Free will and the paradox of predictability

Alexandros Syrakos

Department of Mechanical and Manufacturing Engineering, University of Cyprus,
P.O. Box 20537, 1678 Nicosia, Cyprus

September 24, 2024

Abstract

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 governing laws are known perfectly. This ostensible conclusion has been used to support compatibilism, the thesis that determinism is compatible with free will: supposedly, the PoP reveals that the nature of 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. 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 one's own prediction. The related issue of the impossibility of self-prediction is also discussed.

Please cite this paper as: Alexandros Syrakos (2024). Free Will and the Paradox of Predictability. *Qeios*. DOI: 10.32388/FHH7ZU.2.

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 [Silverthorne and Kisner, 2018, p. 73]).

[1] It is noteworthy that Spinoza does not consider reasons to be causes; otherwise, everyone would know the causes of their actions, since they know the reasons for their decisions. He believes that there are other, hidden causes in the background. In my opinion, this is the natural view, contrary to the suggestion of some philosophers who count reasons among the causes of mental behaviour [Davidson, 1963], making the case for mental determinism trivial since we always have reasons for acting as we do, when acting consciously. However, since in every dilemma there will be reasons for and against each possible choice, it is obvious that reasons do not determine but simply motivate. Which decision is eventually made (and therefore, which 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.

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 understood to a substantial degree: they are the fundamental laws of physics; they govern the behaviour of the countless fundamental particles that compose our bodies and the environment with which they interact. But then the following questions arise naturally: if we know the causes that determine our actions, and we know the mechanics of these causes (expressed in the equations of physics), 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 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[3] of our whole bodies will reveal what the physical laws have determined for our future (overlooking the facts that (a) the required computational effort is, by current standards, so immense that such a simulation is practically impossible, and (b) 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 the possibility of such a scenario disprove physical determinism for humans[4]? 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 readily drawn. 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, under special circumstances, to make predictions even about *inanimate* deterministic physical systems. 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 ("Laplace's demon") who knows the current state of the universe

[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.

[3] This is a kind of computer simulation where each individual atom or molecule of a system (e.g. of my body and its surroundings, in this particular case) is modelled; hence, it is an extremely detailed and accurate, albeit also extremely expensive, kind of simulation. In this and subsequent thought experiments, references to molecular dynamics simulations imply that the predictions account for all of the physics of the predicted system, at any level of detail, leaving nothing out.

[4] To keep things simple, the discussion here assumes that the human body, from a physical perspective, behaves deterministically. If quantum-mechanical indeterminism happens to affect its macroscopic behaviour, and therefore needs to be taken into account, then the simulations' output will be a range of predicted behaviours each associated with 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. A single experiment would then 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.

in all detail and the precise form of the physical laws can deduce the future down to the last detail [Marquis de Laplace, 1814, p. 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 flawed. I will analyse it shortly, but for now notice that the same could be true of the previous libertarian 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 PoP 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, only one of which is lit at a time, 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, indicating his prediction by pressing the button of the corresponding colour. But the device is wired such that pressing the green button causes the red bulb to light up, and pressing the red button causes the green bulb to light up, at the designated future time. So, it seems that despite knowing everything about the universe, including precisely how the light bulb device is wired, 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 vice versa. The frustrator in this case is a purely physical device and functions mechanistically. Hence it could be that our own ability to frustrate predictions made about us is physically 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. [MacKay, 1960, Scriven, 1965, Lewis and Richardson, 1966, Landsberg and Evans, 1970, Good, 1971, MacKay, 1971, Evans and Landsberg, 1972]. Subsequently, this debate was largely forgotten until recently, when interest in it rekindled [Rummens and Cuypers, 2010, Holton, 2013, Rukavicka, 2014, Ismael, 2016, Ismael, 2019, Garrett and Joaquin, 2021, Gijsbers, 2021, Rummens, 2022, Dorst, 2022]. In my opinion, the most accurate and clear analysis of the paradox to date is that by Landsberg and Evans [Landsberg and Evans, 1970], but unfortunately it has not received the attention it deserves. Some recent studies have misdirected their efforts by seeking explanations of the PoP in areas that are completely unrelated to its mechanics. For example, the PoP has been attributed to epistemic limitations of predictors (that they cannot know exactly the initial state or the laws) [Ismael, 2019], or to a purported flexibility of the physical laws [Ismael, 2016, Chapter 7], [Dorst, 2022] who supposedly depend 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 [Rummens and Cuypers, 2010] is mostly correct (see [Rummens, 2022] for a small correction). A couple of authors have noticed a similarity between the PoP and Turning's halting problem [Rukavicka, 2014, Gijsbers, 2021], but this does not shed light on the nature of the PoP, and hence these authors were misled to incorrect conclusions[6].

[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-causation is true, then the mind causes physical events in the body that do not follow from the laws of physics; physical causal closure does not hold [McKenna and Pereboom, 2016, Chapter 10].

