Jean-Pierre Dupuy

Two temporalities, two rationalities: a new look at Newcomb's paradox

Jacques: ...however reluctantly, I always come back to what my Captain used to say: "Everything which happens to us in this world, good or bad, is written up above..." Do you, Monsieur, know any way of erasing this writing? ...
(.....)
The Master: I am wondering about something... that is whether your benefactor would have been cuckolded because it was written up above or whether it was written up above because you cuckolded your benefactor?
Jacques: The two were written side by side. Everything was written at the same time. It is like a great scroll which is unrolled little by little.
You can imagine, Reader, to what lengths I might take this conversation on a subject which has been talked about and written about so much for the last two thousand years without getting one step further forward. If you are not grateful to me for what I am telling you, be very grateful for what I am not telling you.
(.....)
Jacques: ...it would have to be written on the scroll that Jacques would break his neck on such a day and Jacques would not break his neck. Can you imagine for a moment that that could happen, whoever made the great scroll?
The Master: There are a number of things one could say about that...

Diderot, *Jacques the Fatalist*
(1986, pp. 25-6, 30)

Abstract: Several cases of alleged irrational behavior are examined: imitation of others in a situation of uncertainly; non-rational revision of belief in a case of cognitive dissonance; sunk cost fallacy; weakness of` the will. A general characterization of this class of behavior is provided: the agent endowing herself with a power over the past and observing herself as from the outside. These two lectures are shown to characterize the evidentialist choice in Common Cause Newcomb Problems, from Fisher's smoking case to Max Weber's paradox. The rationality of evidentialism is nevertheless advocated. It is furthermore shown that the Backwards Induction Paradox is a Newcomb problem. The rationality and possibility of reciprocal exchange in a case of non-credible promises follows.

Newcomb's paradox[1]

Imagine two boxes. One is transparent and contains a thousand dollars; the other is opaque and contains either a million dollars or nothing at all. The choice of the agent is either B_1: to take only what is in the opaque box, or B_2: to take what is in both boxes. At the time that the agent is presented with this problem, a Predictor has already placed a million dollars in the opaque box if and only if he foresaw that the agent would choose B_1. The agent knows all this, and he has very high confidence in the predictive powers of the Predictor. What should he do?

A first line of reasoning leads to the conclusion that the agent should choose B_1. The Predictor will have foreseen it and the agent will have a million dollars. If he chose B_2, he would only have a thousand. The paradox is that a second line of reasoning appears to lead just as surely to the opposite conclusion. When the agent makes his choice, there is or there is not a million dollars in the opaque box: by taking both boxes, he will obviously get a thousand dollars more in either case.

When people are presented with this problem, no consensus is reached around either solution. Professional philosophers and theoreticians are no more able to agree than anyone else. They seem to fall into the following categories:

- the "no-boxers," those who refuse to choose because they deem the problem incoherent, poorly formulated or insufficiently precise;
- invariant two-boxers, who choose B_2 unconditionally;
- conditional one-boxers who choose B_1 if the confidence to be accorded the Predictor is total;
- those who are either one-boxers or two-boxers, depending on how the problem is specified, the decisive parameter most often being the degree of confidence in the predictive ability of the Predictor. Within this category, there is not even agreement on the critical threshold at which the choice must shift from B_1 to B_2. For some, it is the fallibility or infallibility of the Predictor that makes the difference: it is in the latter case, and only in the latter case, that it is rational to be a one-boxer. Others calculate the threshold by maximizing expected utility--with the expected utility being calculated on the basis of the conditional probabilities that the Predictor did or did not put the million dollars in the opaque box for each possible choice.

In spite of this cacophany, there is a general consensus that Newcomb's problem illustrates in spectacular fashion the possibility of a conflict between two modes of reasoning. The invariant two-boxers reason in terms of a *dominant strategy:* whichever prediction was made, and whichever action taken as a result by the Predictor, more is to be gained by taking two boxes than by taking only one. Those who are one-boxers as soon as they have enough confidence in the predictive ability of the Predictor reason by *maximizing expected utility*. If there is a conflict between these two modes of reasoning, that is because the state of the world (i. e. the presence or absence of the million dollars in the opaque box) depends on the decision probabilistically, but does not depend on it causally (since the determination of the state *precedes* the decision). It is agreed that if there were causal dependence, there would be no basis for dominance reasoning. But the combination of *probabilistic dependence* and *causal independence*, which characterizes Newcomb's problem, does not permit a clear decision in favor of one mode of reasoning over the other.

In my judgment the current state of the controversy is unsatisfying. The positions staked out thus far do not account for all the possibilities, any more than does the distinction between probabilistic and causal dependence. The position that I will defend consists in considering that the solutions B_1 and B_2 are both perfectly legitimate, but that they correspond to two different, albeit inseparable, conceptions of time: projected time and occurring time. In addition, there exists another form of dependence, counterfactual dependence, which, even though it is compatible with causal independence, has the same effect as causal dependence: it eliminates the basis for dominance reasoning. In projected time, the decision between B_1 and B_2 does not, to be sure, cause the state of the world, but it produces it counterfactually, so that one can no longer treat it as "fixed" in relation to the decision, as one does when reasoning in terms of a dominant strategy.[2]

To say that B_1 and B_2 are both legitimate does not amount to saying once again that sometimes one must be chosen and sometimes the other, depending on how the problem is specified or how it is interpreted. The two choices correspond to two forms of rationality that are irreducible to one another and that Newcomb's problem puts into conflict. To make this point as clear as possible, I will delineate the specifics of the problem for only one case: that in which the Predictor is infallible. It is in this case that the conflict is sharpest between the arguments for a dominant strategy and for maximizing expected utility.

Divine foreknowledge and free will.

> ...the great scroll which contains the truth, the
> whole truth and nothing but the truth.
>
> *Jacques the Fatalist* (p. 30)

If the Predictor is infallible, a question immediately arises: is it not inconsistent to suppose at the same time that the agent is able to choose freely? Here we meet up with a very old philosophical question that, rather surprisingly, has recently been revived by analytical philosophers seeking to clarify and to systematize its terms: the problem of the compatibility or incompatibility of divine foreknowledge and human freedom.[3]

I will begin by recalling the "incompatibilist" thesis and presenting three arguments for it. The first one is incorrect, but instructive, the second is the classic formulation, and the third is the argument advanced by present-day incompatibilists.

Let us start by stipulating what will be meant by "God" in what follows. "God" is a proper name that designates a person who possesses in essential fashion the divine attributes. We will limit ourselves to two of these: God is eternal, and He is omniscient. To say that He is eternal is to say that He exists at all times in our temporal framework; to say that He is omniscient is to say that He *believes* to be true all propositions that are true and only those that are true. God possesses these properties *essentially*: that means that in all possible worlds in which He exists, He possesses these properties.

To these definitions must be added two important hypotheses, which we will adopt without question even if some philosophers of note have rejected them:
a) Propositions can be true (or false) at a given time;
b) "future contingent propositions" (that is, propositions relative to free actions taking place in the future) are, like any other propositions, either true or false. "I will present this paper at a symposium a month from now" is true today if and only if it will be true a month from now that I present this paper at a symposium. This hypothesis simply expresses the idea that the future is what it will be, that it is real. It would obviously be a mistake to assume that this hypothesis by itself rules out free will. If there is free will at work in the world, the future *could* be different from what it will be. It is still true that what it will be is in some sense already there. In this sense, the future is no less unalterable than the past.

Let us suppose that a subject S performs an act X at time t_2. The following argument claims to establish that S is not free at t_2 not to do X. The operator [] expresses necessity (meaning that the proposition to which the operator applies is true in all possible worlds); t_1 is any given time prior to t_2.

A1: God believed at t_1 that S would do X at t_2;
A2: If God believed at t_1 that S would do X at t_2,

then [] (S does X at t_2);
Hence A3: [] (S does X at t_2).

A1 results from divine omniscience; A2 is supposed to express the essential character of this omniscience.

This is the argument that Saint Augustine has Evodius make in *De Libero Arbitrio*. It was left to Thomas Aquinas to identify the flaw in the reasoning. God's essential omniscience does not entail A2, but:
E2: [] (If God believed at t_1 that S would do X at t_2, then S does X at t_2).
A2 unduly affirms the "necessity of the consequent," whereas all that we can be sure of is E2, that is, the "necessity of the consequence." The problem is that it is no longer possible to deduce A3. To do so would require that A1 be itself considered necessary. Can an argument be made for this? Yes, by invoking the principle of the *fixity of the past*. It is not in the power of anyone (not even of God) so to act that what happened would not have happened. God believed something at t_1. That is a mental act, a fact that, for every t subsequent to t_1, belongs to the past, and is therefore intangible. At time t_2, A1 is necessary, not, as in the case of a logically necessary proposition, because it has always been necessary, but because it has become so. It was contingent that it become necessary. To use the scholastic terminology, it is "accidentally necessary."

