Unpredictability: A Reply to Cargile and to Benditt and Ross
Author(s): George Schlesinger
Source: *The British Journal for the Philosophy of Science*, Vol. 27, No. 3 (Sep., 1976), pp. 267-274
Published by: Oxford University Press on behalf of The British Society for the Philosophy of Science
Stable URL: http://www.jstor.org/stable/686125
Accessed: 20-08-2014 03:29 UTC

PALACIOS, J. [1957]: 'Revision de la Teoria de la Relatividad', *Revista de la Real Academia de Ciencias Exactas Fisicas y Natureles de Madrid*, **51,** pp. 56–73, 165–73.
PALACIOS, J. [1961]: 'A Reappraisal of the Principle of Relativity as Applied to Moving Interferometers', *Revista de la Real Academia de Ciencias Exactas Fisicas y Naturales de Madrid*, **55,** pp. 201–7.
PODLAHA, M. [1969]: 'Lorentz Theory, Palacios Theory and Interferometrical Experiments', *Il Nuovo Cimento*, **64B**, pp. 181–7.
PODLAHA, M. [1971]: 'Reply to Shamir's Paper: New Experimental Test on Special Relativity', *Lettere al Nuovo Cimento*, **2,** pp. 601–2.
PODLAHA, M. [1974*a*]: 'Uberlegungen zum Verhältnis von Newtonischen Physik zur Spezielten Relativitätstheorie', *Report D.F.G.*, University of Göttingen.
PODLAHA, M. [1974*b*]: 'The Axiomatic Foundations of the Theory of Special Relativity—Reply to Stiegler', *International Journal of Theoretical Physics*, **11,** pp. 69–72.
PODLAHA, M. [1974*c*]: 'Analysis of Invariance of Maxwell's Equations with Respect to Assymmetrical Linear Transformations of Voigt–Palacios–Gordon Type', *Indian Journal of Theoretical Physics*, **22,** pp. 73–9.
PODLAHA, M. [1975*a*]: 'Length Contraction and Time Dilatation in the Special Theory of Relativity—Real or Apparent Phenomena?', *Indian Journal of Theoretical Physics*, **23,** pp. 69–75.
PODLAHA, M. [1975*b*]: 'An Explanation of the Lorentz Transformation by Non-Relativistic Premises', *Report D.F.G.*, University of Göttingen.
PODLAHA, M., ALTHAUS, B. and BENSCH, H. [1975]: 'Zur Problematik der Geschwindigkeitsmessung', *Philosophia Naturalis*, in print.
ROBERTSON, H. P. [1949]: 'Postulate versus Observation in the Special Theory of Relativity', *Review of Modern Physics*, **21**, pp. 378–82.
SCHAFFNER, K. F. [1974]: 'Einstein Versus Lorentz: Research Programmes and the Logic of Comparative Theory Evaluation', *The British Journal for the Philosophy of Science*, **25.** pp. 45–78.
SHAMIR, J. and FOX, R. [1969]: 'A New Experimental Test of Special Relativity', *Il Nuovo Cimento*, **62B**, pp. 258–64.
STRNAD, J. [1970]: 'A Note on the Trouton–Noble Experiment', *Contemporary Physics*, **11,** pp. 59–64.
TRIMMEL, W. S. N., BAIERLEIN, R. F., FALLER, J. E. and HILL, H. A. [1973]: 'Experimental Search for Anisotropy in the Speed of Light', *Physical Review D2*, **8,** p. 3321.
ZAHAR, E. [1973]: 'Why did Einstein's Programme Supersede Lorentz's?', *The British Journal for the Philosophy of Science*, **24,** pp. 95–123, 223–62.

UNPREDICTABILITY: A REPLY TO CARGILE AND TO BENDITT AND ROSS

1 My 'Unpredictability of Free Choices'[1] has given rise to much more than the normal amount of opposition. When so many philosophers feel that something is faulty with a given thesis, then whether or not they can articulate what the fault is and even if the arguments they marshal against it are rather shaky, it could be a sign that they are correct in sensing a deep defect in the reasoning. On the other hand, it could also simply be my conclusion which so upsets people, since they fear it may lead to a picture of human nature repugnant to them. Although I am loth to upset people, I believe that the latter rather than the former is the case, and so I shall argue in defense of my position.