[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 equivalent computational *decision problems* are, e.g. whether an algorithm will

In contrast to our original motivation of using this paradox as a means of disproving determinism and establishing the truth of libertarian free will, many of the above cited works attempted to use the paradox as a means of establishing *compatibilism* – the thesis that determinism and free will can both be true at the same time. 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 and t' are two times such that $t' < t$, then the state of the universe at time t is completely determined by its state at time t' according to a set of laws that govern it. The state of the universe at time t' is itself, in exactly the same manner, determined by the state at any prior time $t'' < t'$, 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 governing laws. Obviously determinism, by definition, is such that any definition of "free will" that is compatible with it can bear only a superficial semblance to what we normally mean by these words, since it implies that 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 PoP 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, and in particular that it actually allows some freedom and hence can be compatible with genuine 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 [Landsberg and Evans, 1970, Rummens and Cuypers, 2010]) 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 prob-

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 cannot exist any general algorithms that can answer them in a finite number of steps (i.e. algorithms that can analyse any other algorithm and its inputs in a finite number of steps and return, say, whether or not that algorithm will ever return the value "0"). The authors of [Rukavicka, 2014, Gijsbers, 2021] 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 something specific 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. For example, consider the hypothetical scenario where humans are indeed deterministic machines, and they have managed to evolve 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, then your task is finished and you have 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 knowing for certain that I will not do it. The case that I will never do the said action is not practically predictable, since the simulations 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.

lem: 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 not part of the predicted deterministic system (an *external* demon) and the case that it is part of it (an *embedded* demon)[7]. The latter case will further be shown to be impossible even if the demon is not asked to invalidate their prediction (perfect 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. On the other hand, just as the PoP 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 inherent in the deterministic laws governing the predicted system, but are external elements that are part of the mental realm and originate from a conscious observer, as noted also by [Gijsbers, 2021]. 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, predictability, and free will

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; it is part of a metaphysical mental understanding of the physical reality, it belongs in the world of meanings and is something that only a mind can do. However, we know by experience that our cognitive capacities have limitations that do not allow us to predict wholly mentally the evolution of complex physical systems; nevertheless, we can resort to the use of physical aids ranging from simple stuff such as pen and paper to highly sophisticated 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 computer memory or screen pixels' colours may denote wind velocity in a weather prediction). We program the computer so that the processes occurring therein mimic those occurring in the original system in some sense so that the evolution of the signifiers will mirror that of the signifieds. The simulated system and the simulator are both governed by the fundamental laws of physics, but since they have very different compositions we must employ considerable intelligence, ingenuity and scientific acumen to ensure that the computer processes indeed mirror those of the original system in some sense that will allow us to draw conclusions about the latter by observing the former. Hence computational science is a very complex and demanding field. It must be stressed 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 electrical charges along circuits, storing them in capacitors, etc.) has nothing to do with what the original, simulated system does (e.g., in the case of

[7] Note that these definitions are different from those used in [Rummens and Cuypers, 2010], where an external demon is one that does not interact with the predicted system (and hence cannot give rise to the PoP), and an embedded demon is one that does (and hence can give rise to the PoP).

numerical weather prediction, moving around air masses, changing their temperature and humidity, etc.). The exact same bit patterns in computer memory may represent completely different things in different applications, and it is up to us to assign meaning to them as it suits us.

Nowadays computational predictions and simulations are ubiquitous in every branch of science and engineering. They are used for weather and climate prediction, for the design of vehicles, buildings, machines, complex materials and drugs, for obtaining insight into complex physical processes, etc. Predictions and simulations have the same ingredients: 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). These ingredients are essentially the constituents of physical determinism, fully determining the evolution of a physical system in a logically comprehensible way, allowing us to predict or simulate it.

Despite the impressive recent 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 [Casalino et al., 2020], then any meaningful molecular dynamics simulation of a whole human body is out of the question for the foreseeable future, perhaps forever. There are also limitations of epistemic nature [Bishop, 2003]: The initial and boundary conditions cannot be known to perfect accuracy (except in simulations of hypothetical scenarios) and this can introduce significant errors in the results. 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 error growth depends on the equations solved, with some equations amplifying 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 [Lorenz, 1969, Zhang et al., 2019]. But even the equations solved (expressing the physical laws) are often not completely accurate, but are only approximate models of the behaviour of actual physical processes.

Do these limitations have any implications concerning the relationship between determinism and free will? To me it is obvious that they have absolutely none. 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. We, on the other hand, are obviously not independent of the laws of physics – but 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 conscious experiences of will, desire, decision, thought, etc. all of which are entirely determined by the impersonal and mechanistic laws of physics. 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 or not our will is ultimately determined *entirely* by factors beyond us; whether or not the deterministic basis of our behaviour is tractable enough to be accurately predictable by us is irrelevant. This has been acknowledged by most philosophers studying the PoP (except [Ismael, 2019]).

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 computing power, still it was logically impossible to predict the future? This would seem to indicate something about the nature of determinism itself, something ontological rather than merely epistemic. If this is the case, then determinism has been hitherto misconceived, and this may raise hope that it is compatible with (genuine) free will after all, especially since unpredictability seems intuitively to be a characteristic of

[8] Despite the free will debate usually focusing on determinism, it is actually epiphenomenalism that is incompatible with free will, of which physical determinism is only a special case. Even if determinism is false, it could be that our will is not free but just random, if the indeterminism is only of the physical, quantum-mechanical type.

