ELSEVIER Annals of Pure and Applied Logic 96 (1999) 117–140

ANNALS OF
PURE AND
APPLIED LOGIC

Self-reference and the acyclicity of rational choice

Haim Gaifman [1]

Philosophy Department, Columbia University, New York, NY 10027, USA

Received 18 April 1997; received in revised form 20 June 1997; accepted 12 March 1998

Communicated by S.N. Artemov

Abstract

Guided by an analogy between the logic of truth and the logic of a rationally choosing agent, I propose for the latter a principle of acyclicity, which blocks paradoxical self-referring reasoning. Certain decision-theoretic paradoxes are used to illustrate what can happen when acyclicity is violated. The principle, however, is argued for directly on grounds of coherence. Among its consequences are certain decision-theoretic rules, as well as a guiding line for setting Bayesian prior probabilities. From this perspective I discuss in the last two sections Prisoner's Dilemma and Newcomb's Paradox. © 1999 Published by Elsevier Science B.V. All rights reserved.

Keywords: Decision theory; Self-reference; Agency; Rational choice; Prisoner's Dilemma; Newcomb's Paradox

1. Introduction

Self-reference in semantics, which leads to well-known paradoxes, is a thoroughly researched subject. The phenomenon can appear also in decision-theoretic situations. There is a structural analogy between the two and, more interestingly, an analogy between principles concerning truth and those concerning rationality. The former can serve as a guide for clarifying the latter. Both the analogies and the disanalogies are illuminating.

Self-reference in situations of choice takes place when the very choice of a certain option affects the satisfaction of the norm that guides the choosing. Paradox ensues when, for each available option, the very fact of its being chosen implies that the guiding norm has not been satisfied. I shall present and analyze two paradoxes of that kind. Following which I will suggest a scheme for representing choosing situations from

[1] Thanks are due to Isaac Levi for illuminating exchanges on topics relating to this paper, and to an anonymous referee for comments that have led me to revise some items in Section 2 and to clarify some arguments in Section 4.

PII: S0168-0072(98)00035-9

the agent's perspective, and a principle of acyclicity for choices, which is analogous to principles adopted by Russell and others, in the context of the logical paradoxes. The principle can serve as a general guiding rule. It implies certain dictums, concerning the prediction of one's own choices, which have been proposed by Levi and Spohn. In the case of Bayesian probabilities, it imposes constraints on the choice of prior probabilities. Section 4 in which the principle and its corollaries are presented is the heart of this paper.

Considerations relating to cycles enter also in the analysis of famous puzzles: Prisoner's Dilemma and Newcomb's problem, to which the last two sections are devoted. To be sure, other aspects, besides cyclicity enter in a crucial way. But viewing these issues in the context of cyclic reasoning makes for a clear analysis.

There is a tendency to bring into the discussion of choice the question of free will. One can however unlink the metaphysical issues by focusing on the agent's perspective. The scheme I am proposing does something of the sort. We bypass thereby the metaphysics of free will, instead of which we investigate the logic of agency.

I should emphasize that the acyclicity principle is not offered here as a remedy for marginal pathologies, but as something inhering in the very logic of a rationally choosing agent. Paradoxes are philosophically interesting in as much as they indicate fault lines in certain modes of thinking. In attempting to solve them we may find principles whose validity extends much beyond the specific, but which could have been easily missed, were it not for the contradictions that arise in extreme situations. The contradictions serve as corroborative evidence for a principle that has, upon reflection, strong direct appeal.

2. The Liar, Strengthened Liar, and the Irrational Man Paradox

Semantic paradoxes arise when some sentence attributes to itself a semantic property – such as being non-true, or false. The simplest is the classical Liar, where the self-reference is direct: 'What I am saying now is false'. The indirect variants involve sentences that attribute semantic properties to other sentences in a way that creates loops. The possibilities are endless. In the classical Liar self-reference is achieved through the indexicals 'I' and 'now'. But it can be achieved without indexicals and this is a cleaner way. The general form is

s is false,

where 's' stands for a description that picks this very same sentence; e.g., 'the sentence written in ... at time ___', with '...' and '___' non-indexical specifications of place and time, such that the sentence written in that place, at that time, is: $\ulcorner s \text{ is false}\urcorner$. We thus get a sentence s such that