2 In this note I shall concentrate mainly on answering the attempted attacks of Cargile,[2] and Benditt and Ross.[3] They all agree on one thing: my argument

[1] Schlesinger [1974]. [2] Cargile [1975]. [3] Benditt and Ross [1975].

that it is best to take both boxes because this is in accordance with what would be the advice of a sufficiently intelligent and well-informed well-wisher, is wrong. Now before attempting to demonstrate that they have produced no valid arguments to support their position, I should like to show that even if my 'well-wisher argument' was completely fallacious there still would be no grounds for their contention that the agent should fare best if he chooses to take box II only.

What are the alleged reasons for claiming that in order to maximise his gain the agent should take box II only? Supposedly we are given an infallible Predictor who if he predicts at t_0—twenty-four hours before t_1—that at t_1 the agent takes box II only, he places $\$M$ in that box, otherwise he leaves it empty. Now how can we in practice be *given* that someone is an infallible predictor in the future as well as in the past? The answer must be that what we could be given is very strong inductive evidence that this is so. What would be the nature of such evidence? It could be suggested for instance that the Predictor in Newcomb's story has played the game one million times before and in the 500,000 cases in which players with widely different backgrounds and temperaments using a great variety of arguments have chosen to take box II only as well as in the 500,000 cases in which they ended up taking both boxes, has without a single exception correctly anticipated their choices. Do we then indeed have strong inductive evidence that we are faced with a Predictor with whose powers the agent must reckon and thus restrain himself and refrain from taking box I so as to ensure, or to say the least increase the chances, that he finds $\$M$ in box II? Not at all! Let me explain.

It is well known that in empirical reasoning it is not the case that it depends exclusively on the nature of a given piece of evidence and on the nature of a given hypothesis whether the former supports the latter. But whether a certain observation confirms a given hypothesis depends also on the theories we hold and in the context of which the confirmation is supposed to take place. I shall presently describe a theory which is by no means absurd and in the context of which the spectacular success of the Predictor provides no indication whatever that it may be better for the agent to choose box II only.

Let: C_1 = The agent chooses to take box II only.
C_2 = The agent chooses to take both boxes.
P_1 = The Predictor predicts that C_1.
P_2 = The Predictor predicts that C_2.
T_1 = The agent has a tendency to do C_1.
T_2 = The agent has a tendency to do C_2.
D_1 = The Predictor diagnoses that T_1.
D_2 = The Predictor diagnoses that T_2.

Now it is maintained that at least 24 hours before anyone makes a choice he has a distinct, and in principle discernible, tendency to end up making that choice. Within that period of time the agent cannot change the tendency he has ingrained in him but by an act of free will he can behave contrary to his tendency. It is also asserted that:

$$T_1 \leftrightarrow D_1 \quad \text{and} \quad T_2 \leftrightarrow D_2$$

which means that our co-called 'Predictor' is really a perfect diagnostician of tendencies and at t_0 can without fail recognise the tendency the agent has to choose at t_1.

Also:

$$D_1 \leftrightarrow P_1 \quad \text{and} \quad D_2 \leftrightarrow P_2$$

i.e. the 'Predictor' always bases his prediction of what the agent will actually choose on his correct diagnosis of what tendency he has to choose. From this it follows of course that:

$$T_1 \leftrightarrow P_1\,(\alpha) \quad \text{and} \quad T_2 \leftrightarrow P_2\,(\alpha')$$

There is no reason why we should not also maintain that:

$$p(C_1/T_1) = p(C_2/T_2) = 0.99\ldots99$$

i.e. that the probability of the agent doing C_i when it is a fact that T_i is extremely high; the vast majority of people who have the tendency to make a given choice end up making that choice. Also, of course:

$$p(C_1/T_2) = p(C_2/T_1) = 0.00\ldots01$$

because very few agents act contrary to their tendency.

