating it will have predicted its future state at time $t = 10$ seconds ($t = 0$ being the instant the calculations begin), and after it has spent 30 seconds calculating it will have predicted its own future state at $t = 60$ seconds; and in general, at time t it will have predicted its state at time $2t$. So, suppose that we are at time t and the computer presents us with its predictions about its own state at time $2t$. What will its state at time $2t$ represent? At time $2t$ the computer will be predicting its own future at time $2 \times 2t = 4t$. Hence the predicted pattern of memory bits at time $2t$ represents the state of the computer at time $4t$. We therefore get two birds with one stone: by predicting, at t, its own state at time $2t$, the computer also has predicted its own state at time $4t$. But, of course, there is more than that, because the predicted state at time $4t$ itself actually constitutes a prediction, representing the future state at $8t$, and so on. So, it turns out that the state of the computer at time t will reveal its future states at all times $2^n t$ for all integer $n > 0$, going out to infinity. This holds no matter how small t is, the consequence being that self-prediction of any future state, even if infinitesimally close to the present, will reveal the whole temporal self-evolution infinitely far into the future. Obviously, this is impossible as it is neither possible to pack infinite information (all

future states) into a finite package (the finite number of memory bits of the computer), nor to calculate this infinite information with a finite amount of effort (the finite calculations performed during time t)[14].

But things are actually even worse. Physical self-prediction as a standalone task is not even something that is well-defined and meaningful. One can realise this by trying to devise an algorithm to perform this task, i.e. to write a computer program whose only task is to compute what it will do in the future. Are we looking for something other than the trivial idle solution, where the program does nothing and the computer's state remains the same as time passes? If so, then it seems that the problem formulation is lacking any driver to move the algorithm in any particular direction, to evolve its behaviour in some way. The task of self-prediction is not well defined at all. It is a task whose definition is based on the future, when, by nature, the future itself depends on the present. We must, in the present, write an algorithm that will predict what the computer state will be in the future, but the state of the computer in the future depends on what algorithm we are programming now. Hence the definition of the task of self-prediction is circular and ambiguous, leaving the task indeterminate. On the contrary, in the usual case of prediction of another, external system the task is determined by the initial conditions and laws that pertain to that system, in which there is (normally) no ambiguity but are well defined. If, on the other hand, self-prediction is not a standalone task but is combined with another task, such as the prediction and/or manipulation of an external system, e.g. the domino configuration, then an additional hard problem arises. If the predictor's complexity is N (number of bits or degrees of freedom) then only $N_s < N$ of these can be dedicated to self-representation for self-prediction, since a non-zero number of components ($N - N_s$) must be dedicated to the other task (prediction and manipulation of the domino configuration). Hence, the predictor must represent the totality of its N degrees of freedom with only a subset $N_p < N$ of these, which is not possible.

7. Final thoughts

We have analysed the PoP in detail in the cases that the predicted system, or counterpredictive device, is a physical system and the demon is either physical or non-physical, and we have seen that the PoP does not reveal any unexpected unpredictability or freedom inherent in determinism, but rather is due to instructing the demon to do something that is physically impossible, something that the deterministic rules do not allow. It is useful to point out that blaming the failure of the demon on the "counterpredictive" device is misleading. There is nothing special about a "counterpredictive device", as long as this refers to a physical system, compared to any other physical system. Rather, it is the

contradictory mental signification rule that is the source of the demon's failure, and relative to which the adjective "counterpredictive" applies. With the right signification rule, any physical object can be made to play the role of a counterpredictive device, be it a domino set, a pebble, a coin, a door knob, even an atom (for example, suppose that you must predict the direction of motion of an atom, and indicate your prediction by pulling the atom in the opposite direction).

That the problem has a mental, rather than physical, origin can be clearly seen if we remove mental terms from its description; "prediction of its behaviour", "indicate its prediction", "revealed to it", etc. are all "intentional stance" vocabulary, according to Dennett's terminology [26]; they attribute mentality to the actors of the PoP. If the problem is instead formulated in "physical stance" language then the mystery disappears. For example, formulated in physical stance language, the domino PoP problem merely comes down to choosing the outcome of the domino experiment from among the choices (6), when neither of these two choices describes a possible outcome.