$$s = \ulcorner s \text{ is false}\urcorner. \qquad (1)$$

In other versions, the equality, '$=$', is replaced by some strong equivalence. 'Not true' instead of 'false', will serve as well, and is preferable for certain purposes:

$$s = \ulcorner s \text{ is not true} \urcorner. \tag{2}$$

By elementary reasoning, each of the assumptions – that s is true, and that s is false – leads to contradiction. At this point the paradox can be blocked by classifying s as a truth-value gap: neither true nor false. The paradox can however strike back. Striking back is straightforward in the case of (2): If s is a truth-value gap then, it is not true, but this is exactly what s says, hence s is true after all, and we are back in the loop. A similar, though less obvious, move is available for (1): If s is a truth-value gap, then it is not false; but s says that s is false, hence s is false after all; in the loop again. The striking-back version has been pointed out by van Frassen [14], who called it the Strengthened Liar.

Now a careful analysis can dispose of the Strengthened Liar as well (a fact which is often missed). A sentence that satisfies either (1) or (2) is not merely a sentence that *happens* to be a gap. It cannot, for linguistic reasons, be anything but a gap. Therefore s does not say *anything*. Anything, that is, which can have a truth-value.[2] It is not meaningless as 'bla bla bla', but it does not *have* truth-conditions. The argument of the Strengthened Liar presupposes however that s says something (namely: that s is false – in (1), that s is not true – in (2), and proceeds to evaluate it as if ordinary truth-conditions apply. The presupposition failing, there is no paradox.

That is not the end of the story. For *we* want to say that s is not true, or that s is not false, as I have been doing right now. But then we find ourselves uttering or writing sentences lacking truth-value: s itself – in the case of (2)), the negation of s – in the case of (1). The paradox now consists not in a contradiction but in a strange limitation that prevents us from expressing an obvious truth in our language. As a rule, the distinction between the unable-to-say paradox and the Strengthened Liar is not clearly made; 'Strengthened Liar' serves for both. I myself used somewhat unclear terminology in [4], where I solve unable-to-say paradoxes (without using metalanguages). I referred to them by 'Strong Liar', though 'Metaliar' would have been more apt. The Metaliar, which is essentially a linguistic paradox, does not have, as far as I can see, decision theoretic analogs.

Our concern, recall, is with self-reference arising in the context of decision making. The self-reference is achieved through conditions that insert the decision itself, the

[2] Usually, a gap occurs when certain factual presuppositions are not fulfilled. A person who asserts 'The king of France is bald' presupposes that there is someone who is the king of France. We may say that this is part of what the sentence implicitly says. No similar presuppositions are available in the case of the Liar. We may try, for (1), the "presupposition" that s is not cyclic e.g., a person who asserts 'The sentence written on the blackboard in room 10 is false' presupposes that what is written there is not this sentence. But it would be strange to regard it as part of what the sentence says, even implicitly. This becomes more apparent when we consider indirect self-reference that involves many sentences. The presupposition of each of the sentences would have to refer to all the others. Finally and conclusively, consider the classical Liar. Somebody asserting 'What I am now saying is false', would have to presuppose as part of his assertion: "What I am saying now is not 'What I am saying now is false' "

fact that this action was chosen, as a factor that affects the norm that determined the choice. A paradox is generated when, for every possible choice, the very fact that it was chosen (and performed) undermines the very norm under which it was chosen.

Now Russell [10] traced the source of the Liar, and many other logical paradoxes, to the fact that certain propositions refer in an illegitimate way to themselves. He proposed to prevent it by imposing severe constraints in the form of a hierarchy of types. The theory, of course, was not proposed merely as a device for blocking paradoxes. He argued for it directly, on grounds of coherence. And although the area has undergone radical changes[3] there remains an important kernel of truth to Russell's view. In a nutshell it is this. The truth-value of a sentence of the form 'The sentence written in ... is ___' , where '___' stands for a semantic predicate, derives, by definition, from the truth-value of the sentence that is written in ... . If what is written there is this very same sentence, there are no truth-conditions; for we enter a non-terminating cycle, or loop. Seen thus the Truth Teller (where '___' stands for 'true') is as faulty as the Liar; although it does not lead to contradiction, it leads to an unending loop. In a somewhat analogous spirit, I shall propose an acyclicity principle that imposes constraints in modeling situations of choice.