Let $[]^S_t$ designate the operator of necessity such that:

$[]^S_t$ (p) means: p is true, and S is not free at t to perform an act such that, if he performed it, p would be false.
It is obvious that: [] (p) ---> $[]^S_t$ (p).

Consider the following line of reasoning:

E1: $[]^S_{t2}$ (God believed at t_1 that S would do X at t_2);
E2: $[]^S_{t2}$ (If God believed at t_1 that S would do X at t_2,
then S does X at t_2);
Hence E3: $[]^S_{t2}$ (S does X at t_2).

E1 correctly expresses the fixity of the past, and the argument is valid assuming one admits the validity of the following inference rule: if S is powerless over p, and is powerless over (p --> q), then he is powerless over q--which seems reasonable.

In its substance, this is the argument of Jonathan Edwards (*Freedom of the Will*, 1745). Its conclusion is that S will, to be sure, do X at t_2 if God foresaw that he would do so, but that he will not do it freely, for it is not in his power to do otherwise. The necessity that appears in E3 is of the same nature as that which appears in E1: it is the necessity of the past, transmitted via the intermediary of the logical necessity expressed by E2.

It is time to indicate why these arguments refer to divine *belief*, and not to divine knowledge--even though it is posited that if God believes that p, then p. The reason is that:
(1) S knows at t that p
entails:
(2) (S believes at t that p) and p.

So that if one were to refer to divine knowledge in E1, and not just to divine belief, one would be postulating from the outset what is to be proved: namely, the

necessity of p--with the paradox (which, to be sure, subsists at the end of the complete argument) that this necessity is the necessity of the past even though p pertains to the future (Hasker, 1989, p. 220).

An article by Nelson Pike, first published in 1965, was to provoke an intense controversy among analytical philosophers, the effects of which are still felt today. Taking up Edwards's argument in a new form, Pike likewise concluded that God's essential omniscience and human freedom are incompatible. I will refer in what follows to the condensed presentation of Pike's reasoning furnished by Alvin Plantinga (1989) and John Martin Fischer (1989).

The point of departure is the following assertion:

(3) God existed at t_1, He believed at t_1 that S would do X at t_2, and it is in the power of S at t_2 to refrain from doing X at t_2,

entails:

(4) Either (4.1) it is in the power of S at t_2 so to act that God would have held a false belief at t_1;

or else (4.2) it is in the power of S at t_2 so to act that God would not have existed at t_1;

or else (4.3) it is in the power of S at t_2 so to act that God would have held at t_1 a belief different from the one that He actually did hold.

One can easily convince oneself of the validity of this premise. The rest of the argument consists in showing that each of the three terms of (4) is necessarily false. It follows that (3) is false, which confirms the incompatibilist thesis.

(4.1) is necessarily false in virtue of God's essential omniscience. The falseness of (4.2) can be derived either by applying the principle of the fixity of the past to the postulated existence of God at t_1; or, if one disputes the applicability of this principle to the existence of God,[4] by invoking another, hard-to-dispute principle that holds this existence to be independent of human action. As for (4.3), its falseness follows from the principle of the fixity of the past applied to the belief that God had at t_1. Here the kinship is clear between Edwards's argument and Pike's: both rely ultimately on the principle of the fixity of the past.

Is it possible to be compatibilist? Can either of the foregoing arguments be refuted? Like the majority of contemporary authors, I will exclude two solutions which, however, are classic in the history of philosophy: the one that consists in denying that the future is real (Aristotelianism); the one that consists in denying that the eternality of God is situated in our temporal framework (Thomism). If the future is not real, God's omniscience does not extend to future contingents; if God does not exist in our temporal framework, His omniscience does not entail the faculty of foreknowledge, and His beliefs are not situated in the past. In either case, the incompatibilist argument collapses. There are good philosophical reasons for excluding these two solutions. I will not set them forth here because there is another of more immediate import for my argument: accepting that Newcomb's problem, with God as Predictor, is well formulated presupposes rejecting Aristotelianism and Thomism alike.

There remains a third solution, associated with the name of William of Ockham. It consists in denying that the principle of the fixity of the past has universal validity. The Ockhamite solution is based on a distinction that is no doubt fundamental, but of which the clarity leaves much to be desired. The current controversy centers on this question. Consider the facts relative to a time t. The object is to distinguish the

"hard facts"--those that are about t, strictly speaking--from those that are merely "soft facts" with respect to t. Thus:

(5) Napoleon entered Iena

is a hard fact about the past; while:

(6) Napoleon entered Iena before I give this paper to my publisher tomorrow

is a soft fact. It is easy to see the reason for this distinction. On October 13, 1806, (5) became accidentally necessary; it is no longer in anyone's power to keep (5) from being necessary. But it is enough that it be in my power to refrain from giving this paper to my publisher tomorrow (a hypothesis that seems hard to challenge) for the fixity of (6) as a fact about the past to be invalidated. It is possible for me to act tomorrow in such a way that (6) does not become (accidentally) necessary.

The Ockhamite ideal would be:

O1: to arrive at a criterion for demarcating hard and soft facts about the past that would be unambiguous in its application and such that:
O2: the principle of the fixity of the past would not apply to the latter, and:
O3: the principle of the fixity of the past would apply to the former.

It is evident what the compatibilist hopes to gain if these three objectives could be achieved. The idea would be to invalidate the incompatibilist arguments by showing that the proposition

(7) God believed at t_1 that S would do X at t_2

is no more than a soft fact about t_1 and therefore cannot be considered accidentally necessary at t_2--thus invalidating both E1 and the argument for the necessary falseness of (4.3).

Alas, the compatibilist must be content with much less. None of the three objectives O1, O2, O3 has been genuinely achieved to this day. Take O1. Intuitively, a criterion does seem to exist in the observation that it is difficult to consider a proposition such as (6) as being a hard fact about the past because its truth depends on the truth of a proposition about the future, namely:

(8) I will give this paper to my publisher tomorrow.

Applied to (7), the same criterion leads to this proposition's also being considered no more than a soft fact about the past, since it entails, in virtue of God's essential omniscience:

(9) S will do X at t_2

which is a proposition about the future.

The problem is that there doubtless is not a single fact about the past, no matter how hard, that does not entail a proposition about the future. Take the example of (5), which corresponds perfectly to the idea we have of what constitutes a hard fact about the past. But (5) entails:

(10) When I visit Iena, I will not be the first Frenchman to set foot there.

By this standard, every fact about the past ought to be considered soft. The Ockhamite compatibilist will maintain, however, that even in the absence of a definitive criterion, a proposition such as (7) should be treated as a soft fact about the past, even though it is exclusively concerned with an event that took place in the past, because it entails a proposition such as (9), that is concerned with the future in the strict sense--which cannot be said of (10) (Plantinga, 1989, p. 193).

A second difficulty is that, the objective O2 not having been achieved either, even if one is convinced that (7) is a soft fact, one still has not shown that (7) could be held not to be fixed at t_2. I will not enter into this debate. I will merely observe that for a compatibilist like Alvin Plantinga, the burden of proof falls in some sense

on the incompatibilists. Assuming that the latter could be convinced that (7) is indeed a soft fact, it would then be up to them to show why the principle of the fixity of the past should still apply.

Be that as it may, it appears to be established that an Ockhamite compatibilist refutes Edwards's argument by refuting E1. What can he say about Pike's argument? He needs to refute the argument for the necessary falseness of (4.3). He therefore needs to attribute to S the power at t_2 to act in such a way that the belief that God had at t_1 was not what it actually was. In such a form, this power seems inconceivable: it would be at worst the power so to act that God would both have had and not have had a certain belief; at best the power to produce the past, to bring it about.[5] Plantinga shows that it suffices to attribute to S a much weaker and more "innocent" power for the incompatibilist argument to collapse: *counterfactual power over the past*. In the case at hand, it can be expressed as follows:

(11) It is in the power of S at t_2 to do something such that, if he were to do it, God would not have had at t_1 the belief that He actually had.

In effect, if (3) is true, S necessarily has this power, in virtue of God's essential omniscience: it is in the power of S at t_2 to refrain from doing X at t_2, and therefore to do something such that, if he were to do it, God would have believed at t_1 that S would refrain from doing X at t_2. For Pike's argument to be valid, (4.3) must therefore be interpreted as (11). But if, precisely, S is endowed with this counterfactual power over the past, there is no longer an argument for demonstrating that (4.3) is necessarily false.

Let us come back to Newcomb's problem, in the case where the Predictor is God. The compatibilist, being led, as we saw, to attribute to human subjects a counterfactual power over the past, cannot but choose B_1, the one-box solution. Like Plantinga, he will affirm that the dominance reasoning leading to the choice of B_2 is invalid. This argument is of the form: if there is a million dollars in the opaque box, then if I took both boxes, there would (still) be a million in the opaque box. In other words: A is true; therefore, if p were true, A would be true. In counterfactual logic, this is inadmissible (1989, p. 203).