Therefore the PoP does not offer any support to the compatibilist cause. The compatibilist believes that a human being is a physical system and perhaps hopes that the PoP proves that, even if deterministic, such a system can act freely, which was shown not to be the case. But does the PoP offer, instead, any support to the libertarian free will cause, as hoped in the beginning of this paper? The short answer is "no", as was already noted in Section 1. This hope rested on the assumption that the PoP reveals an impossibility of prediction only for minds but not for physical systems. However, since "prediction" under the conditions of the PoP is impossible even for a physical system, the fact that prediction of the behaviour of a human being under the same conditions is impossible does not imply that a human is anything more than a physical system.

But, before ending this paper, it is of benefit to consider also the case where the object of prediction is exhibiting mentality, such as a human, because this case has important additional aspects that have not been discussed. Dennett [26] thought that "intentional" and "physical" stance languages are alternative ways of speaking about the same thing, the former being more subjective and the latter more objective. If this were true, then the human case would essentially not be any different than the PoP cases that we already discussed, such as the domino case. However, in my opinion this is far from the truth and intentional stance language statements about humans are usually not reducible to physical stance language (i.e. mentality is not physically explainable). This is the problem of "intentionality" in the philosophy of mind, which cannot be discussed at length here but the reader is referred to [[23], Section 3].

Here we will simply revisit the scenario of Section 1 about the prediction of the behaviour of a human agent, and apply our newly acquired knowledge.

So, suppose again that I consider using a computer to predict, using molecular dynamics simulation, my own future behaviour, with the intention of disproving that prediction so as to prove libertarian free will. The computer has sufficient resources to perform a complete and accurate simulation of my whole body, assuming that its behaviour is determined entirely by the initial and boundary conditions and the laws of physics. The problem is that a necessary and unavoidable ingredient of my plan is that the prediction is communicated to me. But then the PoP problem arises, as the computer will have to exert some influence on me (in the form of visual or auditory signals conveying the prediction) and hence it will not be simply predicting my behaviour but also shaping it. And my brain is most likely wired as a counterpredictive device with respect to those signals and their assigned meaning, so that there is no physical input that the computer can pass to me that can, according to the signification convention (e.g. human language), match my actual future behaviour — my brain will always ensure that I behave differently than what the computer tells me.

So far there is no difference from the PoP for physical systems. Or is there? In fact, there is an important difference. When we considered only lifeless physical systems, the signification rule was something arbitrary, an extraneous element that created the illusion that the counterpredictive device somehow has a reason to frustrate any efforts to predict its behaviour. To reinforce this illusion, we described measures that could be taken such as renaming the slots S_L and S_R to S_{falls} and S_{stands}, naming the variables `"prediction"` and `"choice"` in the computer program etc. But ultimately, the physical behaviour and functionality of the device under study was completely determined by physical causes and had nothing to do with reasons and meanings, such as prediction and counterprediction, which belong in the mental realm. Indeed, the "physical stance" language is the natural language for such a system.

On the contrary, when the predicted entity is a person, then his/her counterpredictive behaviour is not at all illusory but very real; avoiding prediction is precisely what the person intends, it is the *reason* behind his/her behaviour. In this case, therefore, it is reasons and not physical causes that ultimately drive the behaviour of the person. Sure, the prediction is conveyed to the person via physical means, e.g. sound patterns in a language that the person understands, and these sound waves have a physical effect on his/her brain, initiating chemical processes there that are associated with his/her eventual behaviour (although whether they completely determine it is debatable). But it is the *meaning* that has been assigned to these sound waves that matters, not the wave pattern itself; if the person spoke a different

