Erkenn (2010) 72:233–249
DOI 10.1007/s10670-009-9199-1

ORIGINAL ARTICLE

Determinism and the Paradox of Predictability

Stefan Rummens · Stefaan E. Cuypers

Received: 3 September 2008 / Accepted: 19 October 2009 / Published online: 25 November 2009

Abstract The inference from determinism to predictability, though intuitively plausible, needs to be qualified in an important respect. We need to distinguish between two different kinds of predictability. On the one hand, determinism implies *external predictability*, that is, the possibility for an external observer, not part of the universe, to predict, in principle, all future states of the universe. Yet, on the other hand, *embedded predictability* as the possibility for an embedded subsystem in the universe to make such predictions, does not obtain in a deterministic universe. By revitalizing an older result—the *paradox of predictability*—we demonstrate that, even in a deterministic universe, there are fundamental, non-epistemic limitations on the ability of one subsystem embedded in the universe to predict the future behaviour of other subsystems embedded in the same universe. As an explanation, we put forward the hypothesis that these limitations arise because *the predictions themselves are physical events* which are part of the law-like causal chain of events in the deterministic universe. While the limitations on embedded predictability cannot in any direct way show evidence of free human agency, we conjecture that, even in a deterministic universe, human agents have a *take-it-or-leave-it control* over revealed predictions of their future behaviour.

S. Rummens
Institute for Management Research, Radboud University Nijmegen,
PO Box 9108, 6500 HK Nijmegen, The Netherlands
e-mail: s.rummens@fm.ru.nl

S. E. Cuypers (✉)
Institute of Philosophy, Centre for Logic & Analytical Philosophy, Katholieke Universiteit Leuven,
Kardinaal Mercierplein 2, 3000 Leuven, Belgium
e-mail: Stefaan.Cuypers@hiw.kuleuven.be

1 Introduction

In the sixties and seventies of last century, the discovery of, what we call, 'the paradox of predictability' triggered a lively debate on determinism, predictability and free will, mainly in British philosophy. Although it ended without any clear upshot and with a lot of loose ends, we have three reasons to revisit and reopen this debate. First, the introduction of an explicit distinction between two different kinds of predictability, allows us to reformulate the paradox in a way that transparently reveals its underlying structure (Sect. 2). As a result, and this is our second and main point, we are able to show why the earlier explanations of the paradox failed (Sect. 3) and, subsequently, to give a new alternative explanation of its origin (Sect. 4). Third, we believe that the paradox of predictability might have important as yet unnoticed consequences for the contemporary compatibilism–incompatibilism debate on free will. In this paper, we elaborate the first two points, while leaving the third—apart from a promissory note (Sect. 5)—for another occasion.

2 The Paradox of Predictability

Before formulating the paradox of predictability, we introduce some basic terminology. We assume, first, that a universe U is *deterministic* when, for any arbitrarily chosen time t_0, there exists a law-like function f_L which maps the initial state of the universe U_0 at time t_0 in a unique manner onto the state of the universe U_t at any arbitrarily chosen later time t:[1]

$$U_t = f_L(U_0) \tag{1}$$

We assume, furthermore, that determinism implies *external* (*or Laplacean*) *predictability*. This kind of predictability is defined as the possibility of a (God-like) external observer, not part of the universe U, to make predictions of all the future events in U on the basis of its perfect knowledge of the initial conditions U_0 and the law-like function f_L.[2] For our purposes, it is, finally, crucial to distinguish this notion of external predictability from that of embedded predictability:[3]

[1] The function needs to be law-like, for, even in a non-deterministic universe, a non-law-like function between states of a universe at different times can always be defined. Specifying the sense in which this function needs to be 'law-like' is beyond the scope of the present paper (see Russell 1917; Hempel 1961; Boyd 1972; Kukla 1978, 1980; Dieks 1980). We submit that our research presented here is independent from the details of a more specific account of determinism.

[2] The Laplacean observer has a 'view from nowhere' and takes a detached 'objective view'—*sub specie aeternitatis*—of the universe as a whole. For this theoretical standpoint, from which we also wrote this paper, see Nagel (1986). When we say that this external observer and its predictions are 'non-physical', we only mean *outside U*, *not* embedded in the physical universe itself. We remain neutral as to the intrinsic nature of this external standpoint—be it abstract (Platonic), spiritual or material.

[3] Karl Popper (1950, pp. 120–126) makes a similar distinction between *metaphysical determinism* and *scientific predictability*. Yet another version of the same distinction can be found in work of Sellars (1966, pp. 143–144), who distinguishes between *logical* and *epistemic predictability*.

> *Embedded predictability* holds in a universe, U, if there exists a subsystem, S, embedded in U—for instance, a highly intelligent (human) being or a very powerful computer—that is able to predict all the future events in U.

In the past, several authors argued that, especially but not exclusively with regard to non-linear systems, there are serious *epistemic* limitations on the ability of an embedded subsystem of a deterministic universe to make predictions of future events (see, for example, Suppes 1985; Bishop 2003). These epistemic limitations arise, for instance, because a finite subsystem of the universe is unable to make the required accurate measurements of the initial conditions of the universe, or because such a subsystem is unable to represent fully accurately the initial value of a variable, if this value is a (non-computable) irrational number (requiring an infinite string of digits). Though we believe that these arguments show that epistemic restrictions in fact or in practice preclude embedded predictability from holding in a deterministic universe, we want to argue for *the stronger claim* that embedded predictability cannot obtain in a deterministic universe even on the idealizing assumption that the epistemic limitations at issue are removed.[4] Therefore, we propose to lift these restrictions and to assume that the subsystem S in our definition of embedded predictability has infinite possibilities of representing information, is able to gather infinitely precise information about the initial conditions of the universe and is able to make calculations with any desired accuracy in a finite amount of time. We accept, of course, that it is physically impossible to realize these infinite capabilities in a finite subsystem of the universe. However, this reference to the infinite epistemic capabilities of subsystem S is not problematic because we simply use it as a shorthand for saying that the predicting subsystem can have *arbitrarily strong* computing capabilities. Authors who argue for unpredictability in terms of the epistemic limitations of predicting subsystems typically start from a given subsystem with finite capabilities and then construct a corresponding future event that remains beyond the predictive scope of this given subsystem. In contrast, we will identify a single future event that remains unpredictable for all possible predicting subsystems, no matter how strong their computing capabilities.