On the other hand, there exists, for the compatibilist, a valid line of reasoning leading up to the choice of B_1. The description of the problem gives us:

(12) If I took both boxes, then God would have believed that I was going to take both boxes

and

(13) If I took both boxes and if God had believed that I was going to take both boxes, then God would have left the opaque box empty.

From which it follows, with impeccable counterfactual logic, that:

(14) If I took both boxes, then God would have left the opaque box empty.

An analogous line of reasoning leads to the conclusion that if I took the opaque box only, then God would have put the million dollars in it. It follows that I should take only the one box.

This conclusion has a startling implication--namely, that even the most indisputable of hard facts about the past are not immune to our counterfactual power over the past--and therefore cannot be regarded as governed by the principle of the fixity of the past. Suppose that I decide to take only the opaque box. God foresaw this and, therefore, the following proposition is true:

(15) Before I make my choice, there was a million dollars in the opaque box.

At the time that I make my decision, it is nonetheless in my power to perform an act--taking both boxes--such that, if I were to do it, (15) would have been false. Yet

it is impossible to imagine a fact more strictly concerned with the past than (15) (Plantinga, 1989, p. 204).

This result, established in the case of Newcomb's problem, is easy to generalize. It suffices to posit that God is not only omniscient, but that He is endowed with providential power--which, from a theological standpoint, seems reasonable. God, foreseeing the future, acts in accordance with this foreknowledge. In these conditions, notes Plantinga, rare indeed are the facts about the past that can resist our counterfactual power. It is not at all inconceivable, for example, that it be in my power today to do something such that, if I were to do it, Napoleon would not have entered Iena on October 13, 1806. The objective O3 appears in these conditions to be out of reach.

At this stage, the decisive weapon in the compatibilist arsenal against the incompatibilist argument has become perfectly independent of the Ockhamite distinction between hard and soft facts about the past. This weapon is the postulate that human subjects are endowed with a counterfactual power over the past. One may say, more generally, that sometimes the compatibilist refutes the universal validity of the principle of the fixity of the past by invoking the distinction between hard and soft facts and sometimes by invoking the hypothesis of a counterfactual power over the past. At one extreme, we have facts such as (6), whose claim to be concerned with the past is illusory: there is no need to resort to counterfactual power over the past to disprove their fixity. At the other extreme, facts such as (5) or (15) are so obviously concerned with the past in the strict sense that it would be futile to try to present them as soft facts about the past; only the hypothesis of a counterfactual power over the past is capable of invalidating the principle of the fixity of the past. There are doubtless a whole range of intermediary cases, with a fact such as (7) occupying a position midway between the two extremes. It is possible to formulate a strategic principle to which the compatibilist seems to adhere:

CS (Compatibilist Strategy): the more it appears possible to find arguments in favor of the softness of a fact about the past, the less it is necessary to resort to counterfactual power over the past in order to deny the fixity of that fact; and vice-versa.

Essential omniscience and non-essential omniscience.

Now let us imagine a non-divine Predictor, *omniscient but not essentially omniscient*. Does this hypothesis constitute a threat to human freedom? An incompatibilist argument exists, analogous to that of Edwards:

P1: $[]^{S}_{t2}$ (The Predictor foresaw at t_1 that S would do X at t_2);

P2: $[]^{S}_{t2}$ (If the Predictor foresaw at t_1 that S would do X at t_2, then S does X at t_2);

Hence P3: $[]^{S}_{t2}$ (S does X at t_2).

Intuitively, it would appear easier to refute this argument than the argument in favor of theological determinism. The fact that you can foresee exactly what I am going to do does not seem seriously to threaten my free will. Yet P1 seems harder to refute than A1. The only argument that we had to deny that A1 is a hard fact about the past was that A1 entails that S will do X at t_2. We can no longer carry out this deduction in the present case because the Predictor is not essentially omniscient. To be sure, one could resort to attributing to S a counterfactual power over the past in

order to refute P1. That would be an awfully unwieldy and implausible weapon. In virtue of the CS principle, a much better option is available.

It is obviously P2 that is not justified, owing to the non-essential character of the Predictor's omniscience. This can easily be shown by applying an argument analogous to Pike's to the present case. Here, it is not by asserting that (4.3) is conceivable that this argument can be refuted, it is by doing the same for (4.1), reformulated as follows:

(16) It is in the power of S at t_2 to do something such that, if he were to do it, the Predictor would have made a false prediction at t_1.

This assertion appears much more "innocent" than the one that consists in saying that the Predictor would then have made a prediction different from the one he actually made. Can it not be said, however, that something of a counterfactual power over the past subsists in (16)? The hypotheses made include:

(17) The Predictor was omniscient at t_1

or

(17') The Predictor made a correct prediction at t_1.

But (16) entails:

(18) It is in the power of S at t_2 to do something such that, if he were to do it, the Predictor would not have been omniscient at t_1 even though he actually was.

(17), the fact about the past whose (accidentally) necessary character is thus denied, is in reality only the softest of soft facts about the past--just as soft, no doubt, as (6). That is why, in conformity with CS, the appearance of counterfactual power over the past assumed by (18) is illusory. The power asserted by (18) is, in truth, inherent in the distinction that we make between essential omniscience and non-essential omniscience.[6]

To sum up: the compatibility between the non-essential omniscience of the Predictor and human freedom in no way entails a counterfactual power over the past. In the case of Newcomb's problem, the absence of this power renders the dominance reasoning fully valid. One must therefore choose the two boxes.

We thus obtain what we have been seeking to establish. The hypothesis that the Predictor is omniscient is insufficient to dictate the agent's choice. He must opt for either the one box or the two boxes depending on whether this omniscience is essential or not.[7] It might seem that we thereby end up with a solution of the usual type, where the rational choice depends on the conditions of the problem--even if this time the bifurcation of these conditions is very unusual and too subtle for the needs of a practical philosophy. Not at all. Two conceptions of rationality are in conflict here, adapted to two conceptions of our relationship to time: one that entails a counterfactual power over the past, and one that does not.

Two forms of temporality.

Economists or decision theorists may wonder how this seemingly theological discussion concerns them. If they are put off by the recourse to God, never mind. As Laplace would have said, we have no need for that hypothesis. One can very easily do without God: it suffices, paradoxically, to posit that His existence is necessary. For in that case, "God believes that p is true" and "p is true" are logically equivalent.

To Edwards's argument in favor of theological determinism there corresponds an analogous argument in favor of logical determinism (or fatalism):

L1: $[]^{S}_{t2}$ (It was true at t_1 that S would do X at t_2);

L2: $[]^{S}_{t2}$ (If it was true at t_1 that S would do X at t_2, then

S does x at t_2);

Hence L3: $[]^{S}_{t2}$ (S does X at t_2).

Can the fatalist argument be refuted? Note that the proposition:

(19) It was true at t_1 that S would do X at t_2,

simply reflects the principle of the reality of the future, and L2, a requirement for consistency in the application of this principle. It is therefore L1 that must be refuted. The majority of compatibilists deem it significantly easier to do that than to refute E1. One may argue about whether a belief that God had twenty years ago is a hard fact about the past; it seems clear that (19) can only be treated as a very soft fact about the past. As CS indicates, L1 can be refuted without resorting to a counterfactual power over the past.

It is possible, however, to "harden" the fatalist argument. Consider the "Great Scroll" hypothesis.[8] It can be expressed in the following two propositions:

GS1: It was written at t_1 that S would do X at t_2;

GS2: [] (If it was written at t_1 that S would do X at t_2,
then S does X at t_2).

For the fatalist conclusion to be avoided, it must be shown that GS1 was not accidentally necessary at t_2. But GS1 appears to be a very hard fact about the past--even if one can argue, as in the case of an essentially omniscient God, that its truth depends on the truth of a fact about the future. In virtue of CS, the compatibilist argument requires recourse to a counterfactual power over the past in a very strong sense--as strong, no doubt, as the counterfactual power over the providential actions of God in Newcomb's problem.

The conception of time that I call "projected time" includes the following four interdependent characteristics:

PT1: reality of the future;
PT2: free will;
PT3: "Great Scroll" hypothesis ("hard" inscription in the past of PT1);
PT4: counterfactual power over the past.

These four traits make up a coherent, non-contradictory whole. The Great Scroll hypothesis (PT3) should be taken as at once metaphorical and variable in strength. It designates a "hard" or strict form of inscription in the past of the postulate of the reality of the future (PT1); hard enough to keep above a certain critical threshold the force of the counterfactual power over the past (PT4) necessary for free will (PT2) to be safeguarded from the danger that such strong inscription (PT3) represents for it.

We have now studied a number of different forms of inscription in the past of the reality of the future: "God foresaw that...," "God foresaw that... and acted accordingly," "the Predictor foresaw correctly that...," "it was true that...," "it was written that..." The critical threshold can be determined by submitting the corresponding conception of time to a test analogous to Newcomb's problem: we know we are in "projected time" if dominance reasoning is invalidated.

