Dr. Izolda Takacs

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

Even today, there is no precise definition of free will – only mere hypotheses and intuitions. This is why this paper will approach the question of free will from a negative perspective, depicting a scenario in which free will seemingly exists. Subsequently, it will attempt to refute this scenario (as a necessary condition for free will). The absence of *free will* might seem absolute if scientific determinism holds true. Therefore, the goal of the study is to present a logical argument (paradox) that demonstrates the impossibility of an omniscient (P) predictor (scientific determinism), highlighting its inherent self-contradiction. This paradox reveals that the prediction (P = C) by a (P) physical agent of itself is objectively impossible. In other words, even a fully deterministic agent in a deterministic universe cannot predict its own future state, not even in a Platonic sense.

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

1. Introduction

If the doctrine of scientific determinism were true – i.e., 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. For in this case, an omniscient predictor *P* (later christened Laplace's demon) could, in principle, exist, e.g. a supercomputer, that, given all initial conditions and knowing the direction, mass, velocity and laws of physics of all the particles of the universe, could calculate each future event in advance with a desired precision, based on a deductive-nomological model) (Popper, 1950a, 1995). Thus, all events, including the lines of this 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 paper, I present a strong logical argument (paradox) that clearly demonstrates why such an omniscient predictor *P* cannot exist, even in the Platonic sense. It follows that scientific determinism is not only a falsifiable doctrine – primarily due to quantum mechanics –, but also a self-contradictory one. Therefore, the logical paradox can be seen as a much stronger argument. The paradox of predictability shows that the physical agent's prediction (*P*) of itself (*P* = *C*) is objectively impossible.

The argument itself first appeared in the wake of Karl Popper's writings from the early 1950s (*Indeterminism in quantum physics and in classical physics I-II,* 1950), as well as in the studies reflecting on these (MacKay, 1967; Scriven, 1965), also within the context of free will. 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 theory is correct and scientific determinism is only *prima facie* deterministic.[1] *Nota bene*, Popper's impossibility of self-prediction does not mean (as most authors suggest) that it merely proves that determinism is not equivalent to complete predictability – or that even if it exists, there is an objective explanation for feeling free – but rather implies that the whole universe is indeterministic.

Popper's main argument is that there are two obstacles to the self-prediction of physical systems: a physical obstacle and a (Gödel-type) logical obstacle. And because of these limitations, classical mechanics also has an indeterminacy, essentially similar to that known from quantum mechanics, which, although, not sufficient is a necessary condition for free will. Popper's centre argument has resurfaced in contemporary philosophy and has been followed by an ongoing polemic. Therefore, after presenting the paradox, in the second part of the paper I will present this current debate through two approaches: the arguments of Rummens & Cuypers (2010) and Gijsbers (2023). Their writings (*Determinism and the paradox of predictability, The Paradox of Predictability*) focus on two fundamental paths to the paradox of predictability. One view (Rummens&Cuypers, 2010) holds that the computational process in the physical world (usually some process that gives rise to prediction) is such that, for some principled (but physical) reason, it can only reflect on itself by 'influencing' the computational process itself, causing the 'prediction' to turn out false. The other, more general approach (Gijsbers, 2023) is purely "logical". According to him, the paradox of predictability is related to the impossibility of self-prediction and is based on a substantial unpredictability, since Alan Turing's theorem on the halting problem "logically" follows directly from it. Thus, both ideas take as given that a paradox exists with respect to scientific determinism, but the issue in dispute concerns its source (i.e., the main question is whether we are talking about a substantive or non-substantive unpredictability). In this paper, I will argue that the second argument, i.e., the "logical" argument, is the decisive one.[2]

2. Main concepts, the topic in general

[1] I address this question along the lines of Gijsbers' thesis on the undecidability problem (Turing's halting problem).

[2] 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.

The notion of scientific determinism, as outlined in the introduction, is synonymous with predictability. Predictability means that all events, in principle, can be accurately predicted according to the methods of science. Formally, we can speak of scientific determinism if and only if the following statements are true:

(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, $S_1 ... S_n$ 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, i.e. known in advance (Boyd, 1972, p. 431). *In sum*, if we take the deductive nomological model as a starting point, then scientific determinism is equivalent to predictability.