language then that particular sound wave pattern would not have had the effect that it has, and another pattern, which expressed the same meaning in that other language, would. And in general, whatever physical means was employed to communicate the meaning of the prediction to the person would have a counterpredictive effect ultimately by virtue of the meaning that it conveyed and not by some physical property. Since a person ultimately behaves based on reasons rather than physical causes, and since reasons, unlike physical causes, do not have determining power but merely motivational, it follows that rational agents, minds, have free will. Of course, this position is controversial, as materialists will contend that reasons are reducible to physical causes, but in my opinion this is impossible. An effort to reduce reasons to physical causes and meanings to physical events is tantamount to tryning to explain metaphysics in terms of physics, when metaphysics is the explanation and understanding of physics; it is like trying to lift oneself up by pulling their own bootstraps. Hence, in my opinion, this effort is as vain as that of ancient alchemists or geometers who sought to transmutate lead into gold or to square the circle. A full discussion of this issue is beyond the scope of the present paper and the reader is referred to [23]. It must be admitted though that this is not a conclusion that follows from the PoP per se, but rather stems from contemplation of the problem of intentionality which is surely manifest in, but not particular to, the PoP. Hence the PoP does not by itself constitute a particular weapon in the Cartesian dualist's / libertarianist's arsenal, a new hard problem for physicalism.

Finally, let us also consider self-prediction with respect to a person. If it is acknowledged that a person is a Cartesian substance with free will, then obviously self-prediction is, strictly speaking, impossible because it is precluded by free will. If, on the other hand, a person is just a physical system, as physicalism contends, then again self-prediction is impossible for the reasons discussed in Section 6. But we can explore this a little further. Returning to the aforementioned scenario, suppose that in order to avoid the boundary conditions impasse I decide not to obtain a prediction of my future behaviour using a computer, but to deduce it myself using my knowledge of my brain's structure and of the principles of physics, biology, neuroscience, etc. That is, I choose to perform the calculations mentally, in my head, instead of using a computer. If I managed to do that, it would be a case of self-prediction, with the outcome depending on the initial conditions and the laws only, without external influences. Of course, such a task is formidable and the processes in my brain are likely so complex that I could not keep track of them so as to predict the physical behaviour of my brain, let alone of my whole body. The materialists, who believe that I am but a physical system, may contend that this is precisely due to the impossibility of

physical self-prediction, discussed in Section 6: my brain tries to self-predict itself, but it doesn't have the resources to represent and outpace its own self.

However, some deeper contemplation reveals that this is not a clear-cut case of self-prediction, but there is some sort of dualism at play. In pure mental self-prediction (i.e. where a mind predicts its own self) I would be trying to think about what I will think next on purely mental terms, discarding any physical substrate of thought; and in pure physical self-prediction (i.e. where a physical system "predicts" its own self) my brain would have physical structures, e.g. neural networks, that would, according to an inexplicable mapping, be representing (although representation is something that requires a mind) the physical structure of my whole body, including their own selves, and chemical processes would be evolving these structures so as to represent the future state of my body. Of course, both self-predictions are impossible; the former due to the absence of deterministic mental laws that govern thinking[15] and the recursive nature of trying to think what new thought the current thought will produce, and the latter due to the complexity limitations discussed in Section 6. But now we instead seem to have a sort of hybrid mental-physical self-prediction where I, as a mind, try to think, to mentally deduce, the evolution of the physical structure of my body. From one perspective this could be seen as plain prediction rather than self-prediction, since the mind is so different from the body. Any obstacles to such prediction that are due to self-reference, recursion, complexity — the issues that were discussed in Section 6 — are only indirectly at play, due to the correlation between mind and body, a correlation that appears to be necessarily contingent since it is not deducible from or implied by physics or reason [23].

If mental events in the mind and physical events in the body (brain) are not completely correlated, and in particular if it turns out that I can mentally deduce how I am supposed to behave in the future according to the present physics of my body, and then decide not to behave this way, without these thoughts and decision leaving a physical footprint on my brain that frustrates my prediction, then this would indeed disprove epiphenomenalism and prove free will. However, whether or not this is the case is not something that can be logically deduced from the PoP as a philosophical argument, but requires empirical evidence. Hence the self-prediction PoP version again offers no strong support to the libertarian thesis.

