One Approach to the Necessary Conditions of Free Will
Logical Paradox and the Essential Unpredictability of Physical Agents

Dr. Izolda Takács

Eötvös Loránd University (ELTE), Budapest
e-mail: izolda.t@hotmail.com

Even today, we lack a precise definition of free will and continue to rely primarily on intuitions about what it might entail. This paper therefore takes a negative approach to the problem. It introduces a dramatic scenario – scientific determinism – in which free will could not possibly exist, and then seeks to refute this view by exposing a logical contradiction: the paradox of predictability. If scientific determinism necessarily entails a reality in which free will is impossible, then refuting scientific determinism is a necessary condition for the possibility of free will. The paradox shows that self-prediction ($P = C$) by a physical agent (P) is objectively impossible. That is, even an agent governed entirely by deterministic processes cannot predict its own future states – not even in an abstract, Platonic sense.

Keywords: free will, necessary condition, paradox of predictability, Gödel, Turing, logical paradox

1. Introduction

If the doctrine of scientific determinism were true – that is, if every single act were part of an essentially mechanistic, coherent and determinate structure in space-time – it would be the greatest obstacle to free will. In this scenario an omniscient predictor P (later christened Laplace's demon) could exist in principle; for example a supercomputer – that, given all initial conditions and knowing the positions, masses, velocities of all the particles of the universe together with the laws of physics – could compute each future event in advance with any desired degree of precision, based on a deductive-nomological model (Popper, 1950a; 1995). Under such conditions, all events – including the lines of this very paper – would have been determined and predictable since the formation of the first quarks, from the point when the force-carrying bosons (gauge bosons) compelled the quarks to interact.

In this article, I present a logical argument (a paradox) that demonstrates why an omniscient predictor P cannot exist, even in a Platonic sense. From this it follows that scientific determinism is falsifiable not primarily because of quantum mechanics, but because it is internally self-contradictory; hence the logical paradox constitutes a far stronger argument. The paradox of predictability shows that the physical agent's prediction (P) of itself ($P = C$) is objectively impossible.

The argument itself was first introduced in Karl Popper's early writings (*Indeterminism in Quantum Physics and in Classical Physics I-II,* 1950), and in subsequent studies that reflected on his articles in the context of free will (Scriven 1965; MacKay 1967). Popper derived the impossibility of self-prediction from Kurt Gödel's first incompleteness theorem; so that his argument depends on whether Gödel's theorems on which he based his reasoning are true. If they are true, and can be correctly applied to scientific determinism – against the Laplace demon – then Popper's argument is correct and scientific determinism is only *prima facie* deterministic.[1] *Nota bene*: according to Popper, the impossibility of self-prediction does not merely show – as most authors suggest – that determinism is simply not equivalent to complete predictability, or that, even if determinism does obtain, there is an objective explanation for our sense of freedom; rather, it suggests that the universe as a whole is indeterministic. In summary, Popper argues that there are two obstacles to the self-prediction of physical systems: a physical obstacle and a *Gödel-type* logical obstacle. Because of these limitations, classical mechanics exhibits an indeterminacy – essentially similar to that found in quantum mechanics – which, although not sufficient, is nevertheless a necessary condition for free will.

Popper's central argument has resurfaced in contemporary philosophy and has sparked an ongoing polemic. Thus, after outlining the paradox, the second part of this paper presents the current debate via two approaches: the arguments of Rummens & Cuypers (2010) and of Gijsbers (2023). Their articles – *Determinism and the Paradox of Predictability* and *The Paradox of Predictability* – trace two fundamental routes to the paradox of predictability. One view (Rummens & Cuypers 2010) holds that any computational process in the physical world – usually that produces a prediction – can, for principled (yet physical) reasons, reflect on itself only by influencing the very computation, thereby causing the prediction to turn out false. The other, more general approach (Gijsbers 2023) is purely logical. According to Gijsbers, the paradox of predictability is inseparable from the impossibility of self-prediction and based on *substantive or in-principle* unpredictability[2], since it follows logically from Alan Turing's theorem on the halting problem. Thus, both ideas take it as given that a paradox exists with respect to scientific determinism; what is at issue is its source – namely, whether the underlying unpredictability is substantive or non-substantive. This paper argues that the second argument,

[1]I address this question along the lines of Gijsbers' thesis on the undecidability problem (Turing's halting problem).
[2]This paper uses the term "substantive unpredictability" to refer to what is also called intrinsic or in-principle unpredictability – unpredictability arising from structural or logical constraints..