I believe that determinism itself can be understood as the concept desribed above. Not least because of the definition of quantum mechanics as opposed to total predictability. More precisely 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). Thus, it is worth starting from this assumption and from the fact that scientific determinism is a logically much stronger doctrine than determinism itself. The latter, after all, leads to a trivially true, symmetrical argument (Kukla, 1978, 1980; Holton, 2013).[3]

Thus, if the case for scientific determinism holds, then free will – the concept of which intuitively implies in any case something that cannot be predicted (Kant, B 578), i.e. that no predictor can calculate in advance (necessary condition) even in theory – is merely an illusion, an epiphenomenon. We therefore feel free only because we are unable to calculate in some way exactly what will happen in the future, given our limited knowledge compared to the Laplacean demon, the omniscient super-intelligence.Hence, the subjective sense of freedom may derive from the objective fact (E. Szabó, 2002; Grünbaum, 1972) that our knowledge of the facts of the moment is severely limited.

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

[3]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).

While the ‘Newtonian scientific worldview’ outlined above – scientific determinism –, had to be accepted by thinkers in the pre-Heisenberg era because it was synonymous with scientific validity (Popper, 1995, p. 47), we now live in an age where quantum mechanics has replaced this worldview. It should be noted, however, that even the indeterminism ‘guaranteed’ by quantum mechanics does not provide a sufficient explanation for the existence of free will. On the one hand, because we cannot yet identify any form of within probability, since if we claim that predestination deprives us of free will, so does randomness. On the other hand, in the field of quantum mechanics, there are certain limits to our overall epistemic capacity (E. Szabó, 2002; Takács, 2013), aspects of it that we cannot fully understand, which limits our epistemic access.

While it is true that free will has not been salvaged by quantum mechanics, its empirical argument could still serve to falsify scientific determinism (given that the empirical evidence for quantum mechanics is quite strong, based on observations and experiments that suggest that the behaviour of particles in the world is fundamentally uncertain and probabilistic [Takács, 2013], thus challenging the deterministic view of the universe).

However, as the subject of this paper attempts to point out, it is not only the indeterminism, as we known from quantum mechanics that can explain why scientific determinism is false. But there is another line of reasoning that can show, on a more fundamental level, that scientific determinism cannot be fulfilled.

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

3.1. The paradox

The basic thesis of the authors mentioned in the Introduction (Scriven, 1965; MacKay, 1967; Rummens&Cuypers, 2010; Gijsbers, 2023) is that the paradox arises because, even if we assume that in a deterministic universe *U* there exists an omniscient physical predictor *P* that could, in principle, predict with the desired accuracy all future decisions of another physical system *C*, a paradox can be used to demonstrate that *P* could still face a situation – by revealing the prediction to agent *C* – that could, in turn, change the outcome of its prediction.

According to the central argument (Scriven, 1965), if a person or a robot is motivated to counter-predict, their decision cannot be computed even by an omniscient predictor, and his prediction will always prove false. The reason is that when a person or system learns (or replicates) a prediction about itself, it can also falsify it.

Thus, in such a specific situation, the failure of the prediction is guaranteed by the so-called counter-predictivity. For example, if a friend of mine predicts that I will order paella at the restaurant and tells me so, and then, upon receiving this information, I will choose something else in order to refute the prediction. It is this act – the *revelation* of the prediction and the fact that agent *C* can thus refute any prediction that *P* makes about it (the *counter-predictive mechanism*) – that will ultimately make it impossible for *P* to make an accurate prediction.

What exactly is the paradox?