Already in 1950, Karl Popper has made an attempt to make this stronger case against embedded predictability. Although we think that his highly complex and at times overly technical arguments are of genuine interest, we believe that they are ultimately, as Popper himself seems to admit, not fully convincing (Popper 1950, p. 193).[5] A more successful attempt to make the stronger claim can be found

[4] Both external and embedded predictability are types of *conditional* predictability—predictability in theory or in principle—as defined by Honderich (1988, p. 337): 'If something is conditionally predictable, it could be predicted *if* certain knowledge existed, the having or recording of which knowledge is logically or conceptually possible, although perhaps factually or nomically impossible.' Since Honderich does not distinguish between external and embedded predictability, as we do, he fails to notice that his arguments against 'determinism without predictability' (1988, pp. 347–350) do not pertain to embedded predictability but only to external predictability.

[5] Popper (1950) has tried to show that there are unavoidable limitations on the ability of a predictor to predict its own future. He elaborates three arguments. The first and the third ones essentially suggest that a physical predictor is unable to describe its own describing activities and that it cannot contain sufficient information about itself to make accurate predictions. These arguments deal with the kind of epistemic limitations we have excluded from the start. The second argument makes use of Gödel's incompleteness

implicitly in work of Donald MacKay (1960, 1961, 1967, 1971, 1973) and in a more explicit way in that of Scriven (1965) and P.T. Landsberg and D.A. Evans (1970, Evans and Landsberg 1972). These authors argued for the thesis that, conditional on some minimal assumptions about the universe U, there are fundamental limitations of a non-epistemic nature on the ability of any subsystem S to make predictions and that, therefore, even in the deterministic universe U, embedded predictability does not hold. They established this conclusion by constructing a paradoxical situation in which a subsystem S makes a prediction which is *unavoidably self-defeating*. Though we believe that their arguments provide a convincing example of a situation in which embedded predictability is impossible, we also believe that their accounts of this paradoxical situation are unwieldy and that, consequently, these authors failed to fully grasp the real origin of the paradox and the general nature of the unpredictability involved. Therefore, we proceed, first, by presenting our reconstruction of the paradox which allows us to lay out its underlying structure.[6]

Suppose that, at an initial time t_0, the subsystem S_1 is asked to predict the future action, A, at a later time t_2, of another subsystem of the universe U, S_2. Assume that this action is that of making a choice between 0 and 1, such that either $A = 0$ or $A = 1$. Imagine, for instance, that subsystem S_2 is a human being who has to cast a vote in a referendum at time t_2. Assume, further, that S_1 has to make its prediction, call it P, at a specific time t_1, whereby $t_0 < t_1 < t_2$. Importantly, P is an identifiable physical event: S_1 is requested, for instance, to write down or print out the number 0 or 1 on a slip of paper. S_1's predictive task can be formulated as

$$P = A. \tag{2}$$

The paradox of predictability arises when the following two conditions obtain. First, there is some kind of *interaction* between the two subsystems, S_1 and S_2, such that the prediction P is revealed to the subsystem S_2. Imagine, for instance, that the human being is informed about, or secretly finds out the prediction of its future behaviour. Call this the revelation condition:

Footnote 5 continued
results (1950, pp. 179–188). Here Popper suggests that a computer will be unable to predict at time t whether it will itself be able to answer, at a later time t', the mathematical Gödelian question regarding the consistency of the information it has available at that later time. Since this Gödelian question is undecidable, the same holds for the physical question at the prior time t regarding the decidability at the later time. The problem with this argument, however, is that there is no reason whatsoever to suppose that the predictor should even be willing to consider, at the later time t', the Gödelian question Popper has in mind. Since this question is a mathematical one and not a physical one, it suffices to construct a better predictor which refuses to consider mathematical questions by generating some error-message. The generation of this error message (as a physical event) will be perfectly predictable at the earlier time t and so no convincing counterexample which disproves embedded predictability has been generated.

[6] Landsberg and Evans' account is cumbersome in the sense that it makes reference to the brain state of the predicted agent. Whereas this is justified in view of their attempt to criticize Donald MacKay, it obscures the logical simplicity and generality of the result. Scriven's account is unwieldy because it is couched in utilitarian language and it conflates two different issues. As Lewis and Richardson (1966, p. 74, n.3) confirm, the issue of unpredictability in possession of full information is conceptually entirely distinct from the randomizing contrapredictive stratagems used in a case of limited information (which is the main focus of Scriven's paper but not our present concern).

$$P \text{ is revealed.} \tag{3}$$

Second, assume that the predicted subsystem S_2 operates on the basis of, what Scriven calls, a *contrapredictive mechanism.*[7] This means that the subsystem S_2 always aims to defy any prediction made of its future actions and, as a consequence, always does the exact opposite of what is predicted. Imagine, for instance, that the human being has a recalcitrant personality and that it always does the opposite of what people expect. Formalized, the contrapredictivity condition is:

$$\text{If } P \text{ is revealed, then } A = \textit{not-P}. \tag{4}$$

Putting both conditions, (4) and (3) in a modus ponens format directly generates the paradox. Given that the actual behaviour of the (contrapredictive) subsystem S_2 (after revelation) is exactly the opposite of the predicted behaviour P, that is,

$$A = \textit{not-P}, \tag{5}$$

S_1's predictive task, from (2) and (5), must be described by the unsolvable equation:

$$P = \textit{not-P}. \tag{6}$$

When the revelation and contrapredictivity conditions obtain, if S_1's prediction at t_1 is $P = 0$, then S_2's action at t_2 will be $A = 1$; if S_1's prediction at t_1 is $P = 1$, then S_2's action at t_2 will be $A = 0$. The paradoxical nature of the situation thus resides in the fact that any prediction made by S_1 is inevitably self-defeating.