Footnotes

[1] As a sidenote, it is noteworthy that this statement shows that Spinoza (naturally, in my opinion) does not consider reasons to be causes; otherwise, everyone would know the causes of their actions, since they know the reasons behind their decisions. He believes that there are other, hidden, causes in the background. This is an interesting observation in view of the fact that some philosophers count reasons among the causes of mental behaviour [27], making the case for mental determinism trivial, since we always have reasons for acting as we do, when acting consciously. However, as I argue in [[23], Section 5.1], since in every circumstance we find ourselves in we will be presented with alternative choices of action for each of which there will be reasons for and against, it is obvious that reasons do not determine but simply motivate. Which decision is eventually made (and therefore, which of the available reasons will weigh more in our minds) must be determined by other factors, the most likely of which are physical causation, if physicalism is true, or free will, if libertarianism is true.

[2] I am limiting the discussion here to physical determinism, but the same argument could be used against any sort of determinism provided that we could, even in principle, know a priori the causes of our actions and their mechanics, be they physical or otherwise (e.g. theological determinism), so as to deduce what they determine our actions to be in the future.

[3] To keep things simple, the discussion here assumes that the functionality and behaviour of the human body, from a physical perspective, are deterministic. If quantum-mechanical indeterminism does happen to affect this functionality and behaviour at the macroscopic level, and therefore needs to be taken into account in the simulations, then the simulations' output will be a range of predicted behaviours each assigned a certain probability. But this does not change the core of the argument, as one would then be able to disprove epiphenomenalism by behaving in ways that do not follow the predicted probability distribution. In this case, a single experiment would not suffice but a range of experiments would be needed where the subject chooses to always behave, say, in the least probable manner so as to disprove the predicted probability distribution (unless they can behave in a way that has zero probability, in which case a single instance of such behaviour would suffice).

[4] Laplace was not the first to note that; several others noticed this consequence of determinism at around the same time or earlier, as it was realised that it was an implication of classical mechanics.

[5] Agent-causal libertarianism is the thesis that persons are the ultimate originators of their (primarily mental, but indirectly also physical) actions. Their actions derive from their free will and not (or at least not completely) from the physical laws. In contrast, in physical determinism a person's actions are ultimately wholly determined by the past (even before that person existed) and the physical laws. If agent-causality is true, then the effects of free will, although originally manifested in the mind, will permeate into the body since many of our decisions involve physical action, and even those that do not are reflected in the physical processes of the brain (mental thoughts and brain processes are correlated). Hence in agent-causal libertarianism not all physical events that occur in our bodies are physically explainable — physical causal closure does not hold. See [[28], Chapter 10] or [[23], Section 5.2] for more details.

[6] The halting problem is the question of whether a computational algorithm will terminate in a finite number of steps or continue to infinity. Other similar computational *decision problems* are, e.g. whether an algorithm will ever output the result "0", whether it will ever return the square of one of its arithmetic inputs, etc. All such problems are known to be *undecidable* (Rice's theorem), which means that there do not exist any general algorithms that can answer them in a finite number of steps (i.e. algorithms that can receive any other algorithm and its inputs as input, analyse them in a finite number of steps, and return, say, whether or not that algorithm will ever return the value "0"). The authors of [16][12] seem to think that this result decouples predictability from determinism, proving that determinism does not imply predictability: the algorithms examined are deterministic, since their function is determined precisely by the instructions that comprise them, yet there are certain things about them that we cannot predict. But this is a special kind of predictability that is not normally expected of determinism anyway. The state of an algorithm after one, a hundred, a million or a trillion steps is precisely predictable; what may not be predictable is whether the algorithm will do some specific task over a span of potentially infinite steps. It is reasonable to expect that the situation is similar for deterministic physical systems: their initial state, the external influences, and the physical laws determine in a predictable way the state of such a system after one second, one hour, a century, or a trillion years. But whether a system will *ever* perform a certain action or exhibit a certain feature, although determined by the same factors, is potentially unpredictable in a finite amount of time. The problem is that if the system will never act in the specified manner or exhibit the feature in question then no matter how far into the future we advance our prediction we will never obtain a definite answer, since we cannot preclude the possibility that the sought action or feature will be exhibited at a future time, beyond the extent of our current prediction. For