To easily imagine what the paradox or counter-predictive mechanism consists of, suppose that there is a superintelligent predictor, let us call it SIP9000. There is a machine box in front of it, on top of which a red and a blue light bulb are placed. (This is roughly the example of Holton, 2013, pp. 96–97). The operation of the light bulbs, i.e. their switching on and off, is controlled by this machine box. SIP9000 is given a challenge: its task is to predict which of the light bulbs applied to the machine will light up blue or red at a given time, say noon – only one can light up at a time, they cannot light up at the same time. Since SIP9000 is a super-intelligent predictor, it will be able to: i. Have complete information about the operation of the machine box that controls the light bulbs. ii. With as much computing power as it wants, as much as it needs to compute accurately. iii. It also has all the knowledge of how the universe works. It will also know all its laws. iv. To make things even easier, it will be guaranteed that the universe is definitely determinate. Moreover, SIP9000 will be told in advance that the only purpose of the makers of the machine was that its prediction would certainly be wrong (Holton, 2013). So SIP9000 has this knowledge, too, cannot understand surprises.

The task is simple, it only has to predict / calculate whether the blue or the red light bulb will light up at noon, but cannot keep its prediction a secret. There is also a button in front of it, which you must turn on if you have calculated with all your knowledge that the blue bulb will light, and leave off if you have calculated that the red bulb will light. And to be a true prediction, it is of course a requirement that you make your prediction one minute before the time given. For example, if you predict that the blue light bulb will come on at noon, you must press the button at 11:59 to make your prediction:

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

This challenge looks easy, but it's all too good to be true. The hunch turns out to be true, because above the light bulbs there is a sensing device, let's call it the Negator. It's a simple little circuit that anyone can easily build. The 'Negator', when it detects that the button that SIP9000 uses to indicate that the blue one is on, turns on the red bulb at noon, and vice versa. The process is that the sensor-negator always waits until the prediction is made and published by the forecaster, i.e. until 11:59. Then it changes the result to the opposite of the prediction at noon.

t_1: end of calculation time, result of calculation: P = BLUE < t_2: press the 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' machine? In what case can the paradox be resolved?

The main thesis is that even in a perfectly deterministic Universe, it is easy to build a system ('Negator', counter-predictor) that can thwart the prediction of a physical predictor P for a physical system C, and therefore not all events E can be predicted. The latter can be simply proved by assuming – for the sake of *reductio ad absurdum* – that the omniscient predictor P exists. If this is so, then the prediction P_n for all events E_n will be true: $E_n = P_n$. And the contradiction will occur because it can be proved by a so-called contra-predictive mechanism (negator) that there can exist an event when $P_n \neq E_n$. But since we have already assumed that $\forall(P_n = E_n)$, így egy $P_n \neq P_n$ $(P_n = \neg P_n)$, we will encounter a contradiction.

3.2. The predictor

From the above, the capabilities of the predictor in question have been clearly demonstrated: The predictor 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 also not difficult to see that such a predictive ability is omniscient, universal, because P is capable of what the system it is intended to predict, C, is capable of (see [i]). That is, P can read, simulate and understand the meta-theory governing C's behaviour, the principles governing its operation, its decision processes, etc. '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 simply understand why it is important for universal prediction that the same decision process which agent *C* performs can be simulated by predictor *P*, consider the following.

There is a computer *P*, which can only calculate arithmetic problems, and another machine, which is a *C* chess automaton and can only calculate chess steps. In this case, it is obviously not possible to ask the chess automaton for arithmetic steps, and vice versa. Hence, a predictor *P* that wants to compute the steps of a chess automaton *C* must in any case contain a description of the chess automaton. And if predictor *P* wants to predict two computers, a chess machine and a machine performing arithmetic computations, it is obvious that *P* must be able to simulate both its own and the other two machines and have a meta-theory that contains the description and algorithm of the operation of both machines in order to make an accurate prediction. This is the only way to talk about universal prediction.

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

To illustrate the paradox, the authors Rummens and Cuypers first separated 1) *embedded predictability* and 2) *external predictability.* Then they stated *a priori* that the paradox can only exist in the first case, i.e. if the predictor is inside the physical universe *U* (Rummens&Cuypers, 2010).