3 How Not to Explain the Paradox (Away)

We briefly elucidate our reconstruction of the paradox of predictability against the backdrop of the earlier 1960–1970 discussion. We believe that, until now, no convincing explanation of the underlying structure of the paradox has been given. To dispel some of the misunderstandings which have clouded the debate, four important clarifications are in order.

First, the self-defeating structure does not, as some have mistakenly thought (DeWitt 1973), depend on some supposedly unique ontological characteristics of cognitive events, or on the presence of some special kind of free will. There is nothing extraordinary about a contrapredictive mechanism. We can safely assume that such a mechanism can be instantiated by a variety of biological, psychological or electronic mechanisms. Contrapredictivity could, for instance, easily be programmed into a computer, requiring the generation of output 1 if and only if receiving input 0, and vice versa (Landsberg and Evans 1970, pp. 356–357; Good 1971).

Second, the self-defeating structure is not, as some authors wrongly assume (Goldman 1968, pp. 148–149), attributable to some lack of knowledge on the part of

[7] Scriven (1965) considers human beings and assumes that they could be *contrapredictively motivated*. Psychological evidence for the existence of such a type of motivation in humans can be found in the theory of 'psychological reactance' (Brehm 1966). Because we want to provide a more general account by leaving it open whether S_2 is a human being or a machine, we prefer to talk about a *contrapredictive mechanism*, thereby highlighting the deterministic nature of the pertinent mechanism.

the predictor S_1. Additional knowledge about the physical details of the interaction or the workings of the subsystem S_2 will not change the logical outcome that the predictive task yields the same unsolvable equation $P = not\text{-}P$. The fact that this equation is unsolvable does not, however, generate some kind of indeterminacy. We assumed universe U to be deterministic and, therefore, what will happen at t_1 and later at t_2 is uniquely determined. In order to know what will happen, we could make the relevant knowledge complete and add the assumption that if the predictor S_1 fails to arrive at a unique solution for P, it will make the 'default' prediction

$$P = 0. \tag{7}$$

In this case, the prediction at t_1 will be $P = 0$ and, therefore, the action at t_2 will be $A = 1$. This uniquely determined course of events will be clear to everybody, *even to the predictor* S_1. However, this knowledge does not help the subsystem S_1: it also knows that it is unable to generate the correct output $P = 1$, for this prediction would again be self-defeating and bring about action $A = 0$.

Third, it is not the case, contrary to what several authors assume (Lewis and Richardson 1966; Good 1971; also Roberts 1975), that the paradox has something to do with limitations on the speed with which the predictor and the predicted system can make calculations concerning each other's behaviour. These authors seem to have in mind a scenario in which the predictor is able to make a second and secret prediction, P', which would be correct (say, $P' = 1$). Yet, when this secret prediction is revealed, the predicted system will decide to act contrapredictively as regards this second prediction, which will require a third secret prediction, P'' (=0), which in its turn could be revealed, etc. This to-and-fro scenario generates a series of predictions and contrapredictive intentions, such that the subsystem that calculates the fastest will outsmart the other subsystem at the time of action, t_2.

However, this kind of consideration does not genuinely alleviate the paradox. There are three possibilities. We could, first, suspend our assumption that the predicting subsystem has infinite calculating capabilities and interpret this to-and-fro scenario as an actual to-and-fro process, whereby each prediction and each contrapredictive calculation takes a finite amount of time. Whether, in this case, the final prediction made by predictor S_1 before time t_2 will be correct, depends entirely on the contingent characteristics of the calculating and interacting mechanisms. Although, in this scenario, the stronger claim that we can identify a single future event which is unpredictable for all possible predicting subsystems no longer holds, the argument against embedded predictability as such remains. Remember that to demonstrate that embedded predictability cannot obtain, it suffices, strictly speaking, to show that, for every would-be predictor S_1, we can think of at least one situation in which it will fail to make a correct prediction. In this scenario, we can, therefore, secure the paradoxical result simply by *stipulating* that the predicted subsystem S_2 calculates faster than the predicting subsystem S_1.

The second interpretation is also straightforward. If we consider the to-and-fro scenario as an infinite series in a finite amount of time (by assuming, for instance, that both subsystems are able to make infinitely fast calculations), then it simply captures the typically oscillating nature of the paradoxical equation $P = not\text{-}P$ (Wormell 1958). We are again just saying that this equation has no solution and that the predictor will have to resort to the default prediction $P = 0$, which is self-defeating.

On a third interpretation, we could make the asymmetrical assumption that the predicting subsystem S_1 is capable of making infinitely fast calculations, whereas the predicted subsystem S_2 is only capable of finitely fast ones. Whilst this third scenario is the only one in which the proposed 'solution' of the paradox actually works, and S_1 indeed always outsmarts S_2, this resolution comes at a high price. Not only does the asymmetry of the idealizing assumption seem arbitrary, it also remains a fact that by choosing an arbitrarily fast predicted subsystem S_2, we can make the time-gap ε which remains between prediction P and action A arbitrarily small. The claim that S_1 can actually 'predict' the behaviour of S_2, given a very small value for ε, reduces then to the claim that S_1 can fulfil its predictive task simply by waiting until S_2 is no longer capable of reacting to any relevant information from its environment, due to a lack of time. Since this scenario makes the idea of a prediction rather meaningless, we have introduced in our version of the paradox the plausible requirement that the prediction has to be produced at a specific time t_1 ($<t_2$). This assumption simply captures the idea that a real prediction needs to refer to an event that is in some significant sense situated in a finitely (as opposed to infinitesimally) remote future.