The "inscription" metaphor--the Great Scroll hypothesis--is somewhat deceptive. It is certainly inevitable: GS2 tells us that if it is written that Jacques will break his neck on such and such a day, then Jacques will break his neck. But PT4 tells

us that Jacques has the counterfactual power to change what is written on the great scroll. He has, at all times, the power to do something such that, if he were to do it, the "script" would have been different.

By contrast, what I call "occurring time" includes the following four characteristics:

OT1: reality of the future;
OT2: free will;
OT3: "soft" inscription in the past of OT1;
OT4: absence of counterfactual power over the past.

Here too, the traits OT3 and OT4 are at once interdependent and open to variation. The only requirement is that free will be safeguarded and that, in the Newcomb test, dominance reasoning be validated.

In occurring time, the Great Scroll metaphor does not come into play. It could certainly be that an infallible Predictor has written in advance the scenario of things to come. At every moment, however, the agents have the power to act in such a way that, if they were so to act, they would render inaccurate the predictions of the supposed Predictor.

What distinguishes occurring time from projected time is not a greater openness of the future ("nothing is written in advance," etc.). In both kinds of time, the future is at once real (PT1-OT1)--what it will be is "already there"--and open (PT2-OT2)--it could be different from what it will be. It is in the relationship to the past that the essential difference lies. The past is so to speak more fixed in occurring time (OT4) than it is in projected time--the reason being that in occurring time the inscription of the future in the past is softer (OT3).

Projected time is more "paradoxical" than occurring time, in that it entails a power over the past. The agent acts following a scenario prepared in advance, but since he is free, he can pull himself up to the level where the scenario is written and exercise a kind of power over it--the power that we call counterfactual. This "bootstrapping" or doubling expresses the demiurgic character of projected time.

Despite its sophistication, projected time corresponds to one of the forms assumed by the human experience of time--that of a subject executing a plan he has drawn up for himself, at once author and actor. It is an essential dimension of temporality, but not, as I shall undertake to show further on, the only dimension.

From lung cancer to the spirit of capitalism.

> He could no longer hide his excitement, the importance
> he attached to this encounter, and he promised in the
> event of success to bestow a reward on his coachman, as
> if, by inspiring in him a desire to succeed that would
> be added to the one within himself, he could so act that
> Odette, in case she had already gone home to bed, would
> nevertheless be found in a restaurant on the boulevard.
>
> Marcel Proust, *Un amour de Swann*

The invariant two-boxers--those who adopt the dominant strategy however much confidence they accord the Predictor--have armed themselves with two powerful weapons in the controversy that pits them against their adversaries--those who become one-boxers as soon as they have enough confidence in the Predictor's predictive ability: the first weapon is a particular formulation of the theory of

rational decisions, namely *causal* decision theory; the second is the discovery or invention of a class of problems, also called Newcomb problems, for which it seems intuitively obvious that the rational solution is to opt for the dominant strategy.

Causal decision theory formalizes a very simple intuition: when we choose among different possible courses of action, we should base our choice on the desirable character of our action's consequences *alone*--of that of which it is the *cause*, and not of that for which it merely provides *evidence*. When, in Newcomb's problem, the one-boxer justifies his choice as being the one which maximizes the mathematical expected utility, what it actually maximizes is not the probable utility of the state of the world that his action will bring about--it is the probable utility of the state of the world of which his action manifests the existence. Everything takes place as if the agent examined himself from an external vantage point, discovered along with us his own decision and sought to make as welcome as possible the news that this discovery brings him. If I were to act in this way, he reasons, my action would be *evidence* that such or such a satisfying feature of the world must be present. It is therefore in this way that I will act. Irrational, magical behavior, retorts the advocate of the causal theory.

There would thus be two competing theories of rational decision-making. The first, the classical theory, requires maximizing an expected utility calculated in the following way: for every action, one multiplies the utility of each of the possible states of the world, given that the action takes place, by the conditional probability of the state, given the same thing, and one takes the sum over the entire set of states. Using conventional notation, it is a matter of choosing the action A that renders as large as possible:

(20) $U(A) = \cdot_i \, p\,(S_i / A)\; U\,(A, S_i)$

where the S_i's represent the possible states of the world.

This traditional theory is known as the *evidential* decision theory. It concerns itself only with the overall information that the existence of A provides about the odds of S_i, without being interested in whether the dependence of S_i on A in each particular case is causal or not.

The *causal* theory proceeds in exactly the same manner as the evidential theory, except that it is interested only in those probabilistic relationships between A and S_i that correspond to a causal dependence of S_i on A. There exist diverse formulations of the causal theory. However, all of them obviously have in common that in the case of causal *independence* between A and S_i, the probability to be taken into account is $p(S_i)$. The weights used in calculating the expected utility are thus independent of the action, and if a dominant strategy exists, it will be adopted.

The causal theory therefore leads in every case to the two-box choice in Newcomb's problem--whereas the evidential theory leads to the opposite choice if one has sufficient faith in the predictive ability of the Predictor.

But why adopt the causal theory? Is the irrationality of the evidential theory so obvious? Yes, assert the invariant two-boxers, and they prove it using a class of examples which, although they have a structure analogous to that of Newcomb's problem, are more susceptible than it is to being decided by simple good sense.

It is well known that there is a high correlation between tobacco consumption and lung cancer. Suppose that it is discovered that the reason for this correlation is not, as was thought up until now, that tobacco consumption *causes* lung cancer. The cancer is caused by a certain genetic configuration which also happens to predispose people to smoke. Do smokers still have a reason to quit smoking? Their decision will in no way change the fact that they have, or do not have, the lethal gene. It would be absurd for them to deprive themselves of the pleasure they derive from indulging in

their vice. This situation has the same structure as Newcomb's problem. Between the decision and the state of the world, there is, at one and the same time, probabilistic dependence and causal independence. Only this time, a posited *common cause* for the decision and for the state of the world explains this combination more plausibly or realistically than the recourse to a hypothetical and providential Predictor. As to the structure of gains and losses, it likewise conforms to Newcomb's problem, the pleasure of smoking corresponding to the thousand dollars in the transparent box, and lung cancer, to the absence of the million in the opaque box.

Since there is causal independence, the causal theory concludes in favor of the dominant strategy: continue to smoke, as good sense dictates. Not so the evidential theory: giving up smoking is a low price for the smoker to pay for an indication that he does not have the lethal gene. But this behavior appears utterly irrational.

Here is a similar example, inspired by Gibbard and Harper (1985). Valmont, an up-and-coming executive in a multinational corporation, has taken a personality test in the hope of receiving an important promotion. He knows that success on the test hinged on displaying the killer instinct of a ruthless competitor, and he is not sure whether his performance measured up. Late Friday, Valmont learns that the promotion decision has already been made, but that it will only be communicated to him on Monday.

It is almost time to go home when Valmont gets wind of an indelicacy committed by one of his subordinates. He has a choice between showing no mercy or looking the other way. The latter solution presents certain advantages: it will spare him an unpleasant scene with his employee and allow him to leave the office earlier. But he reasons that if he shows himself to be ruthless, it will be good evidence that his personality is such that he must have passed the test. Yet is it not obvious that this behavior would be irrational, since Valmont knows that the promotion decision has already been made?

The advocate of the causal theory concludes his brief as follows: in these examples and others of the same kind, where our considered intuition permits a decision that is beyond debate, the evidential theory is found wanting, while the causal theory proves consonant with common sense. Recourse to the causal theory alone is therefore justified in a case like Newcomb's problem, where our intuition is no help to us.

How can the advocate of the evidential theory respond to this attack? In fact, he has a powerful argument at his disposal. If we have a clear intuition of the rational solution in the two examples examined above, that is because, unlike Newcomb's problem, they have a minimum of realism and correspond (especially the second) to experiences which are not foreign to our daily lives. Now, it is precisely insofar as they possess this modicum of realism that one may make the following assertion concerning them. Contrary to what the causal theory alleges in order to undermine its rival, it cannot be the agent's knowledge of the decision he makes that provides him with an indication of the existing state of the world. Even before he acts, he must be able to draw this inference from his predisposition to act in such or such a way. The inveterate smoker can tell himself that he already possesses enough evidence of the high probability that he has the cancer gene and that his decision to keep on smoking or to quit will not affect his estimate of this probability. He therefore has no reason to quit. If Valmont feels inclined to show his subordinate no mercy, that is what will reassure him about his chances of having passed the test. He therefore no longer has any reason to come down hard on his employee.