Their central argument is that the coexistence of three necessary conditions generates the paradox of predictability. If any one of these is not satisfied, then no paradox is generated. It requires the aforementioned (a) intrinsic predictability (embeddedness in the physical universe), (b) revelation (causality), i.e. that the prediction about itself is somehow known by the system, or the predictor being forced to reveal its prediction. The last necessary condition is the (c) counter-predictive mechanism, i.e., the act of going against the prediction after it has been revealed (Rummens&Cuypers, 2010, p. 237).

It is important to point out that an essential element of Rummens and Cuypers' thought experiment on the paradox is that they have removed all obstacles, which they call epistemic limitations, from the path of the internal predictor. In other words, they assumed that the internal predictor possesses infinite knowledge and has infinite computational capacity. They also put aside the impossibility of obtaining information about events outside the (space-time) light cone, etc. They further assumed that it could complete its computation in finite time with

the desired accuracy. 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).

Their argument is that even if all the latter conditions were true, the embedded 'omniscient predictor *P*' would not be able to predict all events, because – as I deduced in the light bulb example – in such a case the embedded predictor would always be faced with an unsolvable system of equations.

Suppose that at an initial time t_0 (a), the subsystem S_1 embedded in the universe is asked to predict the future action E of another subsystem S_2 of the universe U occurring later at time t_2. This action is: $E = 0$ or $E = 1$. S_1 has to make his prediction (P) at time t_1, where $t_0 < t_1 < t_2$. P is a physical event, i.e. S_1 has to print the number 0 or 1 on a piece of paper. The prediction task of S_1 can be formulated as $P = E$. The above condition is satisfied if there is some interaction between S_1 and S_2, say S_1 is informed of a prediction about its future behavior. This is the condition of revelation (b): prediction P is revealed. E.g. Jacob learns before the vote that his neighbor predicted that he will definitely vote Republican. The first two conditions (a, b) of the paradox are thus satisfied. Now suppose also that the subsystem S_2 is counterpredictive (c.) i.e. S_2 always tends to contradict predictions about its future actions (it always does exactly the opposite of what it was predicted to do).

Jacob learns of his neighbour's prediction and votes Democrat just to contradict it. If P is revealed, then $E = not\ P$, but since we have already assumed that $P = E$, it is a contradiction (Rummens&Cuypers, 2010, pp. 234–237).

Thus, the paradoxical nature of the situation is that any prediction of the future action of E at time t_2 of another subsystem (S_2) of the universe U embedded in the universe U will inevitably be self-refuting (Rummens&Cuypers, 2010, p. 237).

Rummens and Cuypers do not consider the contradiction to be a substantial unpredictability. This is explained by the fact that the prediction is otherwise correct. P correctly predicted C's system, was able to accurately predict C's choice in advance, but after its prediction turned out to be correct, C acted contrary to it. They argue that if C had not known about the prediction, P's prediction would have been valid. On the other hand, predictor P also knows that once it reveals its correct prediction, agent C will act in spite of it. Therefore, under such conditions, P will never be able to get in front of the system, so its omniscience is only sufficient for this case.

Moreover, it is believed that the paradoxical situation would not even arise if the predictor were an external (non-physical observer) prediction (external predictability). This is because if the external P^* predictor (demon-like being) is not part of the universe, then it has no causal relation to the system C (conditions a, b are not satisfied).

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 biggest problem with their argument is that they did not seriously consider the situation where an agent C and a predictor P are not two separate entities, but agent C is also the predictor P. For prediction is, in fact, dual in concept. On the one hand it denotes the so-called hetero-predictability ($P \neq C$) (MacKay, 1967; Grünbaum, 1971, p. 314), which refers to the predictability of one (C) system of physical agents by another (P) agent and its limits, and, on the other hand, the so-called self-predictability, when one (P) physical agent predicts about itself ($P = C$).

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 affects agent C in any way, P can no longer be independent of C, so 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 oonward 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 should also predict about itself, but this is precisely what it is unable to do (Popper, 1950a).

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

To easily understand where Rummens and Cuypers may be incorrect, suppose that we have a predictor P, defined by the pair of authors' conditions, which wants to make a scientific prediction about its own decision ($P = C$). P is a physical predictor embedded in the mechanical universe U, where P has the infinite knowledge necessary to make the correct prediction, can complete its computation in finite time with the desired accuracy, and therefore its prediction is a correct conclusion.