Importantly, the determination of a specific time t_1 ($<t_2$) at which the prediction has to be made *immediately blocks all three versions of the to-and-fro scenario*. The determinate, finite lapse of time between the prediction and the action allows the predicted subsystem to respond to P and the unpredictability of A by S_1 at time t_1 is restored. Since A is thus an event which remains unpredictable, independently from the specific characteristics of the predicting subsystem, our stronger claim that embedded predictability cannot obtain even on the idealizing assumption that S_1 has infinite calculating capabilities still stands in our version of the paradox.

Four, and in line with our first two clarifications, the paradox does not reveal some kind of 'essential unpredictability in human behaviour', as Michael Scriven (1965) wants us to belief. To begin with, neither his version of the paradox, nor ours essentially presupposes that the nature of the predicted subsystem S_2 is human. Neither of both versions makes a distinction between human and non-human agency. In addition, and more importantly, the situation we have sketched remains throughout fully determined and, therefore, no essential unpredictability is implied at all (Suppes 1964, p. 144). As we will show, Scriven's claim is based on a serious flaw in his own argument. Close scrutiny of where it goes wrong is worthwhile, since it helps to reveal the real origin of the paradox.

Scriven's version of the paradox differs from ours in one design feature. Although he too presupposes a contrapredictive mechanism, he rejects the idea of an actual interaction between the predicting and the predicted system. Instead, he assumes that the predicted system has access to the same information I and has the same predictive capacities as the predicting system. In this way, the predicted system can itself make the very same prediction P (without the need for interaction or revelation) and, subsequently, act contrapredictively. As we will explain in the next section, we believe that this difference in the design of the assumptions is in itself not really significant and does not actually change the structure of the paradoxical situation.

Vital to our purposes is that Scriven fails to make the crucial distinction between embedded and external predictability. Scriven proceeds in his argument *e contrario ad absurdum*. He adopts the assumption that action A can be logically derived from the available information I and then argues that the contrapredictive behaviour of the predicted system leads to an immediate falsification of this assumption:

> It immediately follows that I either implies an *incorrect* prediction as to which A will be chosen by S_2 (that is, contains false information) or *none*: hence S_2's choice cannot be rationally predicted by S_1 from I. (Scriven 1965, p. 415)[8]

Though the second part of this crucial passage—S_2's choice cannot be predicted by S_1 from I—is correct, the first part is not and, *pace* Scriven, it does not follow from the contrapredictive behaviour of S_2. Scriven's reasoning goes wrong because he fails to distinguish between two different kinds of predictability. On the one hand, Scriven talks about the implication of A by I, thereby referring to the *logical inference* from I to A. On the other hand, Scriven talks about the actual deduction of A from I by S_1 or S_2, which is a *physical event*. As will become clear in the next section, the impossibility of the actual deduction of A by S_1 does not necessarily imply that the available information is insufficient to logically infer A. On the contrary, even in the paradoxical situation at issue, the deterministic nature of universe U guarantees, contrary to what Scriven supposes, that it must be possible to provide sufficient information from which to make the correct inference to A. Scriven conflates the external kind of predictability with the embedded one. From the standpoint of external predictability there is no mystery at all. On the basis of the relevant information about the initial conditions and the laws of universe U, an external observer can predict that S_1 will make a certain prediction P at time t_1 (in our default case, $P = 0$) and can also unequivocally predict that the action A at time t_2 will be the exact opposite of what has been predicted ($A = 1$). From the external standpoint, human behaviour in U remains inferentially predictable in principle. So, the first part of the quoted passage above—I either implies an *incorrect* prediction as to which A will be chosen by S_2 (…) or *none*—is false.

The paradox of predictability does not arise because action A cannot be logically inferred from the information available, but because the embedded subsystem S_1 cannot deduce its prediction without thereby invalidating it. More specifically, S_1's prediction is not an external logical inference. The paradox, we conjecture, arises because S_1 is itself an embedded part of universe U and, therefore, its prediction a *physical event* which is itself part of the causal chain of events.

4 Predictions as Physical Events

Exploring the distinction between external and embedded predictability somewhat further, we can clarify why and how the paradox of predictability is inextricably connected with the embeddedness of S_1 as a subsystem of universe U. This furnishes

[8] We have adapted Scriven's notation to fit our own.

the material for our alternative explanation of the paradox's origin and its underlying structure.

Our paradox is a specific instantiation of a general situation of embedded prediction in which a subsystem S_1 makes a prediction P at time t_1 about a future action A at time t_2 of another subsystem S_2. To fully describe this general situation from *the standpoint of external predictability*, we introduce some additional equations. Recall equation (1): if the universe U is deterministic, U's state at a certain time, t, is a law-like function of U's initial state at t_0, whereby this law-like function defines a necessary and thus unique chain of events. Given that prediction P is an element of U_1's state and that action A is an element of U_2's state, we can, therefore, introduce two law-like functions, g and h, which describe how P and A, respectively, are uniquely determined by the initial conditions of the universe at time t_0. Using, in addition, the symbol P^* to refer to the prediction made by the external observer, the situation for this external observer is captured by the following set of three equations.

$$P = g(U_0) \tag{8}$$

$$A = h(U_0) \tag{9}$$

$$P^* = A \tag{10}$$

These equations mean the following. Equation (8) describes how prediction P, as a physical event at time t_1, is uniquely determined by the state of U at time t_0 and the law-like function g. Equation (9) describes, similarly, how action A, as a physical event at time t_2, is uniquely determined by the state of U at time t_0 and the law-like function h. Equation (10) describes the predictive task of the external predictor, who has to make sure that its prediction P^* adequately predicts action A. Importantly, equation (10) has *no physical* interpretation because the prediction P^* is not itself a physical event, but just the name for the prediction made by the external (and thus non-physical) observer.