i.e., the "logical" argument, is the decisive one.[3]

2. Main concepts, the topic in general

The notion of scientific determinism, as outlined in the introduction, is synonymous with predictability. Predictability here means that all events, in principle, can be accurately computed by scientific methods. Formally, scientific determinism obtains if and only if the following conditions hold:

(1) Systems of a class *C* behave deterministically, then (2) there exists a set *L* of deterministic laws for systems *C*, such that (3) for any event *E* in system *C*, there would exist a set $S_1 ... S_n$ that describe deterministically sufficient preconditions for *E*. (4) Furthermore, based on $S_1 ... S_n$ and *L*, one could, in principle deductively predict the event *E* exactly in advance. (5) And similarly, any event could in principle be calculated exactly, in advance (Boyd, 1972, p. 431).

In sum, on the nomological model, scientific determinism is equivalent to predictability. I believe one should take determinism itself to be this very concept. Not least because quantum mechanics is defined in explicit opposition to full predictability. More precisely, as Popper notes, because "Heisenberg's argument against determinism is based upon the implicit assumption that determinism entails predictability from within, with any desired degree of precision" (Popper 1995, p. 36). It is therefore reasonable to start from that assumption, along with the fact, that scientific determinism is logically much stronger doctrine than determinism itself. The latter, after all, leads to a trivially true, symmetrical argument (Kukla, 1978, 1980; Holton, 2013).[4]

Thus, if scientific determinism holds, then free will – whose concept, intuitively, always denotes an act that cannot be predicted (Kant, B 578), i.e. one that no predictor could compute in advance (necessary condition) even in theory – is only an illusion, an epiphenomenon. We therefore feel ourselves free merely because, in one way or another, we cannot calculate exactly

[3]I would like to thank László E. Szabó, whose valuable feedback as an opponent greatly contributed to further clarifying the arguments in this paper. I am also grateful to Daniel Kodaj for his excellent insights during our consultations and discussions on the topic, and for all of his critical comments.

[4]If the universe never repeats itself, then any class of events A has an effect B, where B is defined as the class of events that occur when the universe has a unique property and follow the occurrence of A within some time interval d. Since the universe is unlikely to actually repeat itself, these remain very weak arguments. Moreover, the doctrine of determinists does not impose any constraint on the possible sequence of events in the world (Kukla 1978, 143).

what will happen in the future; our knowledge is limited when compared with that of Laplace's demon, the omniscient super-intelligence. Hence the subjective sense of freedom may arise from the objective fact (E. Szabó 2002; Grünbaum 1972) that our grasp of the facts at any given moment is severely restricted.

2.1 Why challenge scientific determinism in the age of quantum mechanics?

While the "Newtonian scientific worldview" outlined above – scientific determinism – had to be accepted by thinkers in the pre-Heisenberg era because it was regarded as synonymous with scientific validity (Popper 1995, 47), we live in an era in which quantum mechanics has replaced that worldview. Yet it should be noted, that even the indeterminism "guaranteed" by quantum mechanics still fails to provide a sufficient account of free will. First, within bare probability we can locate no genuine *will*: if predestination deprives us of free will, sheer randomness does so as well. Second, in the field of quantum mechanics, there are certain limits to our total cognitive capacity (E. Szabó 2002; Takács 2013), aspects of which we cannot fully grasp, and this limits our epistemic access.

Although free will has not been salvaged by quantum mechanics, its empirical results could nevertheless be used to falsify scientific determinism. After all, the experimental evidence for quantum mechanics is exceptionally strong: observations and measurements indicate that the behaviour of particles is fundamentally uncertain and probabilistic (Takács 2013), thereby challenging a deterministic view of the universe. However – as this paper seeks to show – it is not only the indeterminism revealed by quantum mechanics that explains why scientific determinism is false. A separate line of argument demonstrates, on a more fundamental level, that scientific determinism cannot hold.

3. The argument against scientific determinism: the paradox of predictability

3.1. The paradox

The authors cited in the Introduction (Scriven 1965; MacKay 1967; Rummens & Cuypers 2010; Gijsbers 2021) uphold the fundamental thesis that the paradox arises because, even if we assume that in a deterministic universe *U* there exists an omniscient physical predictor *P* capable of predicting every future decision of another physical system *C* with any desired

precision, it can still be shown – by revealing *P*'s prediction to *C* – that this very disclosure may prompt *C* to change its decision, thus rendering the prediction invalid.