In other words, there exists in these examples a variable which the agent can observe within himself before making a decision: the disposition to act in a certain way, which "screens off" the effective decision from the state of the world. It is

exclusively through this disposition that the probabilistic dependence on the state of the world is realized:
- either, as in the example of tobacco and lung cancer, this dependence is observable statistically in the actions themselves, but it is supposed in general that these actions reflect the dispositions, the agents acting before any deliberation that might lead them to act otherwise;
- or else, as in the case of Valmont, the dependence is assessed subjectively, through introspection, but with the agent again relying on his initial impulse to determine his disposition to act, and not on what more or less in-depth reflection would lead him to decide.
In every case, the disposition to act is the only source of evidence concerning the hidden variable constituted by the state of the world. Once this evidence has been analyzed and an inference drawn about the state of the world, there is nothing in the decision itself that would allow one to modify or clarify this inference. The decision is devoid of any evidential value. In consequence, the evidential theory leads to the same conclusion as the causal theory. The counter-examples meant to prove the superiority of the latter over the former are thus defused. If the rational choice seems obvious here, it is not because the obviousness of such superiority is manifesting itself, it is because the structure of the situation is such that decision theories as different in spirit as these reach the same conclusion.

The ball is now in the court of the causal theorists. They have a choice between two strategies. The first consists in denying that it is always possible to find a screen-variable in a disposition to act that could be identified before any decision. It is possible to specify the "common cause" examples in such a way as to eliminate the screen variable. One may suppose for example that Valmont can only assure himself of the ruthlessness of his personality by actually firing his employee; or that, when they learn of the existence of the cancer gene, it is not until they know that they have not decided to quit that the heaviest smokers will be able to realize that they have the gene in question. Only the actual decision in the framework of a given Newcomb problem has evidential value. Now, assert the advocates of the causal theory, when these examples have been specified in this way the intuitive result continues to favor the dominant strategy--since the common cause, being located in the past, is independent of the actual decision. But here, the evidential theory, lacking a screen-variable, comes to a different conclusion and contradicts intuition.

The second strategy consists in turning the force of the screen-variable argument against those who employ it. Not only is the argument admissible, concede the advocates of the causal theory, but it is *always* true. There is always a screen-variable, even in the original Newcomb problem, for never can a free and rational action be evidence for an agent of a state of affairs located in the past and unknown to him. In these conditions, the evidential theory will never conclude differently from the causal theory. The latter is therefore left to occupy the stage alone (Eels, 1982).

An advocate of the evidential theory in spite of it all, Paul Horwich (1987) is the one who has gone the furthest to defend it from the sustained attacks of the rival theory. Against the second strategy, he denies that there is always a screen-variable. The argument for this claim is flawed, he shows, because it neglects one possibility: that the disposition to act is only known to the agent after he has determined what it is rational to do--in which case the knowledge of his disposition to act is no longer of any use to him. There are therefore situations, and the original Newcomb problem is one of them, in which it is impossible to identify a screen variable. Now, in these situations, Horwich asserts against the first strategy of the causal theorists, it is untrue that our intuition leans toward the dominant strategy. These cases without a screen-variable are too strange, too far removed from our experience, for our

intuition to have anything to say about them. The causal theory thus has no convincing counter-example with which to oppose the evidential theory. Either one is dealing with a "realistic" case, and both theories come to the same conclusion because there is a screen-variable; or the case is "strange," there is no screen-variable, and intuition no more favors the causal theory than it does the evidential theory. It is only possible to come to a decision by appealing to general principles to which it appears desirable to submit the chosen decision theory. Horwich enumerates four such principles (simplicity; stability; absence of arbritrariness in the relationship between past and future; normative consistency) and shows that on these four tests the evidential theory comes out ahead of its rival.

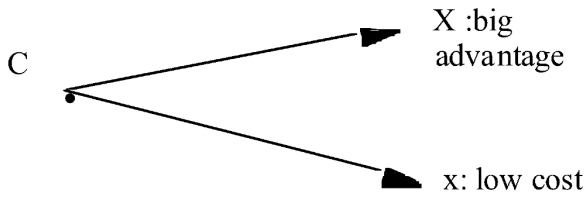
I have reviewed at length the present state of the discussion in order to make clear how completely the terms of the debate are changed by the solution proposed here.

It seems to me, first of all, that if one sticks to the postulate that all the authors whom we have examined share, namely that rationality is univocal, then there is no object that corresponds to the notion of a "common cause Newcomb problem." Let me stipulate that, following all our authors, I take the term Newcomb problem to mean a decision situation in which one has, at one and the same time:

a) causal independence and (strong) probabilistic dependence between the decision and the state of the world;

b) a structure of gains and losses analogous to that of the original Newcomb problem.

In a common cause Newcomb problem one should therefore find a state of the world C that is the cause of both a very favorable state of affairs X and a moderately costly decision x:



We will grant that Not-C causes Not-X and Not-x. The following chart compares the two examples considered above to the original Newcomb problem (in which C *has no* causal role in regard to x):

	Tobacco	Valmont	Original Newcomb
C	Absence of the gene	Ruthless nature	Predictor's prediction
X	Absence of cancer test	Success on the	$1 M in the opaque box
x	Decision to quit smoking	Ruthless conduct	Renunciation of the $1,000

The first question is whether it is really x, the actual decision that C determines, or whether it is not rather a disposition to act, D(x)--whose observation by the agent

can lead him to make the opposite decision. It seems to me that there are three cases to be considered a priori, depending on the level of deliberation at which one stops to establish the disposition to act:

C1: What C (or Not-C) causes is a disposition to act identifiable at the level of an irrational impulse, before any deliberation. D(x) is then indeed an indication of C and serves as a screen-variable. It is therefore rational for everyone, those who have C and those who do not, to do Not-x (since the utility of Not-x is superior to that of x, and since it is the only element of decision remaining.[9] The problem is that the weapon constituted by the screen-variable is too powerful: it works too well because it dissolves Newcomb's problem. When the agent makes his decision, he already knows the state of the world. The only thing that matters about his decision is its immediate utility.

C2: What C (or Not-C) causes is a disposition to act identifiable at the level of the rational decision that would be made "in normal circumstances"--understood to mean a simple decision context, one without Newcomb features. Then only the immediate utility of the decision would count, and everybody would do Not-x. Such a disposition to act therefore cannot be an indication of the presence or absence of C. This case must be rejected.

C3: What C (or Not-C) determines is the very decision that the agent makes rationally in the context of the Newcomb problem. Which means that there is no screen-variable. Now, remember that we are supposing rationality to be univocal. Whether it is the rationality of the causal theory or of the evidential theory, it will lead everyone, C's and Not-C's alike, to act in the same way. Therefore the actual decision cannot be an indication of the presence or absence of C--which contradicts the problem as given. In other words, there is no supposedly univocal theory of a common cause Newcomb problem that is not self-refuting: any such theory, accepted by the agents, leads them to act in a way that invalidates the opening hypotheses.

With no possibility other than dissolution into triviality or self-refutation, the common cause Newcomb problem seems threatened with extinction.

It is nevertheless important to save it. Indeed, I believe that it corresponds to a fundamental human experience. Now, it is possible to put it on a firm footing if one abandons the postulate that rationality is univocal. Let us grant the existence of two conceptions of rationality, associated respectively with what I call projected time and occurring time. Let us further suppose that those who possess C have a strong propensity to place themselves in projected time, and those who do not possess C, a strong propensity to place themselves in occurring time. Finally, let us grant, before demonstrating it, that in a manner analogous to the original Newcomb problem, projected time leads one to choose x, and occurring time, Not-x: the high probabilistic dependence between C (or Not-C) and x (or Not-x) posited at the outset is thereby guaranteed. At last we have a common cause Newcomb problem and a theory of the problem such that the solution confirms the basic premises.

It will be well to keep in mind the significance of the hypothesis underlying this solution: it is those who are already assured of "success" (C, and therefore X) who are the most disposed to a mode of rationality which makes it seem worthwhile to acquire, at moderate cost, the signs of success x, even though the actual determinants of success are beyond our control.

How can the existence of two modes of rationality be justified in the case of a common cause Newcomb problem? Such a duality would contradict something accepted by all our authors, whichever side they are on, as intuitively obvious: namely, that in the "realistic" cases it is Not-x, the dominant strategy, which must be chosen. This intuition is, as we saw, taken by those on the causal theory's side to be the mark of its superiority and by those on the other side to result from the implicit

working of a screen-variable. I believe for my part that this intuition is quite simply deceptive, but that if one had to seek its source, one would have to look in still another direction--that of the principle of the fixity of the past. If the choice of the dominant strategy seems more rational in the case of the common cause than in the original problem, that is because C appears more solidly inscribed in the past, and therefore more "fixed," than in the first case. There are two reasons for this:

a) The fact in question is more concrete than a mental event taking place in the mind of a hypothetical Predictor;

b) C is the cause of x, which is obviously not the case in the original Newcomb problem (where it is rather x which appears--wrongly, needless to say--as the cause of C).