As this set of equations makes clear, an external predictor has no problem whatsoever in understanding and describing the chain of events in the paradoxical situation at hand. On the basis of its knowledge of g, h and U_0 (and its calculating abilities), this observer is perfectly capable of determining the values of P and A and, thus, to fulfil its predictive task (10). If we apply the general description to the specific case of the paradox, equation (8), in our default case, simply becomes the default prediction $P = 0$ and equation (9) on the revelation and contrapredictivity conditions becomes equation $A =$ *not-P*. Clearly, these equations allow the external observer to make correctly (10) the prediction $P^* = 1$.

Turning to the *standpoint of embedded predictability*, the more general situation can be described by a similar yet different set of three equations.

$$P = g(U_0) \tag{8$'$}$$

$$A = h(U_0) \tag{9$'$}$$

$$P = A \tag{11}$$

The first two equations (8′) and (9′), are identical to the first two equations of the previous set, (8) and (9), and mean the same. They describe how prediction P and action A, as physical events, are uniquely determined by the initial state of U at time t_0 and the laws describing its evolution. The third equation (11), however, is different and describes the predictive task of the embedded predictor S_1. In contrast to equation (10), and this is a point of great consequence, equation (11) has a *physical* interpretation. Prediction P at time t_1 can only be successful if it adequately describes the future action A at time t_2. In our case, this means that the number produced by the predictor S_1 at time t_1 has to be the same as the number chosen by S_2 at time t_2.

As the equations make clear, the fact that P, in contrast to P^*, is a physical event has a dramatic impact on the ability of the embedded predictor to fulfil its task. Whereas the first set of equations (8)–(10) provides three equations from which to determine the value of three variables (P, A and P^*), the set of equations ($8'$)–($9'$)–(11) provides three equations from which to determine the value of only two variables (P and A). In the second set of equations, the variables are thus *overdetermined* and, consequently, a simultaneous solution of all the equations is no longer assured. Overdetermination possibly implies unsolvability.

The explanation of this essential difference is the following. The deterministic nature of the universe implies that the first two equations of each set already uniquely determine the value of P and A. The third equation of the second set, (11), imposes an *additional physical constraint* on an already fully determined universe. Equation (11) thus overdetermines the variables P and A. Prediction P is successful only if a certain relation exists between prediction P as a physical event and the predicted action A as a physical event (whereby P has to provide an adequate representation of the future action A). Whether this relation actually obtains and, thus, whether a successful prediction ($P = A$) is possible, cannot be guaranteed a priori but depends on the specifics of the laws and conditions of the universe and of the situation at hand.

By contrast, in the case of an external predictor, this uncertainty does not arise. In this case, the deterministic nature of the universe guarantees a priori that a successful prediction P^* is always possible. The first two equations (8) and (9) allow the external predictor to determine the values of P and A. Although it is thereby perfectly possible that these two values turn out to be different ($P \neq A$), this only means that the external predictor observes that the embedded predictor is unsuccessful. This observation, however, in no way affects the ability of the external predictor to make its own successful prediction $P^* = A$.

Although we have no general theory to offer about the conditions under which the set of equations ($8'$)–($9'$)–(11) is unsolvable and, therefore, an embedded prediction impossible, we would like to draw attention to a specific category of situations which seem particularly prone to overdetermination. We have in mind situations in which action A becomes an explicit function of prediction P. By this we mean that function $h(U_0)$ can be written as function $f(g(U_0))$, so that equation ($9'$) becomes

$$A = f(P). \tag{12}$$

When this new equation for A is combined with the additional physical constraint, (11), it becomes apparent that this category of situations is captured by the *self-referential* equation (compare Landsberg and Evans 1970, pp. 348–350):

$$P = f(P). \tag{13}$$

To be sure, there are cases in which even equation (13) remains solvable. Imagine, for instance, a situation similar to the paradoxical one, apart from the fact that the predicted subsystem S_2 is no longer characterized by a contrapredictive mechanism but by a *fatalist mechanism* which always acts in accordance with a revealed prediction:

$$\text{If } P \text{ is revealed, then } A = P. \tag{14}$$

In this case, the actual revelation of P generates a situation in which f becomes the relation of *identity* and equation (13) straightforwardly reduces to the (solvable) equation

$$P = P. \tag{15}$$

Nevertheless, it is clear that this is an exceptional case and that, generally, self-referential situations as expressed by (13) will not yield solutions.[9] The paradoxical situation itself provides a case in point. The presence of the contrapredictive mechanism, as described by

$$\text{If } P \text{ is revealed, then } A = \textit{not-P} \tag{4}$$

implies that the functional dependence f of action A on prediction P is the one expressed by *negation*. Hence, the self-referential equation (13) becomes the unsolvable equation described earlier as

$$P = \textit{not-P}. \tag{6}$$

As these two examples make clear, self-referentiality, as functionally expressed by (13), can arise as the result of a *direct causal interaction* between the predicting subsystem S_1 and the predicted subsystem S_2. In our two examples, we have assumed that the revelation condition, (3), obtains and, thus, that S_2 somehow receives information about the prediction P after it has been made. In these situations, the functional dependence f of action A on prediction P can properly be understood as a description of *the causal impact the prediction has as a physical event*. In the paradoxical case, the causal impact of the prediction as a physical event is such that it inevitably falsifies the prediction itself.