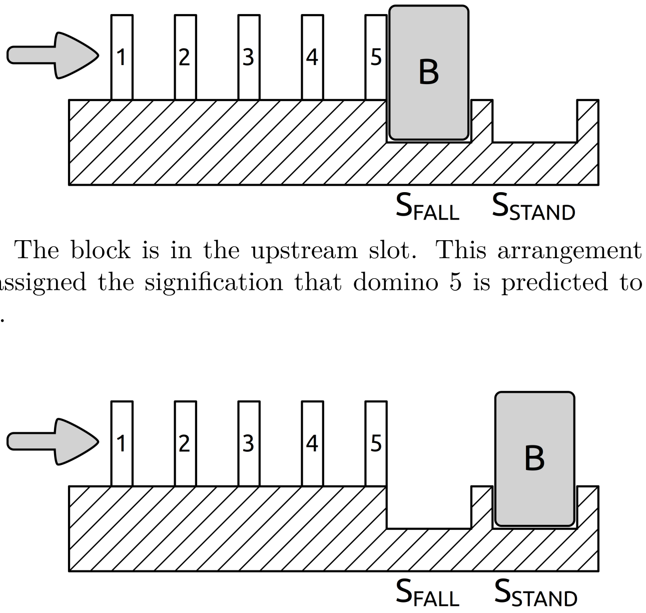
freedom. This appears to have been the main motivation for many of the philosophers who studied the PoP who, being physicalists, want to uphold determinism, but at the same time, naturally, would wish that we have free will. They mistakenly perceived the PoP, perhaps driven by hope and enthusiasm, as demonstrating that there exist completely deterministic systems whose future is impossible to predict even if their initial state and the laws are known precisely. 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 and drawing a conclusion of unpredictability of this sort from the PoP is a fallacy, as will be shown next. 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).

3 Analysis of the PoP for an external demon

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 $i = 1, 2, \dots, N$ ($N = 5$ in the setup of figure 1), a table with a flat surface on which the dominoes are placed followed by two slots S_{FALL} and S_{STAND}, and a solid block B which can fit into any of the two slots. Initially, the dominoes are placed standing in succession, close enough to each other such that if one falls it will topple the next one as well. So, a law of this universe is that if D_i falls then D_{i+1} will fall as well, for $i = 1, 2, \dots, N-2$. 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_{FALL}; but if B is located in S_{FALL} 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_{FALL} and the case that it is initially located in S_{STAND}. There are no internal factors in this universe that could cause the block B to move from its initial position; it is too heavy, and locked in place by the slots, to be moved by domino D_N falling onto it.

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 the upstream slot, S_{FALL}, then D_N will remain standing, and if B is in the downstream slot, S_{STAND}, 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_{FALL}; and if you predict D_N to remain standing, place B in S_{STAND}. With this rule, obviously it becomes impossible to make a correct prediction, because that would require that either B is in the upstream slot S_{FALL} and the last domino falls, or that B is in the downstream slot S_{STAND} 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; the paradox seems to arise artificially. Before we analyse it, 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; 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



(a) The block is in the upstream slot. This arrangement is assigned the signification that domino 5 is predicted to fall.

(b) The block is in the downstream slot. This arrangement is assigned the signification that domino 5 is predicted to remain standing.

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

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_{FALL} or in S_{STAND}) 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 (by moving the block to the appropriate slot in order to indicate his/her prediction). 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. In our case, the demon could choose between two different possible futures, where the last domino falls or remains standing, by placing the block in the downstream or upstream slot, respectively.

3. However, the demon is not free to interfere however he/she pleases but must act according to an imposed rule 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 imposition of the new rule, which must be satisfied on top of the physical laws, makes the total system of laws/rules *overdetermined*; moreover, the imposed rule is, by design, incompatible with the physical laws, making it impossible for both the rule and the laws to be satisfied at the same time. While the physical laws dictate that either B is in the upstream slot and the domino remains standing or the block is in the downstream slot and the domino falls, the imposed rule demands the

opposite. The demon is therefore not asked to do something that he/she does not know how to do (as the PoP assumes), but something that is impossible. 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 so paradoxical. The paradoxy disappears if we express the instructions to the demon in a language stripped from mental notions such as prediction and signification. Compare the following two sets of instructions:

- Since this is a deterministic, and therefore predictable, system, predict whether the last domino will stand or fall by placing the indicator block in the S_{STAND} or S_{FALL} block, respectively.

- Place the block in the appropriate slot so that either (a) the block is in the upstream slot and the last domino falls, or (b) the block is in the downstream slot and the last domino remains standing.

There is nothing paradoxical about the second instruction being impossible to carry out. It asks for the realisation of physically impossible scenarios. On the other hand, the impossibility of carrying out the first instruction, given the nature of determinism, may seem more counter-intuitive; one may thus be surprised that the task proves to be impossible, and make the mistake of identifying determinism, rather than the instruction itself, as the source of the problem. But what both instructions ask the demon to do is essentially the same thing. The difference is that the first instruction is adorned with the extraneous attribution of meaning to various physical components of the setup: the block is characterised as an "indicator" about a prediction, and similarly the names of the slots, S_{FALL} and S_{STAND}, assign to them a corresponding significance. But these meanings that we assign to these elements are not part of their intrinsic nature; they do not derive from the deterministic laws of this miniature universe. They are mere conventions that we, as conscious entities, choose to externally assign to them, and are thus not guaranteed to be consistent with the deterministic laws. In fact, in this particular case, they are inconsistent by design.

Let us see a couple more instances of the PoP in other contexts. The first is a mathematical example. Suppose that the value of x satisfies the following equation (the equation could describe the deterministic evolution of a physical system, and x could be the future value of some quantity):

$$x - 1 = 0 \tag{1}$$

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:

$$x - 1 = 0 + x_p \tag{2}$$

where x_p 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 $x = 1 + x_p$, one for each value of x_p that you choose. Nevertheless, the way that your task is communicated to you makes it appear as though it concerns the solution of (1). Furthermore, the rules require of you to communicate your solution in a way that affects the solution itself; in particular, you are required to reveal your solution by assigning it to the variable x_p (x_p is meant to be the "predicted" value of x). In other words, your task is to choose the value of x_p such that

$$x_p = x \tag{3}$$

Comparing equations (2) and (3) we can see that it is impossible to satisfy them both, whatever the values of x and x_p (if, according to (3), we substitute x for x_p 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: Originally, without tampering, prediction of the state of the last domino is straightforward (step 1), as is the solution of eq. (1). Then (step 2) the domino system is tampered with by allowing multiple outcomes depending on where the demon chooses to place the block, and similarly the equation in modified into the form (2) which has multiple solutions depending on what value the demon chooses to assign to the variable x_p. Finally (step 3), the demon's freedom is restricted to choosing among a set of impossible scenarios involving mutually exclusive physical or mathematical conditions.