According to Scriven's central argument (Scriven, 1965), if a person or robot is motivated to counter-predict, then even an omniscient predictor cannot compute their decision, and its prediction will always prove false. This is because, once an individual or system learns – or replicates – the prediction made about itself, it can act to invalidate that prediction. Thus, in such circumstances, the failure of the prediction is guaranteed by what is called "counter-predictivity." For example, suppose a friend predicts that I will order paella at a restaurant and tells me so. Armed with this information – and intent on disproving their prediction – I will choose something else instead. It is precisely this act – the *revelation* of *P*'s prediction and *C*'s capacity to refute any prediction made about it (the *counter-predictive mechanism*) – that ultimately makes it impossible for *P* to issue an accurate prediction.

What exactly is the paradox?

To easily imagine what the paradox, or the counter-predictive mechanism, consists of, let us suppose a superintelligent predictor – let us call it SIP9000. In front of it stands a machine (a simple box-shaped device), atop which sit two light bulbs – a red one and a blue one. (This roughly follows Holton's example, 2013, pp. 96–97). The device controls the light bulbs' operation – that is, their switching on and off. SIP9000 is then given a challenge: its task is to predict which bulb, blue or red, will be lit at a specified time – say, noon – since only one can be on at once. As a superintelligent predictor, SIP9000 can: (i.) have complete information about how the device works; (ii.) use as much computing power as it needs, for a perfectly accurate calculation; (iii.) possess all knowledge of the universe's workings, including every physical law; and to make things even easier, (iv) it will be guaranteed that the universe is fully deterministic. Moreover, SIP9000 is told in advance that the device was designed solely to make its prediction fail (Holton, 2013), so it cannot be taken by surprise.

The task is simple. As I mentioned, SIP9000 must predict whether the blue bulb or the red bulb will light at noon. But it cannot keep its prediction secret. A button is placed in front of it: SIP9000 must press the blue button if, on the basis of its complete knowledge, it has calculated that the blue bulb will light, and press the red button if it has calculated that the red bulb will light. To count as a genuine prediction, SIP9000 must make its choice exactly one minute before the specified time. For example, if it predicts that the blue bulb will light at noon, it must press the blue button at 11:59 to record its prediction.

t_1: end of calculation time < t_2: articulate the prediction: 11:59 < t_3: Noon.

Prima facie, this challenge looks easy– but it's all too good to be true, and that suspicion proves correct. Above the light bulbs there is a sensing device, let us call it the 'Negator'. It's a simple little circuit that anyone can easily build. The 'Negator', on detecting that SIP9000 has pressed the blue button (predicting blue), lights the red bulb at noon – and vice versa. The 'Negator' waits until SIP9000 reveals its prediction at 11:59, then at noon flips the outcome – lighting the opposite bulb.

t_1: end of calculation time, result of calculation: P = BLUE < t_2: press the blue button at 11:59: P = BLUE < '*NEGATOR'* → t_3: NOON: P = RED

→ P = not P, contradiction

In this case, how did the predictor's omniscience help? Is there any way to outsmart such a 'Negator'? Under what circumstances can the paradox be resolved?

The main thesis is that even in a perfectly deterministic universe, one can easily build a device (the 'Negator', or counter-predictor) that thwarts any prediction by a physical predictor P about a physical system C. Consequently, not all events E can be predicted. The latter can be proved by *reductio ad absurdum*: assume an omniscient predictor P exists, so that the prediction P_n for every event E_n is correct:

$$E_n = P_n$$

Yet the contra-predictive mechanism ('Negator') guarantees some event for which:

$$P_n \neq E_n$$

Since we assumed:

$$\forall_n (P_n = E_n),$$

we arrive at the outright contradiction:

$$P_n \neq P_n (P_n = \neg P_n).$$

3.2. The predictor

From the foregoing, the capabilities of predictor *P* are clear:
P is "a.) aware of all universal physical laws, b.) can perform all relevant calculations of mathematics and logic, c.) is a physical predictor, and d.) is part of the physical system it wants to predict" (Popper, 1995, p. 71).