Rummens and Cuypers argued that the paradox is not a substantive unpredictability because if a predictor *P* kept its prediction secret from agent *C* (causal relation [b] would not be satisfied), it could not falsify the prediction. E.g. I (with all knowledge) predict that Jacob will vote for the Democrats. I write it down on a piece of paper, but I don't tell him, I mail it to him. Jacob will receive this after the vote and will see that my prediction was correct. That in itself would be a perfectly plausible argument, except that the whole paradox, the predictability problem, doesn't go that far. Even so, the question remains relevant: can an omniscient, physical predictor *P* accurately predict what it will calculate for itself?
Because, if predictor *P* examines both possibilities (*yes / no*), it will also arrive at the paradox by answering '*no'*.

Theorem I: *Although the prediction made by a physical predictor P about another, completely independent physical agent C may be correct, even then P cannot know what its own prediction will be regarding what it will compute for itself, and this follows from the due impossibility of self-prediction.*

For the counterargument, I must explain that the paradox does not merely extend to the stage mentioned by Rummens and Cuypers, but rather arises from the impossibility of self-prediction. And I will then demonstrate that the paradox itself is the direct consequence of an essential unpredictability. However in order to justify the logical argument regarding the self-prediction of physical systems, thereby refuting Rummens and Cuypers' thesis, I will first show what the impossibility of self-prediction of physical systems means and why it occurs.

5.1. The impossibility of self-prediction

For my thought experiment I will use the arguments of MacKay (1967) *mutatis mutandis*. For it was MacKay who demonstrated Popper's basic idea of the impossibility of self-prediction through conscious human agents. This is significant because Rummens rejects the argument for substantive unpredictability – Gijsbers' thesis that the paradox can be traced back to Turing's halting problem (see below) – on the grounds that it is an exclusively artificial, formal, mathematical construct, and not at all relevant to physical (human) agents (Rummens 2024). Therefore, before presenting the logical arguments for substantive unpredictability, I will present my counter-argument by applying it to human agents.

Ad 1. MacKay's fundamental idea was to first postulate (MacKay, 1967; Watkins, 1971) that the human brain operates as mechanically as clockwork – that is he imagined an extreme case in which mechanistic brain theory develops into a fully deterministic science. *"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 acquires knowledge (learns a prediction about itself), it inevitably causes a change (disturbance) in the physical brain = mechanistic brain theory. This, in turn may, explain why an agent's previous prediction about its own decision may become obsolete once it becomes aware of its own decision as new (additional)[4] information.

Ad 3. Finally, MacKay's basic idea serves as a counter to Rummens' arguments because, in his precise definition of prediction, Rummens emphasized that a prediction denotes a physical event occurring in space-time. Predictor P actually performs the 'calculation' 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 computation and lasts for a certain time (time *t*), the change in physical brain state occurs after the prediction is made, after the prediction (new data) is revealed.

Following MacKay let us accept the following two conditions, *mutatis mutandis*:

1. The mechanistic theory of the brain is true: the brain works in a completely mechanistic way, with the determinism of a computer. 2. the agent's brain function is not observed by lab technicians (as in MacKay), but by the agent itself, on a computer (the ability to compute the agent's own prediction). Then, let us also suppose the following:

A physical agent *P*(eter) wants to compute a future decision, for which it has all the necessary inputs for the prediction (: all the data required for prediction, with complete knowledge of the effects of environmental forces on *P*(eter)'s nervous system). According the

[4]In a certain sense – according to scientific determinism –, we could not account for new information if all the data were available, and all predictions could be deduced from it. However, prediction is a process, meaning that the result is not immediately known to the agent (even with complete knowledge of the data), but emerges later through a computational process at some point in time. In this sense, awareness of a prediction (including awareness of one's own prediction) can be interpreted as new information.

definition of scientific determinism, the following will hold true for P(eter)'s prediction: *P*(eter)'s brain S_1... S_n and the set *L* of deterministic laws will be able to deductively calculate his future decision (prediction *P*) in advance and accurately, and the prediction *P* will necessarily be true. That is, let us accept the premise (also adopted by Rummens and Cuypers) that prediction *P* follows logically from the conjunction of *L* and *S*. Therefore, the following meta-linguistic claim will be true:

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' basic assumption that the predictor *P* has all the information necessary to make an accurate, correct prediction (in fact, there are no epistemic constraints in its way, it can complete its computation within finite time with the desired accuracy).