As another example consider the following computer program:

```
program predict_me

print "What will I choose (true / false)?"
read prediction

if prediction == true then
    choice = false
else
    choice = true
end if

print "My choice is: ", choice

end program predict_me
```

The program asks the user to input a prediction about its output, in the form of a boolean value (true/false), and then outputs 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. The context makes it appear as if it is impossible to predict the outcome, when in reality what is impossible is to match the values of the variables "prediction" and "choice".

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 asked 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 using 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 for some time before letting it fall; a timer is started and you are asked to predict, using your knowledge of physics, the reading of the timer at the 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 timer to display "$t = 2$ seconds" at impact, then you must hold the pencil for 2 seconds after the timer was started and then let it go). But then for your prediction to be true the pencil must fall with infinite speed, something impossible.

In all the previous examples, the prediction rules that give rise to the PoP consist of heterogeneous parts joined together artificially and unnaturally. 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) 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. Hence, there is nothing precluding that the signifying action and the signified predicted future state are physically incompatible. It is noteworthy that such a mapping can exist only within 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 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 bulb. In the domino example we named the slots S_{FALL} and S_{STAND} alluding to their significance. In the mathematical example the prediction variable is named x_p, a deliberate resemblance to x. And in the computer program example we chose referential 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 the upstream slot and D_N falls} and {B is in the downstream slot and D_N remains upright}, neither of which is possible. In the mathematics example it amounts to finding values for x and x_p 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 execution 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. The predictive significance of elements of these systems is a non-inherent 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: the "rules of the game" of prediction are rigged so as to, by exploiting the deterministic rules that govern the system, always require of the demon to physically invalidate his own prediction.

4 Analysis of the PoP for an embedded demon

Up to this point it was assumed that the demon was an outsider, not part of the deterministic system under study. Now we will consider the case that the demon is part of the system he is tasked with predicting. This has the advantage that there are no external influences on our extended system; it can be 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. Part of the demon's task is now to predict his own behaviour, since he is part of the predicted system.

Where did the sense of paradoxy arise from in the first place? Let us return to Laplace's vision, which encompasses the entire universe. If by "universe" we mean everything in existence, including the demon, then there are no external influences (since there exists nothing outside of the universe), and if that universe is deterministic then its initial state together with its laws determine all its future states. Therefore, if the demon can indeed predict the future based on the past and the laws, he should be able to foresee everything, including the existence of the counterpredictive device, how it works, his own existence and actions, the fact that he will be asked to predict the frustrator's behaviour, and the outcome of this prediction. So, on one hand we have a perfectly predictive demon, and on the other hand a perfectly counterpredictive device. Something has to give; both cannot coexist. And since it is seemingly easy to imagine a counterpredictive device, the obvious inference is that a perfectly predictive demon cannot exist. Hence, proponents of compatibilism have rushed to the conclusion that determinism is ontologically unpredictable, and that this is because determinism is not fully binding but allows some freedom.

This chain of reasoning is flawed at several places, and it will be analysed in what follows. Essentially, this scenario is as inconsistent as the previous one about an external demon, only that the problem is manifested in the initial conditions rather than in the boundary conditions. It is instructive to analyse the embedded demon case separately for at least two reasons.

Firstly, consider again the libertarian philosopher who came up with the idea of proving that we have a free will, transcending of the deterministic physical laws, by using a computer to simulate and predict his future behaviour and then deciding to act differently than what was predicted. In this scenario, the philosopher is the predicted system and the computer is the (external) demon. After some contemplation, the philosopher realises that his idea will not work, because there is no way for the computer to communicate its prediction to him without invalidating it, as was explained in the analysis of Sec. 3. But then it dawns upon him to circumvent this problem by discarding the computer and making his own prediction of himself in his head, mentally, from his knowledge of the physical structure of his body and the laws, without external aids (let us overlook the insurmountable difficulty of such a task); in this way he will receive no external stimuli that may invalidate the prediction. Would this succeed in proving that humans have free will? The answer is no because, just like for the external demon case, the experiment does not seem to require a conscious being but it could be performed with an inanimate system. Indeed, instead of discarding the computer we could instead discard the philosopher, and program the computer to predict its own future self and to subsequently act contrary to its own prediction. Hence the PoP can arise whether or not the embedded demon is a conscious being or not. We will see precisely where such an experiment would stumble in the analysis that follows.

Secondly, the embedded demon case is even "more" impossible, so to speak, than the external demon case because, as it turns out, physical self-prediction is in and of itself an impossible task due to complexity considerations, even it is not accompanied by counterpredictive requirements. That is, a computer cannot predict its own future state, let alone both predict it and invalidate the prediction. This impossibility of self-prediction is not something particular to the PoP, but holds in general. Why this is so will be explained in Sec. 5.