In order to represent matters graphically, let me assert that any player who plays Newcomb's game is in one of the following four classes:

	T_1	T_2
C_1	a: very large $p(P_1) = 1$	c: very small $p(P_1) = 0$
C_2	b: very small $p(P_1) = 1$	d: very large $p(P_1) = 0$

Since class a is much larger than class c, then, if all we are given is that our agent is in a class characterised by C_1 then we know he is much more likely to be in a than in c or characterised by T_1 rather than by T_2 or (because of (α) and (α') by P_1 rather than P_2). Also when we are given no more than that the agent is in class C_2 then he is much more likely to be in d than in b. It follows therefore that:

$$p(T_1/C_1) \gg p(P_2/C_1) \quad \text{and} \quad p(P_2/C_2) \gg p(P_1/C_2)$$

Thus the probability that the Predictor will make the right prediction is vastly greater than that he will err in his prediction. This fits very well with our past observation of the Predictor's performance.

We observe however, that vertical movement only is possible for the agent, that is, he can move from a to b or b to a and from c to d or d *to* c but not, for instance from c to a or d to a. Thus if it is the case that the agent is in the class characterised by T_2 then by going against his tendency and doing C_1, all he can achieve to transfer himself from d to c—and not from d to a—which is of course of no use to him at all since the Predictor who always bases his prediction on his diagnosis of the tendency the agent has, will still predict that C_2 and leave box II empty. Thus he will give up the \$1,000 of box I without in the slightest increasing his chances for receiving $\$M$. Also, if it is a fact that the agent is in a class by T_1 then by actually doing C_2 he does not in the least jeopardise his

chances of getting the $\$M$ and he can only gain \$1,000 by transferring himself from *a* to *b*—he cannot transfer himself from *a* to *d*.

The basis for claiming in the first place that experience supports the contention that it is best to do C_1 has disappeared. Past experience is now interpreted in a manner that it no longer provides any grounds for the agent to say 'I had better be careful and do C_1 so as to increase my chances to receive $\$M$'. And it is not because the perfect record of the 'Predictor' is put down as an incredible coincidence that we say that the agent may ignore the 'Predictor'. This was only necessary as long as we were not aware of the possibility of a theory according to which the 'Predictor's' success is indicative only of his proficiency as a diagnostician. Once this is brought to our attention we realise that it is due to the extremely high correlation between T_i and C_i that he manages to be a predictor at all but actually he has no direct access to the final choices.

3 Now we shall look upon the positive reason why on the pain of a contradiction only, could we maintain that the 'Predictor' is skilful *qua* predictor. I argued because that would imply that it is best to do C_1 while the 'well-wisher argument' implies that it is best to do C_2. This is basically different from the 'dominance argument' which has been employed by some authors to support their contention that it is best to do C_2. Cargile sounds somewhat sceptical and is not entirely convinced that the two arguments are fundamentally different. In my original paper I showed that the two arguments must be different by describing what I called Game 2, in which an Observer replaces the Predictor and in which it is clear beyond doubt that it is best to do C_1. Yet the 'dominance argument' applies equally well here as in Game 1 and would force upon us the obviously wrong conclusion that it is best to do C_2, while the 'well-wisher argument' cannot be applied to this game.

It is not enough however, merely to show that the two arguments are entirely different and that the 'dominance argument' *must be* erroneous. It is also very important for the sake of avoiding Cargile's consequent confusion to show *how* the two arguments differ and *why* exactly the 'dominance argument' is wrong. Let me try.

Box II of course may be in one of two states: E (empty) and F (full). The 'dominance argument' correctly reasons that the agent is better off both relative to F and relative to E if he takes both boxes and gains the money contained in the first box. It, however, happens to be a fact that to be worse off relative to F is preferable than to be better off relative to E. If we assume the Predictor to be infallible, then the only two available outcomes are 'worse off F' or 'better off E' and it is up to the agent to choose between these two. There is nothing in the 'dominance argument' which denies this. Thus, suppose the agent asks himself 'Am I not increasing my chances to find $\$M$ in box II if I restrain myself and refrain from taking box I as well?' There is nothing in the 'dominance argument' which would indicate that he is not increasing his chances of finding money in box II by doing C_1. It supports no more than the contention that should the agent do C_1 then he is to lose the \$1,000 of box I, while doing C_2 will gain him the \$1,000 contained in box I. This in no way excludes the possibility that nevertheless by doing C_1, rather than C_2, that the agent is better off. It is obvious therefore that without violating the 'dominance argument' we may advise the agent to go for the 'worse off F' and do C_1.