It is easy to see that this predictive ability is omniscient and truly universal, since *P* can do everything that the system *C* – which it aims to predict – can do (see pont i). That is, *P* can read, simulate and understand the meta-theory governing *C*'s behaviour: its operational principles, decision processes, and so on. "For every physical event there exists a predictor (it is physically possible to construct a prediction) which is able to reproduce the event in question in another system by reproducing one of the states of affairs which preceded the event" (Popper, 1950a, p. 126). To illustrate why universal prediction requires *P* to be able to simulate *C*'s exact decision processes, consider the following:

Let *P* be a computer that can perform only arithmetic operations, and *C* a chess automaton that can compute only chess moves. In that case, it is clearly impossible to ask the chess automaton for arithmetic steps, and vice versa. Consequently, any predictor *P* that aims to compute the moves of *C* must necessarily contain a full description of the chess automaton. Moreover, if *P* is to predict the behavior of two machines – one a chess automaton and one an arithmetic calculator – then, for accurate prediction, *P* must be able to simulate both itself and the other two machines and possess a meta-theory encompassing the descriptions and algorithms of both. Only under those conditions can we meaningfully speak of a universal predictor/universal prediction.

4. Rummens and Cuypers' arguments: embedded and external predictability

To illustrate the paradox, Rummens and Cuypers first distinguish between *embedded predictability* and *external predictability*, and then argue *a priori* that the paradox can arise only in the former case – namely, when the predictor is part of the physical universe *U* (Rummens & Cuypers, 2010). Their central argument is that the paradox of predictability arises only when three necessary conditions hold simultaneously (a, b, c) – if any one fails, no paradox occurs. These are:

(a) the aforementioned *embedded predictability* (the predictor is embedded within the physical universe);
(b) the *revelation* (causal connection), i.e. the system somehow learns of its own predicted behavior, or the predictor is compelled to reveal its prediction;
(c) and a *counter-predictive mechanism*, meaning that once the prediction is revealed the system deliberately acts contrary to it (Rummens & Cuypers, 2010, p. 237).

This thought experiment also circumvents the common objection that no internal prediction can be successful in principle because the predictor does not have access in time to all the data needed for the prediction (including the extent to which the information obtained has interfered with the other system).

It is important to emphasize that an essential element of Rummens and Cuypers' thought experiment on the paradox of predictability is the removal of every obstacle – what they term "epistemic limitations"– from the internal predictor. In other words, they grant the predictor infinite knowledge and unlimited computational capacity. They likewise set aside the impossibility of acquiring information about events outside its (space-time) light cone. Moreover, they assume it can complete its computation in finite time to any required degree of accuracy. By doing so, they preempt the usual objection that any embedded predictor must necessarily fail, since it cannot gather all the data needed for the prediction in time (including how its own data-collection disturbs the system). Even under these idealized conditions, however, the "omniscient" predictor embedded within the system cannot predict every event: as the light-bulb example shows above, it inevitably faces an unsolvable system of equations (Rummens & Cuypers, 2010).

Suppose that at an initial time t_0 (condition a), the subsystem S_1, embedded within the universe U, is asked to predict the future action E of another subsystem S_2 at a later time t_2. That action can take one of two values:

$$E = 0 \text{ or } E = 1.$$

At an intermediate time t_1, with $t_0 < t_1 < t_2$, S_1 must make its prediction P by physically printing "0" or "1" on a slip of paper. Thus the prediction task is simply $P = E$.

The second condition – revelation (b) – is met if S_1 learns of the prediction about S_2 before t_2. For example, Jacob discovers before the vote that his neighbour has predicted he will vote for the Republicans. Once P is revealed, conditions (a) and (b) are satisfied. Finally, assume that S_2 is counter-predictive (c): it always does exactly the opposite of whatever was predicted. Upon learning P, S_2 enacts

$$E = not\ P\ (\mathrm{E} = \neg \mathrm{P})$$

but since we have already assumed that $P = E$, it is a contradiction (Rummens & Cuypers, 2010, pp. 234–237).

Thus, the paradox lies in the fact that any prediction P that subsystem S_1 (embedded within universe U) makes about the future event E of another subsystem S_2 at time t_2 – when S_2 is likewise embedded in U – is inevitably self-refuting (Rummens & Cuypers, 2010, p. 237).

Rummens and Cuypers do not regard the apparent contradiction as substantive unpredictability. They explain this by noting that, in all other respects, the prediction is accurate. Predictor P correctly predicts system C's choice in advance; yet once C learns of P's accurate prediction, C deliberately acts contrary to it. Rummens and Cuypers argue that had C remained unaware of the prediction, P's prediction would indeed have been correct. Additionally, P knows that as soon as it reveals its correct prediction, C will defy it. Under these conditions, P can never get ahead of the system – its omniscience, predictive power only carries it this far.