4.1 Demon with free will

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. Our main investigation will focus on the case that the demon is part of the deterministic system whose future state he is tasked with predicting, and hence his own behaviour is also deterministic, governed by the same laws as the rest of the system. But before we delve into this, it is worth considering briefly the case that dualism is true and the demon is an immaterial, Cartesian being with free will. In that case, the demon is not governed by deterministic rules but has freedom of choice, and he/she thinks, decides and acts based on reasons rather than causes; the demon's behaviour is indeterminate and therefore unpredictable. Then, going back to the domino example, putting aside the signification requirement, he knows that if he places the block in slot S_{FALL} then the domino D_N will remain standing and if he places it in slot S_{STAND} 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. 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 with complete certainty[9] 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

[9] 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. Strictly speaking, though, since his will is not determined but free, he can always change his mind about what to do and hence he cannot be completely certain about his own future behaviour.

the universe.

4.2 Deterministic demon

Let us now turn to the case that the demon is a physical (deterministic) system, part of the physical 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 included? And 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, a computing machine that predicts its own future state but is also programmed to act as a counterpredictive device, contrary to its own predictions. In fact, there is no substantial difference between these two cases; if the demon is a complex physical object then whether or not the counterpredictive device or anything else is part of it is a matter of convention, of how we define the boundaries of the physical system that is the demon.

For simplicity and convenience, in the discussion that follows the demon will be considered as some sort of computing device, but if physicalism is true then the same conclusions will hold also for a conscious demon, such as a human being. In physicalism, whether reductive or non-reductive, the higher-level, mental properties and events related to a consciousness supervene on the physical structures and processes that occur in the body/brain that gives rise to that consciousness. According to physical causal closure, these physical processes are entirely determined by the laws of physics, which, in a deterministic universe, set a definite, predictable path forward for the evolution of a physical system such as the brain. Since the mental state of that consciousness is supervenient on the physical state of the brain, by predicting the future physical state of the brain according to its physical structure and the laws of physics one predicts also the future mental state of that consciousness. Hence, from a physicalist perspective, there is no loss of generality if we focus only on the physical processes, regarding the demon as a computing machine, and neglect the derivative mental processes that would emerge in the case of a conscious demon.

To proceed, it is noted 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[10]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 $t_+ > t_0$, where t_0 is the time instant when the domino wave reaches D_N, to move the block B to the appropriate slot S_{FALL} or S_{STAND} according to the signification convention. The moving of the block is to take place at a certain time $t_- < t_0$ (and hence will be simulated as well). There are no external influences on this system, and therefore the computer can predict the future from only its initial state and the laws. Since everything is deterministic, whether D_N will be standing or fallen and whether B will be in S_{FALL} or S_{STAND} at t_+ 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 signification rule is still incompatible with the physical laws. But this version of the PoP may perhaps seem more paradoxical than the one where the demon was an outsider, because at first glance the procedure seems straightforward and it is not obvious where it will fail. For given initial locations and velocities of all molecules in the room, including those of the computer, there is a single, definite, determined outcome concerning whether the domino D_N will fall or remain standing at time t_0; that the computer has been programmed to move the block B according to its prediction does not change this fact. Using molecular dynamics software, the computer can compute precisely the motions of all atoms

[10] See footnote 3.

according to the laws of physics and obtain unequivocally the future state. 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 for the external demon case. 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 satisfy this interdependence both ways because of the self-contradicting signifier-signified rule.

To better 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 (an assumption that will subsequently be shown to be false); all the components of the system, and therefore their molecular structures, are clearly defined. If this is so, then the laws dictate a specific, predictable path towards the future, which the computer can predict. Simulation software normally use time-marching algorithms to 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 $t < t_-$ and is now about to predict the state of the system at time instant t_-. At t_-, its future self will move the block B to the appropriate slot according to the signification rule, and in order to do so it will use its own prediction of the state of the domino D_N at time t_+. However, the computer's present self has advanced its own prediction only up to time t_- and hence does not yet know whether D_N at time t_+ will fall or remain standing; therefore, it is unable to deduce whether its future self at time t_- will move block B to slot S_{FALL} or to slot S_{STAND}. As a result, it is unable to advance its prediction beyond time t_-.

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, i.e. how it is programmed, which algorithm it is to execute. This information is embedded in the initial conditions, because the algorithm is physically implemented as a series of bits in the computer memory, and the computer memory consists of molecules whose locations and states are part of the initial conditions. In order to set the initial conditions for the simulation we therefore first need to decide on the algorithm that the computer is to execute, an algorithm that can bring the combined tasks of prediction and signification to successful completion, and the algorithm just described is not up to the task.

One may try to overcome the problem by iteration (a common technique in computational science): program the computer to start its calculations by assuming a future state at t_+, and then repeat the prediction again and again, correcting this assumption according to the results of the previous iteration, until successive predictions are indistinguishable (i.e., in technical language, until the iterations have converged). So, since our calculations have got stuck at t_- because they do not yet know what will happen at t_+, let us begin by assuming 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_-, move the block B to the upstream slot S_{FALL}, according to the signification rules. Continuing the calculations from that point on, when time t_+ is computed, it will be predicted that the domino D_N will actually remain standing, since B is in the upstream slot 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