Unfortunately, these two reasons are worthless. And one will recall the apparently paradoxical result that we reached earlier: the more the inscription in the past of the reality of the future is "hard," "solid," the less the past must be considered as fixed, lest free will be invalidated. Now, it is with this point, of course, that one ought to have begun. Isn't the hypothesis of a common *cause* incompatible with the postulate that the agents are free? We have already defended free will against the twin threats of theological determinism and logical determinism (fatalism). The time has come to defend it against the threat of causal determinism. And this threat is in principle more serious than its predecessors. For there seems to be no more solid inscription in the past of the reality of the future than the existence in the past of the cause of a future event.

Here is how the incompatibilist argument of the Edwardian type would run in this case:

DC1: $[]^{S}_{t2}$ (C occurred at t_1);

DC2: $[]^{S}_{t2}$ (If C occurred at t_1, then S does x at t_2)

Hence DC3: $[]^{S}_{t2}$ (S does x at t_2).

DC1 derives from the fixity of the past, and DC2 from the fixity of the laws of nature.

Can this argument be refuted in the same way that we refuted analogous arguments in the foregoing analyses? It would only seem possible here at a very high price: one would have to deny either the fixity of the past or the fixity of the laws of nature (Fischer, 1989).

The theory of deterministic chaos will allow us to reduce this cost significantly. Let us grant that the deterministic system leading from C to x is a "weakly stable" system, "sensitive to initial conditions." The class of initial states C that lead to x, and the class of initial states Not-C that lead to Not-x, are intimately intermingled. As a result, the process that leads from C (or Not-C) to x (or Not-x) is "complex," in the sense that it is not possible for any human subject to determine, *before* the event takes place, if it is x or Not-x that will emerge. There is no model of this deterministic process that is simpler and faster than the process itself, as it unfolds in time. In these conditions, from the point of view of a subject S, the causal necessity has no importance until such time as it has already become an accidental necessity. It is perfectly legitimate that he reason as a free subject, capable of acting otherwise than he actually does. To put it another way: it is only insofar as there is no screen-variable that the subject can judge himself to be free.

If one grants the foregoing--and not to grant it is to leave the common cause Newcomb problem devoid of meaning--then two possibilities are open to the agent, neither of which seems a priori more legitimate than the other:

a) either he admits DC2, and he cannot but deny DC1. The laws of nature are fixed, the causal inscription in the past of the reality of the future is "hard," and so the subject, seeing himself as free, must endow himself with a counterfactual power over the past fact C. We are in projected time;

b) or else he admits DC1, and he cannot but deny DC2. The past is fixed, so the subject, seeing himself as free, must endow himself with the counterfactual power to invalidate the determinism leading from C (or Not-C) to x (or Not-x). The causal inscription in the past of the reality of the future is soft. We are in occurring time.

In the second case, the rational choice is the dominant strategy Not-x. It is x in the first case.

The solution that I propose here has many implications that contradict what is generally accepted in the present state of the discussion. For example, there is no such thing as a Newcomb problem with a screen-variable, whether or not one is dealing with a "realistic" case. The case may well be realistic, adopting the dominant strategy is not obvious. The case may well have no screen-variable, it is not necessarily "strange" to the point that intuition will have nothing to say. That intuition has something to say does not mean that it comes down clearly in favor of one choice to the detriment of the other. I think one can show that our considered intuition of the solution that this class of problems call for is as ambivalent as the formal solution that I am proposing.

This assertion can be supported by invoking an example of overwhelming theoretical and historical importance to which, surprisingly and rather dismayingly, the current debate almost never refers. I am thinking of the celebrated thesis in the guise of a paradox that Max Weber proposed concerning the "correlations" between the "Protestant ethic," more precisely the ethical consequences of the doctrine of predestination, and the "spirit of capitalism" (Weber, 1930).

What interests me here is only the logical structure of Weber's argument, and not its empirical validity (the extremely controversial nature of the thesis is well known, but one may note that many critics were put off precisely by its paradoxical aspect and did not attempt to look any further). I will render in schematic fashion an argument that is already "ideal-typical." In virtue of a divine decision made for all eternity, everyone belongs either to the camp of the elect or to that of the damned, without knowing which. There is absolutely nothing people can do about this decree, nothing they can do to earn or to merit their salvation. However, divine grace manifests itself by signs. These signs cannot be observed through introspection, only acquired through action. The principal sign is the success obtained by *proving* one's faith in a worldly profession (*Beruf*). This proof is costly, it requires that one work methodically, without let-up, without ever relaxing in the security of possession, without ever pausing to enjoy one's wealth (Weber, p. 157). "You may labour to be rich for God, though not for the flesh and sin," that is how the Presbyterian pastor Richard Baxter exhorted his English flock in the second half of the 17th century (Weber, p. 162). "Unwillingness to work," Weber notes, "is symptomatic of the lack of grace" (p. 159).

We have all the ingredients here for a common cause Newcomb problem. The common cause C is the divine decree. X is eternal salvation; x, the *costly* decision (only moderately so, to be sure, in comparison to the magnitude of the stakes) to acquire the signs of grace--that is, *the decision to consider oneself chosen*. We even have the hypothesis that must be satisfied, as we have shown, in order that a common cause Newcomb problem not collapse into self-refutation: the *nature* of the elect leads them to place themselves in projected time--and that is why they choose x. "The *electi* are," Weber notes, "on account of their election, proof against fatalism because in their

rejection of it they prove themselves 'quos ipsa electio sollicitos reddit et diligentes officiorum'" (p. 232, n. 66).

(Note that what Weber refers to here as "fatalism" is not what we labelled as such in our earlier discussion--logical determinism--but simply the type of rationality, proper to occurring time, which leads to the choice of the dominant strategy Not-x).

The "logical consequence" of this *practical* problem, Weber remarks, should "of course" have been "fatalism" (p. 232, n. 66). Like nearly all the current authors, Weber thus views the dominant strategy as the obvious choice in a common cause Newcomb problem. His whole book is nevertheless devoted to explaining how and why "the broad mass of ordinary men" (p. 110), with few exceptions, made the opposite choice: "But on account of the idea of proof the psychological result was precisely the opposite" (p. 232, n. 66).

The Calvinist doctrine of the masses held it "to be an absolute duty to consider oneself chosen, and to combat all doubts as temptations of the devil, since lack of self-confidence is the result of insufficient faith, hence of imperfect grace" (p. 111). The means of acquiring this self-confidence, the means of assuring oneself of one's state of grace, was "intense worldly activity" (p. 112), thus giving rise to another paradox: here we have an essentially ascetic and "anti-mammonistic" doctrine which condemns the pursuit of money and material wealth, and yet which "ends up in practice by making a moral obligation out of the riches that crown the accomplishment of professional duties" (Besnard, 1970, p. 89).

Even though Weber puts great stress on it, most commentators have missed this key point: the processes that he describes were not contained in the doctrines of the theologians, nor in the rules of morality enunciated by the preachers. Instead, they were "unforeseen and even unwished-for results of the labours of the reformers. They were often far removed from or even in contradiction to all that they themselves thought to attain" (Weber, p. 90). The quest for signs of grace in professional success was perfectly contrary to Calvin's doctrine, as was the subsequent disappearance of the doctrine of justification by pure grace. For Calvin, the believer could know with certainty that he had been touched by grace, from the fact that he heard within himself and believed in the word of Christ--in other words, a "screen-variable" was at work, making it possible to trivialize the problem.

The choice of placing oneself in what I call projected time was in fact the spontaneous response of the mass of the faithful to the anguish induced by the questions: "Am I one of the elect? ...And how can I be sure of this state of grace?" (Weber, p. 110). It was a "psychological" reaction, writes Weber, motivated by the intense *desire* to figure among the elect, by the desire for salvation. We probably have here the motive force underlying projected time, which we ought perhaps to call desired time.

The Lutherans accused the Calvinists of reverting to the dogma of "salvation by works"--to the great dismay of the latter, outraged that their doctrine could be identified with what they most scorned: Catholic doctrine. This accusation amounts to saying that someone who, placing himself in projected time, chooses x, reasons *as if* x were the cause of X--behavior that is *magical* in the true sense of the word, insist the accusers, since it amounts to taking the sign for the thing (x for C). And this accusation is none other than the one which, in our day, the advocates of causal decision theory level at their adversaries, the defenders of the evidential theory.

The Lutherans accused the Calvinists of reverting to the dogma of "salvation by works"--to the great dismay of the latter, outraged that their doctrine could be identified with what they most scorned: Catholic doctrine. This accusation amounts to saying that someone who, placing himself in projected time, chooses x, reasons *as*

if x were the cause of X--behavior that is *magical* in the true sense of the word, insist the accusers, since it amounts to taking the sign for the thing (x for C). And this accusation is none other than the one which, in our day, the advocates of causal decision theory level at their adversaries, the defenders of the evidential theory.