Furthermore, they contend that the paradox would not even arise if the predictor were external to the universe, a non-physical observer. An external predictor P^* (a demon-like entity) not embedded in our universe would lack any causal link to system C – so conditions (a) and (b) fail. Thus, such an external predictor does not have to reveal its own prediction to agent C. So C remains unaware of the prediction, it does not affect it. In this case, if the internal physical predictor's prediction is P_{em} and the external (non-physical) predictor's prediction is P_{ex}, then $P_{em} \neq P_{ex}$ must also be true.

The central flaw in Rummens and Cuypers's reasoning, however, is their failure to consider the case where agent C and predictor P coincide (not two separate entities). Prediction is twofold in concept: on one hand, hetero-predictability ($P \neq C$) (MacKay, 1967; Grünbaum, 1971, p. 314) refers to one agent (P) predicting another one (C) and its limits; on the other hand, self-predictability ($P = C$) refers to an agent predicting its own actions.

Moreover, in the default case – due to physical constraints –, identity ($P = C$) is established as soon as P interacts with C. That is, if P's prediction in any way affects agent C, P can no longer remain independent of C, and the problem of self-prediction cannot be avoided. As Popper wrote, the point is that once system C 'discovers' predictor P (or any system its assigned predictor), i.e., acquires information about it, from that point onward predictor P will no longer be able to predict that system C, because C's future behavior will immediately become a function of predictor P's own behavior. This makes C a part of P, they form a system

($C = P$). Consequently, P ought to predict its own action, which is precisely what it cannot do (Popper, 1950a).

5. If $P = C$, counterargument to Rummens and Cuypers' theses

To illustrate where Rummens and Cuypers may be mistaken, suppose we have a predictor P – defined by their own conditions – that attempts to make a scientific prediction about its own decision ($P = C$). P is a physical predictor embedded in the mechanical universe U; it possesses the potentially infinite knowledge required to make a correct prediction, can complete its computation in finite time to any desired degree of accuracy, so its prediction will indeed be correct.

As noted, Rummens and Cuypers argue that the paradox does not imply substantive in-principle unpredictability because if a predictor P kept its prediction secret from agent C (thus causal relation [b] would not obtain), C could not falsify it. For instance, having full knowledge I predict that Jacob will vote for the Democrats. I write it on a piece of paper, mail it, without telling him. He only reads it after voting and sees I was correct. While perfectly plausible in isolation, this argument fails to capture the full predictability problem. The real question remains: can an omniscient, physical predictor P accurately compute what it will compute for itself? Even if P considers both possible answers – "yes" or "no" – it still leads straight back into the paradox by choosing "no".

Theorem I: Although the prediction made by a physical predictor P about another, completely independent physical agent C may be correct, P still cannot know in advance what it will compute for itself – this follows from the fundamental impossibility of self-prediction.

For the counterargument, first I will show that the paradox does not merely extend to the stage described by Rummens and Cuypers but in fact stems from the impossibility of self-prediction. Next, I will demonstrate that the paradox itself is a direct consequence of fundamental unpredictability. To make this logical case – and thus refute Rummens and Cuypers' thesis – I must begin by explaining precisely what "the impossibility of self-prediction of physical systems" means and why it necessarily arises.

5.1. The impossibility of self-prediction

For my thought experiment, I will employ MacKay's (1967) arguments, *mutatis mutandis*. MacKay was the one who used conscious human agents to illustrate Popper's basic idea: *the impossibility of self-prediction.*

This is significant because Rummens (2024) rejects the case for substantive unpredictability on the grounds that it is purely an artificial, formal, mathematical construct with no bearing on physical (human) agents. He therefore also rejects Gijsbers' thesis (see below) that the paradox ultimately reduces to Turing's halting problem. Before presenting the logical arguments for substantive unpredictability, I will first set out my counter-argument as applied to human agents.

Ad 1. MacKay's fundamental idea (MacKay, 1967; Watkins, 1971) was to postulate that the human brain operates as mechanically as clockwork – envisioning an extreme scenario in which mechanistic brain theory becomes a fully deterministic science. As he put it: "Suppose that all the relevant facts on the workings of your brain could be made available, without disturbing it, to a computer system capable of predicting its future behaviour from these facts and the environmental forces acting on your nervous system" (MacKay, 1967, p. 8). As a first step, let us consider this possibility.

Ad 2. MacKay (1967) also postulated that when a human agent learns a prediction about itself, that very knowledge inevitably disturbs the physical brain – consistent with mechanistic brain theory. This disturbance, in turn, explains why an agent's earlier prediction of its own decision can become obsolete once the decision itself is consciously realized (recorded) as new (additional)[5] information.