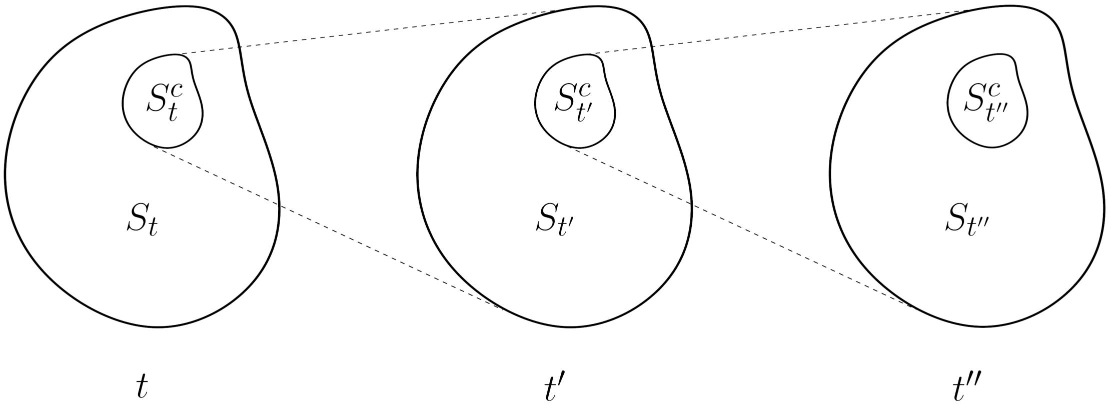


Figure 2: Ideal scenario for the prediction of a system by an embedded demon. The states S_t, $S_{t'}$ and $S_{t''}$ of the system at times $t < t' < t''$ are depicted. Subset S_t^c of S_t (the state of the demon at t) represents the future state $S_{t'}$ of the whole system. Similarly, the subset $S_{t'}^c$ of $S_{t'}$ represents the future state $S_{t''}$ of the whole system.

that the domino will fall, and further 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 simultaneously. This is the most expensive method, since the computer must have enough memory to store all the states of the system at all times $t \in [0, t_+]$ (it may have already occurred to the reader that this is impossible, since the computer itself is part of the simulated system and therefore it must store now in its memory all future contents of that same memory – obviously there cannot be enough storage space available; let us overlook this for the moment, but we will return to it in Sec. 5). However, if the problem has a solution, this procedure will find it.

The state $\mathcal{S}_t$ 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 $\mathcal{L}$:

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

Part of the initial state $\mathcal{S}_0$ of the system is the initial state of the computer (which is a part of the system), which determines its design and functionality, how it is built and programmed. Let $\mathcal{S}_t^c \subset \mathcal{S}_t$ be the part of $\mathcal{S}_t$ that describes the computer. The computer performs its predictions physically, and 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 $\mathcal{S}_t^c$, embedded in $\mathcal{S}_t = \mathcal{L}(\mathcal{S}_0, t)$, which can be interpreted (by a mind under some representation convention) as representing the state $\mathcal{S}_{t'} = \mathcal{L}(\mathcal{S}_0, t')$ of the whole system at a later time $t' > t$ (Fig. 2). At time t_- the state of the system is $\mathcal{S}_{t_-} = \mathcal{L}(\mathcal{S}_0, t_-)$, and its subset $\mathcal{S}_{t_-}^c$, which is the state of the computer at time t_-, must represent the state of the whole system at time t_+, $\mathcal{S}_{t_+} = \mathcal{L}(\mathcal{S}_0, t_+)$, which includes whether the last domino is fallen or standing.

So, let us think how we could design $\mathcal{S}_0^c$ (i.e. how we could program the computer) such that the required tasks are completed successfully. 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_+ the state of the system, $\mathcal{S}_{t_+}$, will necessarily exhibit exactly one of the following:

$$(\ B \text{ is in } S_{\text{STAND}} \text{ and } D_N \text{ falls }) \quad \text{or} \quad (\ B \text{ is in } S_{\text{FALL}} \text{ and } D_N \text{ stands }) \tag{5}$$

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

In addition to equation (4) which is imposed by the laws of nature and from which (5) derives, we want $\mathcal{S}_{t_+}$ to satisfy the signifier-signified convention, imposed by us, which requires that at time t_+ the state of the system, $\mathcal{S}_{t_+}$, exhibits exactly one of the following:

$$(\ B \text{ is in } S_{\text{STAND}} \text{ and } D_N \text{ stands }) \quad \text{or} \quad (\ B \text{ is in } S_{\text{FALL}} \text{ and } D_N \text{ falls }) \tag{6}$$

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 include any of the outcomes (6). Hence it is apparent that no matter how we program the computer, or how powerful it is, it cannot perform the task that we want it to.

To summarise, 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 and simulation software, 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 why 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 around this problem and they all fail, due to the incompatibility of the signification rule with the physical laws. 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)).

5 Self-prediction

But, it turns out that the embedded demon scenario is even more problematic than the one where the demon is external. The reason is that the computer (demon) 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 [Popper, 1950a, Popper, 1950b].

First of all, it should be reminded that, strictly speaking, the physical processes occurring within the computer, such as the transfer of electrons from one place to another, do not constitute a prediction in the full sense. They are *interpreted* as a prediction by a conscious mind, who assigns this meaning to them. That mind is who, using the computer as an aid, actually predicts in the full sense, as it has the metaphysical capacity to understand the physical reality and see meaning in it. For the computer to be useful as a prediction *aid*, it must (a) have an internal structure that can be construed (by a mind) as *representing* the structure of the predicted system (representation is again a mental notion, it is not inherent in the computer structure but is invented by a mind who interprets the computer state 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 processes they mimic.

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 memory, processor, and circuitry 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 memory 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 links it to atmospheric processes; the link 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 vast number of air molecules. Therefore, in practice we conceptually 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 all the aspects of the mechanics/dynamics of the predicted system that are of interest to us within acceptable error bounds.

Now let us consider a computer that must be programmed to predict its own future state. To accomplish this, a mapping must be established between the components of its future self and the components of its present self. The current memory of the computer must store a representation of the future memory's contents (the data), as well as the instructions of how to act on this data to further advance the prediction (the algorithm). The future memory in turn will contain a similar representation of an even later memory. Now, a full molecular dynamics simulation of its own self is out of the question for the computer, since its smallest storage units are bits, 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. A single bit of current memory can store the representation of only one bit of future memory, and no more. If the bit is the smallest unit of current memory that can hold essential and indispensable information, the same must be true for future memory. Therefore, there must be a one-to-one correspondence between the bits of current and future memory in terms of representation.