example, consider the hypothetical scenario where humans are indeed deterministic machines, and they have managed to modify themselves so as to become immortal. Suppose that you perform molecular dynamics simulations of my whole body to find out whether I will ever perform a certain action or task, e.g. discover a general solution to the Navier-Stokes equations, or start smoking, or perform murder, or utter the word "abibliophobia". To this end, you run your software to predict my next 50 years, and according to the prediction I will not perform the said action during that time. But this does not mean that I will never perform it; thus, you continue the simulation to predict the 50 years beyond that, 100 years in total. Still, the simulation says that I will not do it. Unsatisfied, you continue the simulation up to 1000 years into the future but it is still predicted that I will not do it up to that time. What about at year 1001 though? Or at year 2000, or a million years into the future? You cannot know unless you extend your simulation to that point. If at any point during the simulation it is predicted that I will perform the said action at some instant in the future, then your task is finished and you have acquired the sought knowledge, the answer to your question. But as long as the simulation has not yet predicted that I will do it you need to carry on. And if it happens that I will never do it (which you do not know), then you will need to continue your simulation perpetually, infinitely far into the future, without ever knowning for certain that I will not do it. The case that I will never do the said action is not verifiable by predictions, since they cannot cover an infinite time span.

This is all quite interesting, but has no implication on the relationship between determinism and predictability in the usual sense, and certainly has no implication on the relationship between determinism and free will. Neither does it have anything to do with the PoP which purports to show not that infinite prediction is impossible but that it is impossible to predict certain events that will occur in finite, known time.

[7] Computational Fluid Dynamics (the simulation / prediction of fluid flow) is my area of expertise.

[8] It is actually epiphenomenalism that is incompatible with free will, of which physical determinism is a special case, despite the free will debate usually focusing on determinism. Even if the physical world is indeterministic, as quantum mechanics possibly implies, it does not follow from this that we have free will; it could be that we just have random will, if it is entirely governed by the indeterministic laws of quantum mechanics without any ultimate contribution from us as substances. Of course, usually compatibilists offer their own definitions of "free will" which are compatible with determinism, but these are merely namesakes and represent completely different concepts that entail that each human is essentially programmed to think and act as he/she does.

[9] The word "ultimately" is important, because a compatibilist may, by sleight-of-hand, claim that while it is the physics of our bodies that determines our will, we nevetheless are our bodies, and therefore "we" determine our will.

[10] An exception is Karl Popper [24][25] (he did not explicitly refer to the PoP though) who considered determinism to be synonymous with predictability, and held that since perfect predictability is impossible due to epistemic limitations (uncertaintly in the knowledge of the initial conditions and the laws) the universe would not be deterministic even if it were governed by the laws of classical physics. This is a fallacy though, if the term "determinism" has its usual meaning.

[11] The arbitrariness of mappings between signifiers and signifieds is discussed at length in [[23], Section 3].

[12] See my discussion on reasons and causes in [[23], Section 5.1].

[13] In a weak sense, the demon can predict the future because he knows what he has decided to do, and he can predict how his actions in combination with the initial state and the laws will translate into the future state of the universe. But this is more appropriately called "shaping" the future rather than "predicting" it. Strictly speaking, since his will is not determined but free, he cannot definitely predict the future because he cannot predict what he will be willing in the future. Even if he decides to act in a certain way at a future time, he could change his mind before that time arrives.

[14] Except under very special circumstances, e.g. if the computer is completely idle then its current state can be construed also as a prediction of all future states, since they will all be the same as the current one. Likewise, if the computer's state varies periodically, then every now and then it will get the prediction right, specifically at times t that are multiples of the period T, since the state at times $2^n t$ will be the same as that at time t.

[15] A physicalist may protest that our thoughts are determined by the physical processes that occur in our bodies, but even if this were so it would not, stictly speaking, be thoughts themselves that determine future thoughts but their underlying physical events. Hence, purely mental prediction, where one tries to deduce his/her own future thoughts based solely on their current thoughts without regards to any underlying physical basis, is not possible.