Ad 3. Finally, MacKay's fundamental idea also serves as a counter to Rummens' arguments because, in his precise definition, Rummens emphasizes that a prediction denotes a physical event occurring in space-time. According to him, predictor *P* actually performs the 'computation' and stores the result in its memory."This prediction is therefore either physical, a hardware memory record, or a physical brain state, depending on the nature of the predictor" (Rummens, 2024, p. 2099). Since a prediction is a computational process that lasts for a certain

[5]In a certain sense – under scientific determinism – we couldn't register any genuinely new information if every data were already at hand and every prediction could be logically deduced from it. However, prediction is a process, not an instantaneous inference: even with complete data, the agent only learns the outcome later, as the result of a computational procedure at some subsequent moment. In this respect, the conscious realization of a prediction (including becoming aware of one's own prediction) counts as new information.

period (up to time *t*), the change in the physical brain state occurs *after* the prediction (computation) is completed – once the new data have been revealed.

Based on MacKay, let us accept these two conditions*, mutatis mutandis*:

1. *The mechanistic brain theory is true:* the brain functions in a completely mechanistic fashion, with the determinism of a computer.

2. *Self observation by the agent:* the agent's own brain activity is not monitored by external lab technicians (as in MacKay's original scenario), but by the agent itself, using a computer capable of computing its own predictions.

Then, let us also suppose the following:

A physical agent *P*(eter) wishes to compute his future decision and has all the inputs required for that prediction – namely, all data relevant to the prediction, including complete knowledge of the environmental forces acting on *P*'s nervous system. Under the definition of scientific determinism, the following holds true for *P*(eter)'s prediction:

Given the set of brain-state propositions $S_1 \ldots S_n$ and the deterministic laws *L*, Peter's brain can deductively compute his future decision (prediction *P*) in advance and with exact precision – and that prediction will be necessarily true. In other words, we accept the premise (also endorsed by Rummens and Cuypers) that *P* logically follows from the conjunction $L \wedge S$. Consequently, the following meta-linguistic proposition holds:

Proposition (1): 'If *L* and *S*, then *P'* is true. Then the prediction computed by the agent is:

L

S

N (If *L* and *S*, then *P*)

P

In short:

L and *S*, therefore *P* (Watkins, 1971, p. 266).

So far, this premise aligns with Rummens and Cuypers' core assumption – that the predictor *P* possesses all information required for an accurate, correct prediction, faces no epistemic constraints, and can complete its computation in finite time to any desired precision.

What happens next?

Given the scenario above, agent *P*(eter)'s brain will compute his own future decision with perfect accuracy. That is, his brain – operating with the determinism of a computer –, has all the necessary inputs to calculate a correct prediction *P* at time t_x, about what he will decide for a later time t_y, and Peter immediately reads this prediction from the monitor. Note, that no one else (e.g. lab observers) needs to announce the prediction. At time t_x the prediction *P* for time t_y becomes consciously present to Peter, as his own true prediction of his decision at t_y.

It follows that if this prediction *P* appears on the monitor before the agent has performed the action – and if we follow the 'mechanistic brain theory' – then the agent immediately records (becomes aware of) the prediction *P* about himself. However, the awareness of the predicted activity also alters the agent's physical brain state, because, as I described above, according to mechanistic brain theory, whenever a human agent acquires new information, the physical brain state necessarily changes. Therefore, when the agent records a prediction, he may in fact modify the prediction *P* itself, for the following reasons:

From the above premise, it already followed that *L* and *S* (the agent's brain-state propositions) logically imply the correct prediction *P*. However, once the agent becomes aware of *P* at time t_x – thereby augmenting *L* and *S* with *P* as new information (*I*) – which we'll call the strengthened premise *S'* (i.e., $S' = S \wedge P$) – then at time t_y the actual outcome will be *P'* rather than *P*, so the correct prediction is *P'*; otherwise, an inconsistency would arise. Indeed, if we add *I* to *L* and *S* but leave proposition (2) – *L* and $S' \rightarrow P$ – unchanged, an inconsistency follows (Watkins, 1971).

Thus *Theorem II* should correctly read:

L and *S'* hence *P'*
And so on.

This shows that, if we accept the above conditions, a physical agent *P*(eter) can never predict his own future decision at time t_y without that very prediction influencing his future

action. That is, he cannot predict his own behavior without taking his own prediction *P* into account. However, the mere act of discovering *P* means that his earlier prediction will no longer remain accurate. It may have been correct up until time t_x, just before he learned of it, but as soon as he becomes aware of it (memory fixation), the prediction can immediately become obsolete. Moreover, the agent cannot integrate this altered state back into his initial computation, meaning he will never be able to *get ahead* of the process. Therefore, he cannot calculate in advance what he will calculate at the later time t_y, because he can always contradict it. Thus, even if a human agent possessed complete information about his own brain state, thought processes, etc., he still could not predict with certainty what he will do in the future.