The expression "as if" is ambiguous. If one interprets it to mean that both lines of reasoning lead to the same result, then the accusation is well-founded, since in *practice* the two doctrines, Calvinist and Catholic, are indistinguishable (Weber, p. 115-116). But if the interpretation is that the Puritans *really* took the sign for the thing, then the accusation becomes incomprehensible, and perfectly unjustified. For, Weber shows, as is well known, ascetic Puritanism constitutes the final stage in the vast movement of "elimination of magic from the world" that repudiates "all magical means to salvation as superstition and sin" (p. 105). Recalling the "Puritan's ferocious hatred of everything which smacked of superstition, of all survivals of magical or sacramental salvation" (p. 168), he describes how this state of mind engendered in each individual "a feeling of unprecedented inner loneliness" (p. 104) and was at the root of a "disillusioned and pessimistically inclined individualism" (p. 105). More interesting yet from our standpoint, it is this Puritan outlook on life which, Weber shows, "stood at the cradle of the modern economic man" (p. 174), "gave birth to economic rationalism" (p. 259, n. 4) and transformed the "calculating spirit" of capitalism "from a mere means to economy into a principle of general conduct" (p. 261, n. 10).

Thus, not only does the choice of placing oneself in projected time reveal itself, in the paradigmatic case, and despite its paradoxical character, to be perfectly "rational" (just as the choice of placing oneself in occurring time would have been), but, in addition, this rationality turns out to be none other than economic rationality. This excursion into the history of religions brings us back to a discovery that we were already able to make through logical analysis.

One can see how far it is from the truth to assert, as Robert Nozick does in the first article ever published on the subject (1969), that "Newcomb's problem" was "constructed" by a physicist, William Newcomb, in the early sixties. It would be just as far from the truth to say that it was invented by theologians. We are dealing with a fundamental existential problem that rears its head every time we are confronted with absolute uncertainty concerning a variable on which our "salvation" depends. The question then is whether we are ready to pay the necessary price to acquire the *signs* of "election." The whole problematic of "conspicuous consumption" (Veblen), or of the "demonstration" or "sign effects" of consumption (d'Iribarne), could be taken up again in this perspective.

Counterfactual decision theory.

> The calculation which we make in our heads and the one
> recorded on the register up above are two very
> different calculations. Is it we who control Destiny
> or Destiny which controls us? How many wisely conceived
> projects have failed and will fail in the future! How
> many insane projects have succeeded and will succeed!
>
> *Jacques the Fatalist* (p. 29)

The great interest of Newcomb's problem is that it brings to light the fact that our conception of rationality is not univocal, and that it is at the very least ambivalent. Does that mean that the two rival decision theories, the causal theory and the evidential theory, are both acceptable? It is obviously the opposite conclusion

that must be adopted. Neither one is acceptable, for each excludes one of the two modes of rationality.

The ambivalence of rationality does not imply that both decision theories are necessary. There is no reason to exclude the hypothesis that a single theory, by being adaptable to either temporality, could account for both rationalities. Such a theory exists and is even mentioned in the literature. However, the conviction that rationality is univocal is so firmly fixed in the minds of our authors that no sooner is this theory envisaged than it is reduced to one or the other of the two rival theories, thereby losing precisely what gives it its value: its indeterminate character.

The theory in question is the counterfactual decision theory, which can be formulated as follows: choose the action A that renders as large as possible:

(21) $U(A) = \cdot_i \; p(A \; []\text{--->}S_i) \; U(A, S_i)$

where, as before, the S_i's represent the possible states of the world. The counterfactual PÊ[]--->Q (which can be read: if P were true, then Q would be) is true in the possible world w (apart from the "degenerate" case in which P is not true in any possible world) if and only if Q is true in the possible world which, among those where P is true, is the closest to w.[10]

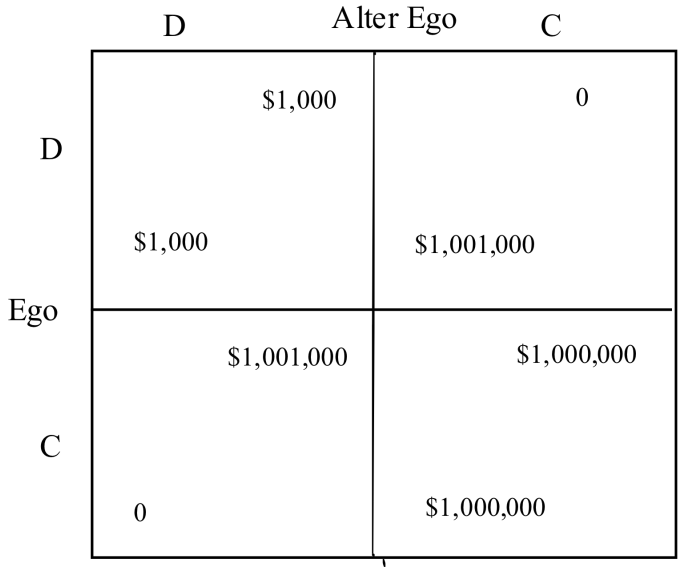
Gibbard and Harper (1985) adopt this theory, but they supplement it with the following principle: if Q is causally independent of P, then p(PÊ[]--->Q) = p(Q). In Newcomb's problem, that amounts to ruling out the possibility of a counterfactual power over the past and therefore to reducing the counterfactual theory to the causal theory. Horgan (1985a) takes up the same theory, sees, following Lewis (1979), that its reference to a relationship of similarity or proximity between possible worlds makes it highly indeterminate, and proposes to eliminate this indeterminacy in such a way that one always has: p(PÊ[]--->Q) = p(Q/P). That amounts to reducing the counterfactual theory to the evidential theory and, in Newcomb's problem, to postulating the necessary existence of a counterfactual power over the past.

The solution that I propose requires that the counterfactual theory be left with its indeterminacy intact: that is precisely what makes for its interest. It is the type of temporality in which one places oneself that makes it possible to evaluate the counterfactual probabilities. In projected time, one postulates a counterfactual power over the past, and the counterfactual theory leads to the same conclusion as the evidential theory, *yet cannot be assimilated to it*. In occurring time, one finds the same kind of identity between the conclusions of the counterfactual theory and the causal theory, without there being identity at the level of the theories.

Prisoner's dilemma and Newcomb's problem.

Lewis (1985) has attempted to show that the prisoner's dilemma is, fundamentally, a Newcomb problem--or, to be exact, two Newcomb problems side by side, one for each of the players. If this thesis is correct, it follows that Newcomb's problem is a structure no more exceptional than the ever so widespread prisoner's dilemma. What are we to make of this thesis in the light of our solution?

Here is a prisoner's dilemma:



As seen by Ego, the game looks like this:

(22) By choosing D (for defection but also dominance), I am assured of having at least $1,000.
(23) I may perhaps obtain a million dollars, but that does not depend (causally) on me.
(24) I will have the million if and only if Alter Ego renounces taking his thousand dollars.

Newcomb's problem is nothing other than (22) and (23) with the following addition:
(24') I will have the million if and only if the Predictor foresaw that I would renounce taking my thousand dollars.

But what better simulation of the predictive process of the Predictor could I imagine than the deliberation of someone who, like my Alter Ego, is placed in a situation exactly similar to mine? Hence the announced equivalence.

Lewis, an advocate of the causal theory, is an invariant exponent of the dominant strategy. In every case, he chooses the two boxes in Newcomb's problem in the same way that he chooses defection in the prisoner's dilemma. Some people are one-boxers when they have enough confidence in the predictive powers of the Predictor, he reminds us. This behavior is exactly like cooperating in the prisoner's dilemma if I am sufficiently convinced that my partner will act in the same way that I do. Now, this attitude, even if it is recommended by some, is clearly irrational in the opinion of Lewis.

One can see right away where his reasoning goes astray. The equivalence that he has demonstrated is purely formal and takes no account of what makes for the obvious difference between the Newcomb Predictor and my Alter Ego in the prisoner's dilemma. The latter is an agent acting freely in his own best interest, the former is a principle, an abstract entity whose sole function is to serve as the predictor of my behavior. If the one-box choice in Newcomb's problem can be considered rational, that is because it is possible to imagine that the agent possesses a counterfactual power over the predictions of the Predictor. In equivalent fashion, in order to make cooperation in the prisoner's dilemma a rational choice, one would have to imagine that I wield a counterfactual power over the actions of a free agent, my Alter Ego, who acts independently of me. That is much more problematic. The

counterfactual power rests, as we have shown, on the hypothesis that the Predictor is infallible in essential fashion, that is to say in all possible worlds; it derives from the postulate that, despite this threat to his free will, the agent remains free. But, if one may envisage the possibility that Alter Ego's choice will always coincide with mine in practice, there is in general nothing to justify holding this linkage to be true in essential fashion. The difference lies in the fact which Lewis judges to be, precisely, "inessential," namely that the prediction by Newcomb's predictor was made *in the past*, as a pre-diction in the etymological sense of anticipation.