However, such a direct causal interaction is not a necessary condition for the functional dependence, (12), to arise. As Scriven's version of the paradox makes clear (see Sect. 3), the existence of a *common causal history* can suffice. Remember that Scriven drops the revelation condition, (3), and replaces it with the assumption that S_2 is itself capable of calculating the prediction that is made by S_1. In our

[9] Alvin Goldman (1968, pp. 141–143) suggests that there are possible deterministic universes in which a full 'book of life' exists (containing the full and correct life story of, for instance, Alvin Goldman) and in which the person whose life is described actually reads this book, without thereby invalidating the predictions. Our equations imply, however, that the possible existence of such a book requires that all the self-referential situations generated by reading the book remain *solvable* and, thus, that several completely *contingent* constraints ('contingent' in the sense of being additional and independent from the deterministic laws of this universe) are met within this universe. The existence of such a book in some universe cannot be excluded a priori, but its existence would be a quirky exception without any philosophical significance for the problems dealt with in this paper.

description, this means that S_2 is capable of solving equation (8′) and to use this knowledge of P to act contrapredictively. This version of the paradox thus generates the same functional dependence of A on P as in our version:

$$A = not\text{-}P. \tag{5}$$

While in this case the functional dependence of A on P arises without any actual causal interaction after time t_1 between the prediction as a physical event and the action itself, this dependence now crucially turns upon the presence of a common causal history. In our account, this common history is encapsulated by the initial conditions U_0 such that both prediction P and action A causally depend on these conditions as expressed by equations (8′) and (9′), respectively.[10]

The presence of a common causal history is a *necessary* condition for Scriven's version of the paradox to work. This can be demonstrated by taking a look at the hypothetical situation in which Scriven's contrapredictive subsystem S_2 were to attempt to outsmart the external predictor by trying to calculate the prediction P* and then act contrapredictively with regard to this prediction as follows:

$$A = not\text{-}P^*. \tag{16}$$

In this case, the externality of the predictor implies that his prediction is not a physical event which causally depends on the initial conditions U_0 and that there is, consequently, no equivalent for equation (8′) from which S_2 could try to calculate P*. Instead, the only information about P* available to S_2 is contained in the predictive task of the external predictor, captured by the combination of equations (9) and (10) as

$$P^* = h(U_0). \tag{17}$$

However, since $h(U_0)$ is, by equation (9), identical to action A, the hypothetical scenario in which S_2 were to use (17) to determine P* and, subsequently, were to use this knowledge to act contrapredictively, as specified by (16), would as a result have an action which is the negation of itself:

$$A = not\text{-}A. \tag{18}$$

The fact that such an action cannot exist demonstrates that subsystem S_2 is unable to act contrapredictively with regard to an external observer with which it does not share a common causal history.

Although our account of the overdetermination involved in a case of embedded prediction has focused, thus far, on the category of more tractable situations in which action A is directly functionally dependent on prediction P (equation 12), the generality of the three equations (8′)–(9′)–(11) indicates that the problem of overdetermination might also be much more general. The laws of our universe seem to be such that all sorts of interactions and functional dependencies between two embedded subsystems unavoidably exist. Physics tells us, for instance, that two of

[10] The *commonness* of this history is an effect of the fact that the predicted subsystem S_2, in order to be capable of completing its calculation of P, needs to acquire sufficient information about the initial conditions which determine the future evolution of S_1. So, gathering the relevant information requires actual interactions—either directly or through intermediaries—between the two subsystems at some point in their pasts.

the fundamental forces in the universe (gravity and the electromagnetic force) have an infinite scope. This means that the actual physical workings of the predicting subsystem (containing biological or electronic devices which calculate and store the prediction) generate causal factors which will inevitably affect the future evolution of the predicted subsystem. Owing to this causal interconnectedness, one can expect that the set of equations ($8'$)–($9'$)–(11) will often generate problems of self-referentiality similar to those of equation (13).

In fact, the ubiquity of causal interactions seems to imply a reversal of our initial problem. Given that functional dependencies are almost inescapable, how then is it at all possible that we are actually capable of making embedded predictions of future events? The answer is, of course, that in many situations when we want to make predictions we are able to discount or shield of our own past and present interactions with the predicted system at hand. For instance, if we want to calculate the orbit of Jupiter and predict its position at a specific time in the future, we may safely discount the gravitational impact of our calculations (as a physical activity) on its orbit. In such a situation, the future event still remains functionally dependent on the present state of the universe, but will be, for all practical purposes, sufficiently *causally uncoupled* from our predicting activities. This uncoupling eliminates the self-referentiality of the situation and allows us to find a single solution that describes and predicts the future event.

Needless to say, however, that such an independence is never complete and only possible to a sufficient extent for some (but not all) specific 'local' problems and purposes. A general and complete causal uncoupling of our past and future activities from the predicted subsystem, would require, among other things, that we were capable of obtaining all the information needed for our predictions *without* actually disturbing the predicted system. This would require, what Popper (1950, p. 129) called, a 'one-way membrane' between the predictor and the predicted system. Because such a membrane cannot physically exist, the possibility of complete causal uncoupling is only available to a Laplacean predictor as a non-physical observer outside the universe.[11]

5 Take-It-or-Leave-It Control

In the 1960–1970 debate, it was widely assumed that the paradox of predictability was relevant for the debate on freedom and determinism. Especially Donald MacKay (1960, 1961, 1967, 1971, 1973) tried to use the paradox as the starting point for a compatibilist theory, according to which, even in a deterministic universe, human brain-states remain 'logically indeterminate' until the moment of 'free' choice. Although MacKay's approach is interesting in many respects, we believe that the works of Landsberg and Evans (1970), Evans and Landsberg (1972)