It is entirely different with the 'well-wisher argument'. To the well-wisher we could put the explicit question: is there any point in taking box II only, rather than both boxes? Does the agent in any way lessen his chances for finding $\$M$ in box II by doing C_2 rather than C_1? We can work out with absolute certainty what his answer would be: there is nothing to gain by doing C_1, the agent cannot in the least increase his chances for obtaining $\$M$ by restraining himself and taking box II only (for his chance is already 100 per cent or in any case will remain 0). And of course, on the assumption that between t_0 and t_1 the state of box II does not undergo any changes, he must be right. We are forced therefore to withdraw the assumption that the predictor is competent, an assumption which leads to an answer conflicting with that of the person who is in a perfect position to judge the issue. From the 'dominance argument' we cannot extract any answer to this question. It therefore cannot force us to revise our assumption that the agent has reason to be concerned that his actual choice may affect the contents of box II.

4 We are in a position now to look at Cargile's objection in detail. He claims that my contention that the agent must follow the advice of the well-wisher ignores the fact that:

> . . . the observer may advise this (*i.e.* to do C_2) for two different reasons: (1) at least you'll get a thousand and (2) you might as well get an extra thousand. If the player subscribes to a theory about the predictor according to which whether it is (1) or (2) depends on his choice, the fact that the observer will always advise C_2 may be irrelevant.

It is indeed true that the competent judge who we know holds the opinion that, in order to maximise his gain, the agent ought to do C_2 may think so either because he sees that there is anyhow no money in box II or because he is already assured that no matter what the agent does, he is going to win the $\$M$ which is in box II, and we do not know which. But if we say that by doing C_2 the agent may bring about, or even just may raise the probability that the observer's deeming C_2 the better choice is because of P_2 rather than P_1, then we are forced to the conclusion that he ought to refrain from C_1. But this is contrary to the views of the one person who is entirely competent to judge this issue and whose opinion is therefore not irrelevant but completely binding.

Or, to put it differently, the agent may raise the question: The well-wisher advises me to do C_2. But by actually doing C_2 am I not raising the probability that he advises me thus because he sees that box II is empty rather than because he sees that there is $\$M$ in it anyhow? But this very question may be put to the well-wisher and we know for sure what his answer would be: No, by doing C_2 rather than C_1 all you do is raise by 100 per cent the probability of winning the $1,000 contained in box I but affect in no way the probability of there being $\$M$ in box II.

Cargile seems to sense the untenability of his position and tries to turn around in the right direction by saying:

> It might be replied that once the sympathetic observer checks the boxes, it is too late for the predictor to do anything, so the fact that the well-wisher will necessarily advise C_2 cannot be irrelevant.

But he immediately sinks into an unfortunate confusion:

> But this is just to reject the possibility of probabilistic reverse causality which is implied in accepting probabilistic independence. This may be 'reasonable', but then, given this attitude, there is no conflicting strategy, and it is not necessarily true that this attitude is correct.

But of course 'this attitude' is not regarded as '*given*'. To begin with it is suggested that we go along with Newcomb's story and definitely assume that we are confronted with a case of some 'sort of probabilistic reverse causality', *i.e.* that the agent's choice at t_1 influences the Predictor's action at t_0. But after it has been demonstrated to us that ultimately this assumption leads to the conclusion that C_1 maximises the agent's gain, which conflicts with the conclusion following the 'well-wisher argument' that it is rather C_2 which does so, we are forced to withdraw our original assumption. Thus our proof is essentially a *reductio ad absurdum* proof. We begin by postulating that the Predictor of our story is competent and demonstrate that the story leads to a contradiction. In order to remedy the situation we reconsider matters realising that one of the assumptions incorporated in our story must have been wrong. This leads us to the conclusion that the wrong assumption was that the Predictor was competent.

5 Benditt and Ross say that there are three major points in my argument:

> First, there is the claim that the well-wisher would wish the player to take both boxes. Second, there is the claim that whatever he wishes the player to do is what is in the player's best interest (this is supposed to be analytic). And third, Schlesinger supposes that what is in the player's best interest is what is rational for him to do.