To sum up: as a rule, the only rational choice in the prisoner's dilemma is that of the dominant strategy. One therefore cannot speak of equivalence with Newcomb's problem, contrary to what Lewis asserts.

It is possible, however, to re-establish an equivalence by *specifying* the prisoner's dilemma in such a way as to make it into a common cause Newcomb problem. This case is considered in the literature, with the usual bifurcation, some choosing cooperation, others defection. One posits that Ego and Alter Ego share a common "nature," without knowing whether that nature causes them to cooperate or to defect. Referring back to our earlier discussion, we need to stipulate: 1) that the posited determinism is "complex" and that there is therefore no screen-variable; 2) that the agents know that they will make the same choice because the determinism in question leads them to adopt the same mode of rationality. It then becomes rational for Ego to choose cooperation (which does not keep defection from remaining rational): this behavior corresponds to projected time, Ego deeming to be (essentially) fixed not only the linkage between the state of the common cause and his choice, but also the linkage between this state and Alter Ego's choice (this last hypothesis is indispensable given the ambivalence of rationality). In considering himself free, Ego thereby endows himself with a counterfactual power over the state of the common cause, and therefore over the choice that Alter Ego makes independently of his own. Note that the real (essentially infallible) predictor of Ego's choice is not in this case Alter Ego's choice, but the *past* state of the common cause.

It is perhaps along these lines that one may be able to explain what seems to be a well-attested fact: in societies with a Puritan tradition, where everyone places himself in projected time and can take for granted that the same is true of everyone else, people generally manage to surmount everyday situations of the prisoner's dilemma type, even "one-shot" cases (traffic congestion, queueing behavior, etc.).[11]

Notes

1 Formulated by a physicist, William Newcomb, early in the 1960s, this problem was first treated in print in 1969 by the philosopher Robert Nozick. Since then, publications on it have steadily multiplied, in philosophy journals and popular science magazines alike. A good survey of the debate can be found in the collection of essays compiled by R. Campbell and L. Sowden (1985).

2 Of all the positions on Newcomb's problem known to me, only the one that Terence Horgan defends in the two essays included in Campbell and Sowden's anthology (especially in the second essay) somewhat approaches the solution I propose here. There are two important differences, however: a) it is only reluctantly that Horgan accepts the existence of two forms of rationality that are irreducible to one another, and he judges that the controversy is destined to remain stalemated. I believe on the contrary that both forms are an integral part of the human experience of time; b) Horgan remains in spite of everything an advocate of B_1, the one-box solution. I judge for my part that while both modes of rationality are equally legitimate, one is more "fundamental" than the other (in a sense that will be spelled out below), and that is the one which leads to the solution B_2.

3 A very good survey of the discussion can be found in the collection of essays compiled by J. M. Fischer (1989).

4 It can be disputed that the existence of God at t constitutes a fact about t *in the strict sense* (cf. below).

[5] Fischer (1989) introduces a distinction between causal and non-causal production of the past, which is not the same as the distinction between the power to bring about the past and counterfactual power over the past. Hasker (1989) denies that this last distinction is valid. I will not enter into this controversy. It is enough for me to note that counterfactual power over the past is not a causal power.

[6] One might posit that, although non-essential, the Predictor's omniscience is such that it remains counterfactually independent of the actions of any human subject. That is the hypothesis that Plantinga proposes with regard to a God who would be omniscient, but not essentially so (1989, p. 196). It is then, rather paradoxically, that, in an Ockhamite perspective, God's omniscience would threaten human freedom. Since we are interested in the case of a Predictor who is not divine, we do not need to consider this possibility.

[7] The authors who, following Nozick, choose the one box in the case where the Predictor is infallible and both boxes in the other cases, are at a loss to justify the resulting discontinuity between infallibility and quasi-infallibility. The solution which I propose at least has the advantage of dissipating this mystery. The limit of a more and more perfect predictive capacity is de facto infallibility, and the two-box choice remains the rational choice right up to and including the limit. The break occurs between de facto infallibility and essential infallibility.

[8] J. M. Fischer (1989, pp. 43-44) refers to this hypothesis in a slightly different get-up, attributing it to an unpublished manuscript by D. W. Widerker, assistant professor at Bar-Ilan University. Poor Diderot...

[9] It is not possible to posit, as some authors do in the tobacco example, that the utility of Not-x (the pleasure of smoking) depends on the presence or absence of Not-C; such that the persistent smokers would be those whom the presence of the gene leads to smoke *because* it gives them pleasure. This hypothesis contradicts the structure of gains and losses that is a defining feature of Newcomb problems.

[10] This formulation is Stalnaker's (1968). A more general formulation will be found in Lewis (1973).

[11] To be justified in treating the prisoner's dilemma as a common cause Newcomb problem, it is not enough to posit in Kantian fashion, as many authors do, that Ego and Alter Ego share the common characteristic of being "rational." They still need to be sure that they share the same mode of rationality. In societies with a Puritan tradition, one tends not to choose the dominant strategy in common cause Newcomb problems, and to cooperate in the prisoner's dilemma. Does this consistency extend even to the original Newcomb problem? In an American sampling (*Scientific American* readers who made their response known), cited by Poundstone (1988), one-boxers outnumber two-boxers by two-and-a-half to one. It would be interesting to compare this result to what one would find in a society with a Catholic tradition.

Bibliography

Robert J. Aumann (1988), "Preliminary Notes on Irrationality in Game Theory," paper presented in "IMSSS Summer Seminar on Economic Theory," *mimeo*, Stanford University (July 14).

Philippe Besnard (1970), *Protestantisme et capitalisme: La controverse post-wŽbŽrienne*, Paris, Armand Colin.

Richmond Campbell and Lanning Sowden, eds. (1985), *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem*, Vancouver, The University of British Columbia Press.

Denis Diderot (1986), *Jacques the Fatalist*, tr. Michael Henry, Harmondsworth, Penguin.

Jean-Pierre Dupuy (1992), *Introduction aux sciences sociales*, Paris, Ellipses.

Ellery Eells (1982), *Rational Decision and Causality*, Cambridge and New York, Cambridge University Press.

John Martin Fischer, ed. (1989), *God, Foreknowledge, and Freedom*, Stanford, Stanford University Press; with an Introduction: "God and Freedom."

Allan Gibbard and William L. Harper (1985), "Counterfactuals and Two Kinds of Expected Utility," in R. Campbell and L. Sowden (1985), pp. 133-158; originally published in Hooker, Leach and McClennen, eds. (1978), *Foundations and Applications of Decision Theory*, vol. I, Dordrecht, Holland, D. Reidel, pp. 125-162.

William Hasker (1989), "Foreknowledge and Necessity," in J. M. Fischer (1989), pp. 216-257; originally published in *Faith and Philosophy*, 2 (1985), pp. 121-157.

Terence Horgan (1985a), "Counterfactuals and Newcomb's Problem," in R. Campbell and L. Sowden (1985), pp. 159-182; originally published in *The Journal of Philosophy*, 78, no. 6 (June 1981), pp. 331-356.

Terence Horgan (1985b), "Newcomb's Problem: A Stalemate," in R. Campbell and L. Sowden (1985), pp. 223-234.
Paul Horwich (1987), *Asymmetries in Time: Problems in the Philosophy of Science*, Cambridge, Mass., The MIT Press.
David K. Lewis (1973), *Counterfactuals*, Cambridge, Mass., Harvard University Press.
David K. Lewis (1979), "Counterfactual Dependence and Time's Arrow," *Nous*, 13, pp. 455-476.
David K. Lewis (1985), "Prisoner's Dilemma Is a Newcomb Problem," in R. Campbell and L. Sowden (1985), pp. 251-255; originally published in *Philosophy and Public Affairs*, 8, no. 3, pp. 235-240.
Robert Nozick (1969), "Newcomb's Problem and Two Principles of Choice," in N. Rescher et al., eds., *Essays in Honor of Carl G. Hempel*, Dordrecht, Holland, D. Reidel, pp. 114-146; reprinted in abridged form in R. Campbell and L. Sowden (1985), pp. 108-133.
Nelson Pike (1989), "Divine Omniscience and Voluntary Action," in J. M. Fischer (1989), pp. 57-73; originally published in *The Philosophical Review*, 74 (1965), pp. 27-46.
Alvin Plantinga (1989), "On Ockham's Way Out," in J. M. Fischer (1989), pp. 178-215; originally published in *Faith and Philosophy*, 3 (1986), pp. 235-269.
William Poundstone (1988), *Labyrinths of Reason*, New York, Anchor Press, Doubleday.
R. Stalnaker (1968), "A Theory of Conditionals," in *Studies in Logical Theory, American Philosophical Quarterly Monograph Series, no. 2.*
Max Weber (1930), *The Protestant Ethic and the Spirit of Capitalism*, tr. Talcott Parsons, London, Unwin Hyman.