Consequently, it can be seen that self-prediction ($P = C$) is a more fundamental aspect of the paradox and extends beyond the scenario outlined by Rummens and Cuypers. From this point onward, all that remains is to emphasize that the paradox – the impossibility of self-prediction – implies a substantive unpredictability.

6. Gijsbers' counterargument

Victor Gijsbers – in his paper *The Paradox of Predictability* – has also seeks to justify substantive / in-principle unpredictability, and thus refutes Rummens and Cuypers' arguments about the paradox. His central claim is that neither the (b) act of revelation nor the (a) embeddedness of *P* in the predictive universe is a necessary conditions for the paradox.

Although Gijsbers too, regards the distinction between external and embedded predictors as indispensable for framing the paradox, he nevertheless insists on completely redefining it. In his view, the external predictor as presented by Rummens and Cuypers – in its original form – cannot resolve the paradox. Rummens and Cuypers define this external predictor as a disembodied (demon-like) observer outside universe *U* who, despite not being part of *U*, makes predictions $[U_t = f_L(U_0)]$ for all future events in *U* based on perfect knowledge of all initial conditions U_0 and the law-like function f_L (Rummens&Cuypers 2010, p. 234).

It follows from this definition, Gijsbers argues that such an external predictor computes future events using the same algorithm as the internal predictor. That is, the external predictor (regardless of whether it is disembodied) also arrives at its prediction via a well-defined reasoning process that takes the universe's initial state (U_0) and the laws of nature as its input.Thus, it will find itself in precisely the same position as the embedded, physical predictor, and the failure of one predictor will necessarily result in the failure of the other (Gijsbers, 2023,

p. 585). Moreover, simply positing that a predictor is non-physical does not, by itself, render it external.

Gijsbers' conclusion is that if an agent *C* is capable of executing a process (algorithm) to compute a prediction *P* of its own behavior, then any predictor using that same algorithm will also fail to predict *C*'s behavior. Thus, the paradox demonstrates a form of substantive / in-principle unpredictability. The inability of the predictor to generate an accurate forecast holds regardless of other features of the system – such as whether the predictor is disembodied or physically embedded. He supports this argument with a rigorous formal proof and argues that the paradox of predictability is structurally identical to Alan Turing's proof of the undecidability of the halting problem. Accordingly, whatever holds for the latter will also hold for the paradox of predictability. For the argument to succeed, he claims, it is sufficient that the system in question – the universe – is capable of executing the (algorithmic) computational process that generates the prediction. "If that condition is met, Turing's formal proof allows us to show that P will not, in general, be able to predict the behaviour of the given system" (Gijsbers, 2023, p. 588) – not even in a Platonic sense.

It is well known that the halting problem asks whether there exists an algorithm (Turing machine) that determine, for any arbitrary algorithm and input, whether the algorithm will halt or run indefinitely. Alan Turing proved that no such algorithm exists that can always correctly solve the halting problem. Gijsbers derives the paradox of predictability from this rigorous formal proof. That is, if the halting problem holds, then no program can compute its own behavior in every case when given its own description as input. Any program capable of predicting its own behaviour could be used to construct a counter-predictive statement, which would inevitably lead to a logical contradiction. On this basis, Gijsbers points out that the same applies in the case of the paradox of predictability: any predictor *P* (even if assumed to be omniscient), whether physical or disembodied, that attempts to predict what it will itself predict regarding a future decision, will inevitably encounter the same kind of logical contradiction.

To summarize:

(I) The paradox of predictability does not depend on the specific necessary conditions proposed by Rummens and Cuypers.

(II) The paradox follows from two sources: (II.1) Directly from Turing's proof that the halting problem is undecidable. (II.2) From a specific material condition – namely, that the (computational) prediction process must be realizable (i.e. modelable)

within the universe in question (Gijsbers 2023, p. 588). In other words, universe *U* must be capable of instantiating the very algorithm that produces prediction *P* (see point i. above). If it can, then no internal or external predictor can ever outwit its counter-predictor.

(III) It further follows from II.1 that, given the structural identity between Turing's proof and the paradox of predictability, there is a rigorous formal proof that no deterministic system *C* can ever be predicted – even by an omniscient predictor *P*. Both scenarios involve a system attempting to predict its own behaviour (self-prediction), and in every case this attempt must fail (Gijsbers 2021).