An obvious possible mapping 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 memories' states now and in the future are exactly the same. This map is applicable only in the trivial case that the computer state does not evolve in time, 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 $a{\cdot}t$ then for a prediction we must have $a > 1$ whereas in this trivial case we have $a = 1$. Clearly then, this mapping is not satisfactory.

But is a better general fixed mapping rule possible? It seems not. First of all, at time $t = 0$, when the prediction begins, the computer must represent its current self, i.e. representation and represented are necessarily one and the same: the current bit pattern construed as a representation represents its own self. Therefore, at $t = 0$ each bit must represent its own self by necessity, according to the aforementioned trivial mapping. Then, for $t > 0$, since the task of prediction requires that the predictor can advance its prediction faster than the predicted system evolves, the computer (which is both the predictor and the predicted) must advance its internal representation of its future state at a pace that is faster than that at which the representation is actually advancing, which is impossible. Its current bit pattern will represent, simultaneously, both its current state (through the identity map) and an evolving and diverging future state[11]. 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 its own self. At best, both the internal representation of the predicted state and the actual current state progress at the same speed, hand-in-hand (the aforementioned trivial representational rule).

Consider the repercussions if self-prediction was indeed possible. A prediction is, by definition, performed faster than the predicted system evolves. So, suppose that the computer predicts its own state 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 has also 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 only one millisecond into the future, will reveal the whole temporal evolution of the self infinitely far into the future (Fig. 2). 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) – except in the trivial case where the computer is idle and all states are the same.

Additionally, if self-prediction is not a standalone task but is combined with another task, such as the prediction and/or manipulation of an external system (as in the domino example) then an additional hard problem arises. If the predictor has N bits or degrees of freedom and N_s of these must be dedicated to the additional task, then this leaves $N - N_s < N$ bits for self-representation, whereas N are required.

But things are actually even worse. Physical self-prediction as a standalone task is not even something well-defined and meaningful. This can be seen by trying to devise an algorithm to perform this task, i.e. to design a computer algorithm whose only task is to compute what its own self will do in the future. Are we looking for something other than the trivial idle solution, where the algorithm does nothing? 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; "predict your own self" is not something that can translate into a concrete set of instructions, it does not even have a starting point. 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,

[11] This would be analogous, for example, to devising an algorithm whereby the same bit pattern in the memory of a weather-forecasting computer always represents both the current weather and the weather at five days later.

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 the predicted system, which are well defined and independent of the predictive system.

6 Final thoughts

Our detailed analysis of the PoP has shown that it 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 laws do not allow. It should be pointed out that blaming the failure of the demon on the “counterpredictive” device is misleading. There is nothing special about a “counterpredictive” device compared to any other object. Rather, it is the contradictory character of the mental signification rule relative to the physical laws that is the source of the demon’s failure, and it is relative to the signification rule that 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 [Dennett, 1981]; 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 of minds but not of 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 under these conditions is impossible does not imply that a human is something more than a physical system.

But before ending this paper it is of benefit to consider a bit further the case that the object of prediction is exhibiting mentality, such as a human, because this case has an important additional aspect that has not yet been discussed. Dennett [Dennett, 1981] 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 is true, then the human case is essentially no different than the PoP cases that we already discussed, such as the domino case. However, in my opinion this is not the case and intentional stance language statements about humans are usually not reducible to physical stance language (i.e. the concept of meaning is not physically explainable) – this is the problem of “intentionality” in the philosophy of mind, which cannot be discussed at length here[12]. 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, through atomistic simulations, my

[12] The interested reader is referred to Section 3 of my unpublished work [Syrakos, 2023].

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 its behaviour to be 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, since my intention is to disprove the prediction, and assuming that this intention of mine is correlated to the physical state of my brain, my brain is wired as a counterpredictive device with respect to those signals. Hence, there is no physical signal that the computer can pass to me that can, interpreted according to the signification convention (e.g. the English language, matching physical sounds to meanings), match my actual future behaviour – my brain will always ensure that I behave differently than what the computer forecasts to 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 or intention to frustrate any efforts to predict its behaviour. This illusion was reinforced by assigning deceptive names to the physical components to which signification was assigned, such as S_{FALL} and S_{STAND} to the domino slots, "`prediction`" and "`choice`" to the variables of the computer program, etc., so as to make it seem as if the signification rule is not a mere convention but something natural, inherent in these components, which has meaning even for the counterpredictive device which avoids being predicted because it likes to be free, so to speak. But in reality, 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; in a different language the sound wave pattern would be different, but it would have the same effect on the person by virtue of conveying the same meaning, provided that the person understands that language. 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 trying 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.

A full discussion of this issue is beyond the scope of the present paper (see [Syrakos, 2023]). 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 / libertarian's arsenal; it is not a new hard problem for physicalism.