What will happen?

Subject to the above, agent *P*(eter)'s brain will accurately calculate its own future decision. That is, his brain, which operates with the determinism of a computer, is able to calculate, after all the necessary information, at time t_x, a correct prediction *P* for t_y at a later time, which the agent immediately reads from the monitor. Note that this does not require anybody (lab observers) to make the prediction. The prediction of *P* for time t_y at time t_x will be known to the agent as its own correct (true) prediction for time t_y.

It is easy to see from this that if this prediction *P* appears on the monitor before the agent has performed the action – and if we follow the mechanistic theory of the brain – then in this case, that the agent immediately records (becomes aware of) the prediction *P* about

himself. However, the awareness of the predicted activity also changes the agent's physical brain state, because, as I described in the conditions, according to mechanistic brain theory, whenever a human agent acquires new knowledge, it always causes a change in the physical brain. Therefore, when an agent records a prediction, it may also alter the prediction *P* itself, for the following reasons:

It has already followed from the above premise that *L* and *S* logically implied the correct prediction *P*. Except that the original premise *L* and *S*, once realized, is also augmented by *P* as new information (*I*) at time t_x – let the two together be, for simplicity, a confirmed *S*, i.e. *S'*. But then, at time ty, the result will be *P'* instead of *P*, so that the prediction *P'* will eventually be true, since, if it were not, it would lead to inconsistency. For, if I add *I* to *L* and *S*, but the proposition (2) remains *L* and *S'* would therefore remain *P*, it would be inconsistent (Watkins, 1971).

Thus Theorem (2) is correctly *L* and *S'*, hence *P'*
And so on.

This demonstrates that, if we accept the above conditions, *P*(eter) a physical agent can never predict his own future decision (at time t_y) without his own prediction affecting his future action. That is, he cannot make a prediction about himself without having to take his own prediction P into account. However, the change – the mere act of discovering *P* – means that his earlier prediction will no longer be accurate. It may have been correct up until the t_x moment before the prediction was known, but once he learns the prediction (memory fixation), it may become immediately obsolete. Yet the agent cannot integrate this altered state into his initial prediction, 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 t_y time, because he would always contradict it. Thus, even if a human agent possessed all information about his own brain state, thought processes, etc., it 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 concept regarding the paradox and extends beyond the point outlined by Rummens and Cuypers. From this stage onward, it only remains to highlight that the paradox – the impossibility of self-prediction – encompasses an essential unpredictability.

6. Gijsbers' counterargument

Victor Gijsbers has also attempted to justify substantive unpredictability, and thus to refute Rummens and Cuypers' arguments regarding the paradox, in his paper *The Paradox of Predictability*. His central argument is that neither revelation (b) nor the embeddedness of *P* in the predictive universe (a) are necessary conditions for the paradox.

Although Gijsbers also considers the distinction between the external and the internal predictor to be indispensable for presenting the paradox, he nonetheless calls for a complete redefinition of it. In his view, the external predictor as described by Rummens and Cuypers, as it stands, cannot resolve the paradox. Rummens and Cuypers' external predictor was by definition a disembodied, external observer who, although not part of the universe *U*, makes predictions $[U_t = f_L (U_0)]$ for all future events in *U* based on perfect knowledge of the initial conditions U_0 and the law-like function f_L (Rummens&Cuypers 2010, p. 234). It follows from this definition, according to Gijsbers, 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 or not) also makes its prediction *via* a well-defined reasoning process, taking the initial state of the universe (U_0) and the laws of nature as its input. For this reason, it will find itself in exactly the same situation as the embedded, physical predictor, and the defeat of one predictor will necessarily result in the defeat of the other (Gijsbers, 2023, p. 585). Moreover, just because a predictor is assumed to be non-physical does not make it external.