[11] The idea that observers necessarily causally influence the system they are observing is, of course, not novel. In the social sciences, for instance, this idea is known as the observer's paradox. However, our results support the additional and central claim that this influence seriously and, in some circumstances, irreparably undermines our ability to predict the future behavior of the observed system.

and Grünbaum (1971) have convincingly shown that MacKay ultimately failed in developing a plausible compatibilist position.[12]

We believe that this failure is not accidental but rather the effect of an underlying assumption which MacKay and most other authors in the earlier debate seemed to have shared. They all assumed a direct conceptual connection between freedom of human actions and their (embedded) unpredictability. We believe, on the contrary, that this direct connection fails to hold and that (embedded) unpredictability is neither a sufficient nor a necessary condition for human freedom. It is not sufficient, since, as we have noted, it is perfectly possible that 'soulless' machines programmed in a contrapredictive manner remain unpredictable. It also seems unreasonable to suggest that unpredictability is a necessary condition for freedom. If someone or something predicts that tomorrow I am going to help my neighbour fix the shed in the garden and if, as a matter of fact, I see no good reason not to do so, it seems implausible to suggest that my freedom to choose is undermined simply because the prediction is indeed in line with my personality and my reasons for action.

While these considerations imply that no direct inference from the lack of embedded predictability to the possibility of human freedom can be drawn, we conjecture that the paradox of predictability leads to results which might be relevant for the compatibilism-incompatibilism debate on free will. Introducing some basic terminology will be helpful.[13] If determinism, as characterized by (1) in Sect. 2, is true, the facts of the past, together with the laws of nature, entail all facts of the present and future. Indeterminism is the denial of determinism. Compatibilism is the view that free will, free action and moral responsibility are compatible with determinism; incompatibilism is the denial of compatibilism. Libertarians are incompatibilists (and indeterminists) who believe that at least some of us, at times, perform free actions for which we are responsible. Hard determinists are incompatibilists who deny that we have the sort of free will required for moral responsibility; hard incompatibilists deny the same, irrespective of accepting determinism or not. Incompatibilists standardly hold that responsibility-relevant freedom can be nothing less than regulative control (the sort of control which involves the existence of genuinely open alternative possibilities), whereas compatibilists argue that such freedom only requires guidance control. At a minimum, our conjecture pertains to guidance control.[14]

[12] MacKay (1960, 1967, 1971) argued in favor of a notion of truth, according to which predictions are only true if they are *revelation-robust* as well as *belief-robust*. This means that (1) the truth value of propositions describing a predicted future state of a system should not depend on whether or not a physical interaction revealing the prediction has taken place between the predicting and the predicted system, and that (2) this truth value should also not depend on whether or not the predicted system believes that the predicted state will actually be realized. Since these assumptions imply that a future state of affairs (the predicted event or action) should not depend upon prior physical events and states of affairs with which it is actually causally connected, this notion of truth seems highly implausible. Indeed, this notion of truth implies, as MacKay specifically intended, that no propositions describing future human actions have a truth value. As Evans and Landsberg (1972) demonstrated, this inflated notion of truth begs all the relevant questions regarding compatibilism. Compare also Honderich (1988, pp. 347–350).

[13] For more, excellent information on this debate and its standard terminology, see Kane (2002, 2005).

[14] This distinction between regulative and guidance control, introduced by John Martin Fischer, has become well known; see Fischer (1994, pp. 131–189). We surmise that our suggestion here can be elaborated and possibly be integrated with Fisher's compatibilist model of free agency as reason-responsiveness (Fischer and Ravizza 1998; Fischer 2006).

To briefly substantiate our suggestion that the paradox of predictability might (at least) be relevant for the compatibilist notion of guidance control, we propose to return for a moment to the self-referential situations discussed in the previous section. Remember that we introduced both a contrapredictive mechanism, characterized by (4), which always goes against revealed predictions and whose outcome is therefore always unpredictable, and a fatalist mechanism, characterized by (14), which always conforms to a revealed prediction and whose outcome is therefore always predictable. Relevant for our purposes is that it is furthermore possible to consider a series of more complex intermediary situations with mechanisms which reject a revealed prediction when other *sufficient conditions for rejection* (*SCR*) hold true and are also revealed:

$$\text{If } P \text{ is revealed and if SCR are revealed, then } A = \textit{not-P}. \tag{19}$$

Situations in which this type of *condition-responsive mechanism* operates in a subsystem S_2, the predictability of action A depends on the additional *SCR*-conditions. If these are in fact revealed to hold, then the embedded predictor S_1 faces the by now familiar contrapredictive self-referential and unsolvable equation

$$P = \textit{not-P}. \tag{6}$$

As explained, this means that S_1's embedded prediction will be inevitably self-defeating. Yet, in case the *SCR*-conditions are not revealed or fail to obtain, a solution for the pertinent equations is not necessarily excluded and, depending on the details of the situation, an adequate prediction might remain possible.

The presupposition that we actually live in a deterministic universe is compatible with the additional and, in our view, plausible assumption that human beings are in possession of (deterministic) psychological mechanisms of the condition-responsive kind, indicated by (19). This means that, even in a deterministic universe, a human being who is confronted with a prediction about its own future behaviour might consider several other conditions and, depending on whether or not these hold, decide to go against or, alternatively, decide to act in accordance with the prediction made. We illustrate the types of situation we have in mind by considering two possible sufficient conditions for rejection. Suppose, first, that the subsystem which predicts our actions does so on the basis of its foreknowledge for the purpose of manipulating us. In this case, if the prediction and the manipulative intentions of the predictor are revealed, we could be motivated to act contrapredictively. Similarly, it could turn out that a prediction of my future behaviour is based on the discovery of some kind of biological or sociological mechanism which determines my actions. In this second case, such a discovery could lead me to reassess what I previously considered as my authentic preferences and my authentic reasons for making my decisions. As a result, I could decide, on the basis of the more extensive information now available to me, that I no longer wish to pursue my biologically or socially determined urges and desires.[15] If, as in these two examples, manipulation or the

[15] For sure, whether or not my condition-responsive mechanisms will be able to override the underlying biological or social mechanisms (whether my thus determined urges and desires are 'controllable' or not) depends on the contingent causal connections between these psychological, biological and social mechanisms realized in my central nervous system that controls my movements. The answers here seem

presence of inauthentic preferences is interpreted as a *SCR*-condition, and if this condition is revealed, then the working of (deterministically) operating condition-responsive psychological mechanisms warrants that the predictions made about our future behaviour will inevitably be self-defeating.