Finally, let us also consider self-prediction with respect to a person. If a person is a Cartesian substance with free will, then obviously precise 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 5. But we can explore this a little further. Returning to the aforementioned scenario where I want to prove libertarianism by falsifying a prediction made about me, suppose that in order to avoid the boundary conditions impasse I decide not to obtain the prediction of my future behaviour using a computer, but to perform it myself, in my thought, using my knowledge of my body'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 too complex for me to keep track of. A materialist, who believes that I am but a physical system, may contend that this is precisely due to the impossibility of physical self-prediction, discussed in Section 5: 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 suggests that this is not a clear-cut case of self-prediction, but there is some sort of dualism at play. It is neither a purely mental nor a purely physical kind of prediction. In pure mental self-prediction (i.e. where a mind predicts its own future mental state) 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, like the computer of section 5, have neural structures that would, according to an inexplicable mapping, be representing (although representation is something meaningful only to a mind) their own future selves, and chemical processes would be evolving these structures so as to represent future states. Of course, both self-predictions are impossible; the former due to the absence of deterministic mental laws that govern thinking[13] 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 5. But now we instead seem to have a sort of hybrid mental-physical self-prediction where I, as a mind, try 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, where the mind is an external demon mentally predicting the physical evolution of the body. Any obstacles to such prediction that are due to tampering, self-reference, recursion, complexity – the PoP issues that were discussed in the previous sections – are only indirectly at play, due to the correlation between mind and body. In my opinion, this correlation is necessarily contingent, not deducible from or implied by physics or reason – but this is a topic of debate, the "problem of intentionality" in the philosophy of mind (more appropriately called the "problem of meaning" in my opinion). It is a vastly more important topic than the PoP, and a contribution of the present paper is that it highlights aspects of the PoP that are related to intentionality or meaning.

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 mental self-prediction version of the PoP can again offer

[13] 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, strictly 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 directly on their current thoughts and their reasoning without regards to any underlying physical basis, is not possible.

no strong support to the libertarian thesis.

References

[Bishop, 2003] Bishop, R. C. (2003). On separating predictability and determinism. *Erkenntnis*, 58(2):169–188.

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

[Davidson, 1963] Davidson, D. (1963). Actions, reasons, and causes. *Journal of Philosophy*, 60(23):685.

[Dennett, 1981] Dennett, D. C. (1981). True believers: The intentional strategy and why it works. In Heath, A. F., editor, *Scientific Explanation: Papers Based on Herbert Spencer Lectures Given in the University of Oxford*, pages 150–167. University of Massachusetts Press.

[Dorst, 2022] Dorst, C. (2022). Laws, melodies, and the paradox of predictability. *Synthese*, 200(1):1–21.

[Evans and Landsberg, 1972] Evans, D. and Landsberg, P. (1972). Free will in a mechanistic universe? an extension. *The British Journal for the Philosophy of Science*, 23(4):336–343.

[Garrett and Joaquin, 2021] Garrett, B. and Joaquin, J. J. (2021). Ismael on the paradox of predictability. *Philosophia*, 49(5):2081–2084.

[Gijsbers, 2021] Gijsbers, V. (2021). The paradox of predictability. *Erkenntnis*, pages 1–18.

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

[Holton, 2013] Holton, R. (2013). From determinism to resignation; and how to stop it. In Clark, A., Kiverstein, J., and Vierkant, T., editors, *Decomposing the Will*, pages 87–100. Oxford University Press.

[Ismael, 2016] Ismael, J. (2016). *How physics makes us free.* Oxford University Press.

[Ismael, 2019] Ismael, J. (2019). Determinism, counterpredictive devices, and the impossibility of laplacean intelligences. *Monist*, 102(4).

[Landsberg and Evans, 1970] Landsberg, P. and Evans, D. (1970). Free will in a mechanistic universe? *The British Journal for the Philosophy of Science*, 21(4):343–358.

[Lewis and Richardson, 1966] Lewis, D. K. and Richardson, J. S. (1966). Scriven on human unpredictability. *Philosophical Studies*, 17(5):69–74.

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

[MacKay, 1960] MacKay, D. M. (1960). On the logical indeterminacy of a free choice. *Mind*, pages 31–40.

[MacKay, 1971] MacKay, D. M. (1971). Choice in a mechanistic universe: A reply to some critics. *The British Journal for the Philosophy of Science*, 22(3):275–285.

[Marquis de Laplace, 1814] Marquis de Laplace, P. S. (1951 (1814)). *A philosophical essay on probabilities.* Dover Publications. Translated by F. W. Truscott and F. L. Emory.

[McKenna and Pereboom, 2016] McKenna, M. and Pereboom, D. (2016). *Free will: A contemporary introduction.* Routledge.

[Popper, 1950a] Popper, K. R. (1950a). Indeterminism in quantum physics and in classical physics. Part I. *The British Journal for the Philosophy of Science*, 1(2):117–133.

[Popper, 1950b] Popper, K. R. (1950b). Indeterminism in quantum physics and in classical physics. Part II. *The British Journal for the Philosophy of Science*, 1(3):173–195.

[Rukavicka, 2014] Rukavicka, J. (2014). Rejection of Laplace's Demon. *The American Mathematical Monthly*, 121(6):498–498.

[Rummens, 2022] Rummens, S. (2022). The roots of the paradox of predictability: a reply to Gijsbers. *Erkenntnis*, pages 1–8.

[Rummens and Cuypers, 2010] Rummens, S. and Cuypers, S. E. (2010). Determinism and the paradox of predictability. *Erkenntnis*, 72(2):233–249.

[Scriven, 1965] Scriven, M. (1965). An essential unpredictability in human behavior. In Wolman, B. B. and Nagel, E., editors, *Scientific psychology: principles and approaches*, pages 411–425. Basic Books.

[Silverthorne and Kisner, 2018] Silverthorne, M. and Kisner, M. J. (2018). *Spinoza: Ethics: Proved in Geometrical Order.* Cambridge University Press.

[Syrakos, 2023] Syrakos, A. (2023). Hard problems in the philosophy of mind, version 1. Queios. doi: 10.32388/VWPLUA.

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