They are quite correct in saying that I subscribe to the first two points. But of course instead of their third point, I should substitute 'What is *known* to the player to be in his best interest is what is rational (in the sense that it will most promote his best interest) for him to do'. This obviously renders their 'counter-example' entirely irrelevant since in the 'simple game' they have devised the agent does *not* know that the well-wisher who is well-informed about the contents of box A and B is of the opinion that he should take box B. Hence he has nothing to go by but the probability judgments available to him according to which taking A is most likely to gain him the money. In our case however the agent knows with absolute certainty that the well-informed onlooker thinks that the agent maximises his gain by taking both boxes. Thus if the agent wants to maximise his gain, he must act in accordance with the former's judgment. Consequently it is wholly pointless to go on arguing as they do that:

> . . . we can and must distinguish between (*a*) what I should do in the sense of what rational policy I should follow and (*b*) what I should do in the sense of what will in fact attain the most desirable result for me.

In a situation like ours where the agent can work out exactly what policy the fully competent judge thinks will bring him the most desirable result, there is no distinction between this policy and the one he ought to follow if he wants to bring about the most desirable result.

6 Contrary to the claims of Benditt and Ross, the fact that the agent knows that in the opinion of the well-informed and intelligent onlooker's judgment it is best to do C_2 is sufficient for the agent to be fully assured that to do C_2 must be the most desirable thing for him to do even though the agent does not know why the well-wisher has this opinion.

Consider the following game: There are two boxes X and Y and a well shuffled pack of ordinary cards. One card is drawn at random and placed in box Y and the other 51 cards in box X. Two dice are rolled and if a double-six occurs the agent is to get a million dollars if he points his finger at the box which does not contain the ace of spades; otherwise he gets a million dollars if he points his finger at the box which does contain the ace of spades. The agent does not know which card has been put in box Y nor the result of rolling the dice and is invited to point a finger at one of the two boxes. The probability that by pointing at box X the agent will get a million dollars is at least $35/36 \times 51/52$ since the probability that the dice showed not a double-six is $35/36$ (and in which case he has to point at the box which does contain the ace of spades) and the probability that the ace of spades is in box X is $51/52$. In other words the probability that by pointing at box X he will win a million dollars is more than 95 per cent.

Now let us suppose that a friend who is an absolutely perfect well-wisher is allowed either to observe the fall of the dice or to examine the contents of the two boxes but not both. The agent knows this with absolute certainty but has no idea which of the two acts his friend was allowed to do. Let us also suppose that the friend is permitted to communicate with the agent but no more than to advise him at which box to point. Consider the case in which the friend says to the agent 'Point at Y'. It is clear that he may have given this advice for two different reasons. Either because he saw that the improbable has happened and both dice showed a six in which case he wins $\$M$ if he succeeds in pointing at the box not containing the ace of spades and there is a probability of $51/52$ that the ace of spades is not in Y, or because he has seen the contents of Y and knows that the single card in it turned out to be the ace of spades in which case there is a probability of $35/36$ that he will gain $\$M$ by pointing at it since the probability of not having a double-six is $35/36$ which is the probability that he wins $\$M$ by pointing at the box that does contain the ace of spades. There is not a shadow of doubt that it is rational for the agent to point at Y. It matters not in the least that he has no idea for which of the two possible reasons his friend has advised him to point at Y, the mere fact that he advised him thus suffices to ensure that by pointing at Y the agent maximises his chances to win the money. Thus it is quite pointless for Benditt and Ross to say:

> . . . in the Newcomb's game . . . what choice one's friend would have one to make does not provide any information about what the situation is.

The agent need not know 'what the situation is', *i.e.* on what information does his friend base his opinion that it is best to do C_2. It is enough for him to know that the final verdict of a perfectly competent judge is that by doing C_1 he can gain absolutely nothing. Doing C_2 must therefore be his best choice.

GEORGE SCHLESINGER
The University of North Carolina at Chapel Hill

REFERENCES

BENDIT, T. M. and ROSS, D. J. [1975]: 'Newcomb's Paradox', *British Journal for the Philosophy of Science*, **27,** pp. 161–4.
CARGILE, J. [1975]: 'Newcomb's Problem', *British Journal for the Philosophy of Science*, **26,** pp. 234–9.
SCHLESINGER, G. [1974]: 'The Unpredictability of Free Choices', *British Journal for the Philosophy of Science*, **25,** pp. 209–221.