Our analysis of the paradox demonstrates that embedded subsystems in a deterministic universe retain, what we call, a *take-it-or-leave-it control* as regards revealed predictions. The specifics of the operating mechanisms of such an embedded subsystem establish the conditions under which it will accept or reject predictions about its future behaviour made by an other embedded subsystem in the same universe. Whether or not this consequence will turn out to be important for the contemporary debate on the compatibility of human freedom with determinism depends on the additional assumption that take-it-or-leave-it control is an essential part of a convincing analysis of free human agency. We conjecture that this assumption holds water, but a further elaboration of such an analysis is beyond this paper's scope.

Acknowledgments Stefan Rummens' research contained in this paper was made possible by a postdoctoral fellowship of the Research Foundation Flanders (FWO). Stefaan Cuypers' research connects with the research project *Alternatives, Belief and Action* (FFI2009-09686), awarded and supported by the Spanish Ministry of Science and Innovation. We are grateful to the FWO *Odysseus* formal epistemology research group of Igor Douven (KU Leuven) for useful discussion. We also thank three anonymous referees for *Erkenntnis* for their helpful suggestions.

References

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

Boyd, R. (1972). Determinism, laws, and predictability in principle. *Philosophy of Science, 39*(4), 431–450.

Brehm, J. W. (1966). *A theory of psychological reactance*. New York: Academic Press.

DeWitt, L. W. (1973). The hidden assumption in MacKay's logical paradox concerning free will. *The British Journal for the Philosophy of Science, 24*(4), 402–405.

Dieks, D. (1980). Discussion: On the empirical content of determinism. *Philosophy of Science, 47*(1), 124–130.

Evans, D. A., & Landsberg, P. T. (1972). Free will in a mechanistic universe? An extension. *The British Journal for the Philosophy of Science, 23*(4), 336–343.

Fischer, J. M. (1994). *The metaphysics of free will*. Oxford: Blackwell.

Fischer, J. M. (2006). *My way. Essays on moral responsibility*. Oxford: Oxford University Press.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control. A theory of moral responsibility*. Cambridge: Cambridge University Press.

Goldman, A. I. (1968). Actions, predictions and books of life. *American Philosophical Quarterly, 5*(3), 135–151.

Footnote 15 continued
to depend on empirical issues relating to the functioning of our brain. There is, however, nothing particularly mysterious or metaphysical about the assumption that a physically realized psychological mechanism could override (or could learn to override) other physically realized mechanisms.

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

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

Hempel, C. G. (1961). Some reflections on "the case for determinism". In S. Hook (Ed.), *Determinism and freedom in the age of modern science* (pp. 170–175). New York: Collier Books.

Honderich, T. (1988). *A theory of determinism. The mind, neuroscience, and life-hopes*. Oxford: Clarendon Press.

Kane, R. (Ed.). (2002). *The Oxford handbook of free will*. New York: Oxford University Press.

Kane, R. (2005). *A contemporary introduction to free will*. New York: Oxford University Press.

Kukla, A. (1978). Discussion: On the empirical significance of pure determinism. *Philosophy of Science, 45*(1), 141–144.

Kukla, A. (1980). Discussion: Determinism and predictability: Reply to Dieks. *Philosophy of Science, 47*(1), 131–133.

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

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

MacKay, D. M. (1960). On the logical indeterminacy of a free choice. *Mind, 69*, 31–40.

MacKay, D. M. (1961). Logical indeterminacy and freewill. *Analysis, 21*(4), 82–83.

MacKay, D. M. (1967). Freedom of action in a mechanistic universe. In M. S. Gazzaniga & E. P. Lovejoy (Eds.), *Good reading in psychology*, (pp. 121–137). Englewood Cliffs: Prentice-Hall, 1971; originally id. (1967), *Freedom of action in a mechanistic universe: Eddington memorial lecture of November 1967*. Cambridge: Cambridge University Press.

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.

MacKay, D. M. (1973). The logical indeterminateness of human choices. *The British Journal for the Philosophy of Science, 24*(4), 405–408.

Nagel, T. (1986). *The view from nowhere*. New York: Oxford University Press.

Popper, K. (1950). Indeterminism in quantum physics and in classical physics. *The British Journal for the Philosophy of Science, 1*(2/3), 117–133. see also 173–195.

Roberts, L. D. (1975). Scriven and MacKay on unpredictability and free choice. *Mind, 84*, 284–288.

Russell, B. (1917). On the notion of cause. In id., *Mysticism and logic* (pp. 132–151). London: Unwin, 1963.

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

Sellars, W. (1966). Fatalism and determinism. In K. Lehrer (Ed.), *Freedom and determinism* (pp. 141–174). New York: Random House.

Suppes, P. (1964). On an example of unpredictability in human behaviour. *Philosophy of Science, 31*(2), 143–148.

Suppes, P. (1985). Explaining the unpredictable. *Erkenntnis, 22*, 187–195.

Wormell, C. P. (1958). On the paradoxes of self-reference. *Mind, 67*(266), 267–271.