(IV) Therefore, if the halting problem is accepted as true, the sole philosophical upshot is that "Turing's proof shows that perfect knowledge of perfectly deterministic and perfectly determinate laws does not imply complete predictability" (Gijsbers, 2023, p. 590), not even in principle.

Gijsbers argues that the very concept of the Turing machine is the true source of the contradiction, since all the properties relevant to the proof are shared by both the machine and the mathematical system it represents. As such, the Turing machine can be regarded simultaneously as a physical device and a mathematical construct – making it easy to extend the indeterminacy of a mathematical problem to certain physical systems, including human agents (Gijsbers, 2023, p. 595).

Finally, if the paradox shows that complete computability is logically impossible, then only one question remains: what further philosophical implications might this have for the concept of free will?

7. Concluding remarks – Paradox and the freedom of will

While the paradox indeed refutes complete predictability, the contradiction itself does little to advance our understanding of free will. Not least because we still lack consensus on what free will actually means. As I noted earlier, some endorse more ambitious libertarian accounts, others more modest compatibilist ones – but simply rejecting the paradox does not directly coincide with any well-known model of free will.

However, if we begin from the premise – outlined at the start of this paper – that free will entails the capacity to perform actions no predictor can calculate in advance, then under this definition a necessary condition for free will is the refutation of scientific determinism.

Moreover, the paradox remains invariant across diverse philosophical "isms," even those in fundamental conflict. Therefore, with the above arguments, we have done no small thing (Holton 2013) in removing the most formidable obstacle to free will. Moreover, we have achieved this without invoking any theory beyond that inherent in the notion of scientific determinism or relaxing its draconian conditions.

Conflict of Interest Statement: The author declares that there are no conflicts of interest related to this study.

References

Boyd, R. (1972). Determinism, laws, and predictability in principle. *Philosophy of Science, 39*(4), 431–450. https://doi.org/10.1086/288466

E. Szabó, L. (2002). *A nyitott jövő problémája – véletlen, kauzalitás és determinizmus a fizikában*. TYPOTEX.

Gijsbers, V. (2023). The paradox of predictability. *Erkenntnis, 88*, 579–596. https://doi.org/10.1007/s10670-020-00369-3

Grünbaum, A. (1971). Free will and laws of human behavior. *American Philosophical Quarterly, 8*(4), 299–317.

Holton, R. (2013). From determinism to resignation; and how to stop it. In A. Clark, J. Kiverstein, J. Vierkant, & V. Tillman (Eds.), *Decomposing the will* (pp. 87–100). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199746996.003.0005

Kant, I. (2018). *A tiszta ész kritikája*. Atlantisz.

Kukla, A. (1978). On the empirical significance of pure determinism. *Philosophy of Science, 45*(1), 141–144. https://doi.org/10.1086/288786

Kukla, A. (1980). Determinism and predictability: Reply to Dieks. *Philosophy of Science, 47*(1), 131–133. https://doi.org/10.1086/288916

Mackay, D. M. (1967). *Freedom of action in a mechanistic universe: The twenty-first Arthur Stanley Eddington Memorial Lecture, delivered at Cambridge University, 17 November*. Cambridge University Press.

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. https://doi.org/10.1093/bjps/I.2.117

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. https://doi.org/10.1093/bjps/I.3.173

Popper, K. R. (1995). *The open universe: An argument for indeterminism from the postscript to the logic of scientific discovery*. Routledge.

Rummens, S., & Cuypers, S. E. (2010). Determinism and the paradox of predictability. *Erkenntnis, 72*, 233–249. https://doi.org/10.1007/s10670-009-9199-1

Rummens, S. (2024). The roots of the paradox of predictability: A reply to Gijsbers. *Erkenntnis, 89*(5), 2097–2104. https://doi.org/10.1007/s10670-022-00617-8

Scriven, M. (1965). An essential unpredictability in human behaviour. In B. B. Wolman & E. Nagel (Eds.), *Scientific psychology: Principles and approaches* (pp. 411–425). Basic Books.

Takács, G. (2013). Fizika a standard modellen innen és túl. *Természet Világa: Mikrovilág, 2013*(1), 3–9.

Watkins, J. W. N. (1971). Freedom and predictability: An amendment to Mackay. *The British Journal for the Philosophy of Science, 22*(3), 263–275. https://doi.org/10.1093/bjps/22.3.263