References

1. ^M. Silverthorne and M. J. Kisner. Spinoza: Ethics: Proved in Geometrical Order. Cambridge University Press, 2018.
2. ^P. S. Marquis de Laplace. A philosophical essay on probabilities. Dover Publications, 1951 (1814). Translated by F. W. Truscott and F. L. Emory.
3. ^D. M. MacKay. On the logical indeterminacy of a free choice. Mind, pages 31–40, 1960.
4. ^M. Scriven. An essential unpredictability in human behavior. In B. B. Wolman and E. Nagel, editors, Scientif ic psychology: principles and approaches, pages 411–425. Basic Books, 1965.
5. ^D. K. Lewis and J. S. Richardson. Scriven on human unpredictability. Philosophical Studies, 17(5):69–74,196 6.
6. a, b, cP. Landsberg and D. Evans. Free will in a mechanistic universe? The British Journal for the Philosophy of Science, 21(4):343–358, 1970.
7. ^I. Good. Free will and speed of computation. The British Journal for the Philosophy of Science, 22(1):48–50,1 971.
8. ^D. M. MacKay. Choice in a mechanistic universe: A reply to some critics. The British Journal for the Philoso phy of Science, 22(3):275–285, 1971.
9. ^D. Evans and P. Landsberg. Free will in a mechanistic universe? an extension. The British Journal for the P hilosophy of Science, 23(4):336–343, 1972.
10. a, b, cS. Rummens and S. E. Cuypers. Determinism and the paradox of predictability. Erkenntnis, 72(2):233–2 49,2010
11. ^R. Holton. From determinism to resignation; and how to stop it. In A. Clark, J. Kiverstein, and T. Vierkant, e ditors, Decomposing the Will, pages 87–100. Oxford University Press, 2013.
12. a, b, cJ. Rukavicka. Rejection of Laplace's Demon. The American Mathematical Monthly, 121(6):498498,2014.
13. a, bJ. Ismael. How physics makes us free. Oxford University Press, 2016.
14. a, b, cJ. Ismael. Determinism, counterpredictive devices, and the impossibility of laplacean intelligences. Mon ist, 102(4), 2019.
15. ^B. Garrett and J. J. Joaquin. Ismael on the paradox of predictability. Philosophia, 49(5):20812084,2021.
16. a, b, cV. Gijsbers. The paradox of predictability. Erkenntnis, pages 1–18, 2021.
17. a, bS. Rummens. The roots of the paradox of predictability: a reply to Gijsbers. Erkenntnis, pages 1–8,2022.

18. a, bC. Dorst. Laws, melodies, and the paradox of predictability. Synthese, 200(1):1-21, 2022.

19. ^L. Casalino, Z. Gaieb, J. A. Goldsmith, C. K. Hjorth, A. C. Dommer, A. M. Harbison, C. A. Fogarty, E. P. Barros, B. C. Taylor, J. S. McLellan, E. Fadda, and R. E. Amaro. Beyond shielding: The roles of glycans in the SARS-CoV-2 spike protein. ACS Central Science, 6(10):1722-1734, 2020. PMID: 33140034.

20. ^R. C. Bishop. On separating predictability and determinism. Erkenntnis, 58(2):169-188, 2003.

21. ^E. N. Lorenz. The predictability of a flow which possesses many scales of motion. Tellus, 21(3):289-307,1969.

22. ^F. Zhang, Y. Q. Sun, L. Magnusson, R. Buizza, S.-J. Lin, J.-H. Chen, and K. Emanuel. What is the predictability limit of midlatitude weather? Journal of the Atmospheric Sciences, 76(4):1077-1091,2019

23. a, b, c, d, e, f, g, hA. Syrakos. Hard problems in the philosophy of mind. Qeios, 2023. doi: 10.32388/VWPLUA.

24. a, bK. R. Popper. Indeterminism in quantum physics and in classical physics. Part I. The British Journal for the Philosophy of Science, 1(2):117-133, 1950.

25. a, bK. R. Popper. Indeterminism in quantum physics and in classical physics. Part II. The British Journal for the Philosophy of Science, 1(3):173-195, 1950.

26. a, bD. C. Dennett. True believers: The intentional strategy and why it works. In A. F. Heath, editor, Scientific Explanation: Papers Based on Herbert Spencer Lectures Given in the University of Oxford, pages 150-167. University of Massachusetts Press, 1981.

27. ^D. Davidson. Actions, reasons, and causes. Journal of Philosophy, 60(23):685, 1963.

28. ^M. McKenna and D. Pereboom. Free will: A contemporary introduction. Routledge, 2016.

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.