Gijsbers's conclusion is that if an agent *C* is capable of the process (algorithm) used to compute the prediction *P* of its behaviour, then any predictor using the same process will be equally unpredictable to that system *C*. Hence, the paradox implies inherent unpredictability, and the inability of a predictor to make an accurate prediction will hold regardless of other properties of the system, such as whether it is a disembodied or a physical predictor.

Gijsbers supports his argument with a rigorous formal proof. He argues that the paradox of predictability is structurally identical to Alan Turing's proof of the undecidability of the halting problem. Consequently, whatever is true of the former will be true of the predictability paradox. For this argument to hold, all that is required, he says, is that the system in question, the Universe, "*must be able to incorporate the process (algorithmic computational T.I.) that generates P's 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).

It is well known that the halting problem seeks to answer the question whether there exists an algorithm (Turing machine) that determines for an arbitrary algorithm whether it halts or runs to infinity for a given input. Alan Turing proved that there is no algorithm that always solves the stopping problem correctly. Gijsbers derived the predictability paradox from this rigorous formal proof. That is, if the halting problem holds true, then no program can compute its own behavior in every case, given its own description as input. Because any program that can predict its own behaviour can be used to construct a counter-predictive statement, leading to a logical contradiction in any case. From this, he pointed out that exactly the same situation occurs in the case of the paradox of predictability, which is why any predictor P (assumed to be omniscient) – whether physical or disembodied –, that attempts to predict what it will itself predict about a future decision will inevitably encounter the same logical contradiction.

To summarize:

(I) The paradox of predictability is not dependent on of the necessary conditions specified by Rummens and Cuypers.

(II) Second, from a certain material condition, which means that the paradox requires that the (computational) process of prediction must be able to manifest/be modelled in the given universe (Gijsbers, 2023, p. 588). In other words, the universe *U* must be able to generate the process that generates the prediction of *P* (see also above under point i). If it can, then it can be stated that there can never be an internal or external predictor that can outsmart the counter-predictor.

(III) It further follows from Proposition II.1 that, by the structural identity that can be drawn between Turing's proof and the predictability paradox, there exists a rigorous formal proof that a deterministic system *C* can never be predicted even by a predictor assumed to be omniscient, *P*. After all, both are associated with a system that tries to predict its own behaviour (self-prediction) and this will always turn out to be false (Gijsbers, 2023).

(IV.) So, if the Halting problem is accepted as true, the only philosophical consequence 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 theory.

Gijsbers believes that the concept of the Turing machine is the real source of the contradiction, because all the properties relevant to the proof are common to the machine and to mathematics. Thus, it can be considered both a physical device and a mathematical construct,

so that it is very easy to cross the boundary of the indeterminacy of a mathematical problem in the direction of the indeterminacy of a certain type of physical entity (even human agents) (Gijsbers, 2023, p. 595).

Finally, if the paradox can demonstrate that complete computability is logically impossible, the question arises as to what further philosophical implications this might have for free will.

7. Concluding remarks – Paradox and the freedom of will

While paradox does indeed disprove total predictability, contradiction does not bring us much closer to free will. On the one hand, we do not know whether the world is metaphysically indeterministic or deterministic, and on the other hand, we do not know exactly what free will means. As I said, there is no consensus on the latter. There are more ambitious libertarian and less ambitious compatibilist notions of free will, but the rejection of paradox itself does not directly overlap with any of the well-known models of free will.

At the same time, if we assume – as I outlined at the beginning of this paper – that free will certainly implies some action that no predictor can predict, then in line with this concept, one of the necessary conditions of *free will* is the refutation of scientific determinism. In addition, the paradox remains invariant to different philosophical *isms* that conflict on fundamental issues. Therefore, the arguments presented above have made a significant contribution (Holton, 2013) in removing the most formidable obstacle to free will. Moreover, we have achieved this without needing to include any theory other than the one implicit in the concept of scientific determinism, or without having to relax the draconian conditions of scientific determinism.

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