2.1. *The Irrational Man*

The following is a variant of a puzzle that appeared in [2], based originally on an idea of G. Schwartz. Mr. Z. offers Adam two boxes, each containing \$10. Adam can choose either $S1$: to take the leftmost box and get \$10, or $S2$: to take the two boxes and get \$20. Before making his decision, Adam is informed by Z. That if he acts irrationally, Z. Will give him a bonus of \$100. It goes without saying that Adam is likely to ask himself whether this a hoax, whether Z. Is trustworthy, what is Z.'s view of rationality, etc. To eliminate these noise factors, assume that Adam believes that Z. is serious, has the relevant knowledge, is a perfect reasoner and is completely trustworthy.

Without the extra piece of information, $S2$ is the rational choice. Adam may therefore go for the bonus by "irrationally" choosing $S1$. But given the present conditions, choosing $S1$ – under the expectation of bonus – would be highly rational. This reasoning behind him, Adam may try $S2$ as the "irrational" choice. But the same argument applies again; to choose $S2$ is rational, if he expects thereby to satisfy the bonus condition.

It might appear as if Adam tries, and fails, to choose *irrationally*. Actually it is the flip side of his failure to choose rationally, that is: to choose in a way that brings (according to his beliefs) the bonus. He fails because by choosing in this way he violates the bonus condition. The choice cancels its own rationality. The paradox can

[3] There is a vast literature on semantic self-reference, which I do not refer to here, as the topic is not the subject of the present work.

be represented in terms of regret: No matter what Adam chooses, he is going to regret it immediately after the choice, because the other choice would seem better.

To sharpen a bit the underlying notion of rationality, consider two preliminary characterizations of rational choice: the strong, (SR), and the weak, (WR). We shall not need, for the purposes of this paper, to go beyond these preliminary notions.[4]

(SR) For every alternative, y, the agent believes that he, or she, does not stand to gain more by choosing y. Formally, if the choice is x, this amounts to: $\forall y\{\mathrm{Bel}[\neg(y \text{ is better than } x)]\}$

(WR) For every alternative y, the agent does not believe that he, or she, stands to gain more by choosing y. Formally: $\forall y\{\neg\,\mathrm{Bel}[y \text{ is better than } x]\}$

Accordingly, we have two notions of *irrationality*. The strong, (IR), is the negation of (WR). The weak, (NR) – also called non-rationality – is the negation of (SR).

(IR) For some y, the agent believes that he, or she, stands to gain more by choosing y. Formally: $\exists y[\mathrm{Bel}(y \text{ is better than } x)]$

(NR) For some y, the agent does not believe that he, or she, does not stand to gain more by choosing y. Formally: $\exists y\{\neg[\mathrm{Bel}(\neg(y \text{ is better than } x)]\}$

Rationality, irrationality and non-rationality play analogous roles to truth, falsity and non-truth. The case of non-rationality and non-irrationality is like a truth-value gap. It takes place when the agent lacks the belief that the choice is optimal, and lacks the belief that it is not optimal. (Here 'optimal' means that no other choice is better.) In Mr. Z.'s statement, 'irrational' can be interpreted in the strong sense (the Liar with 'false'), or in the weak sense (the Liar with 'non-true'). Either way we get a paradox; because acting under the belief that the action gets one the bonus makes this action rational, which implies (via the bonus condition) that one should have acted otherwise.

Having figured all this out, Adam infers that there is no way of satisfying the bonus condition. The 'rational' and 'irrational', which occur in Z.'s statement, do not apply to Adam's rational choices. There is a "rationality gap". Adam can therefore disregard the promise and choose $S2$. He is however not done yet, for the paradox can return: Given Adam's reasoning, $S2$, with its higher payoff, is the only rational choice. Choosing $S1$ would be irrational; hence there seems to be, after all, a way of satisfying the bonus condition. Back in the loop. This is the analogue of the Strengthened Liar.

[4] Thus (a point made by the referee) we assume that the agent is fully apprised of, and can weigh all the options; this need not hold in complicated situations, such as chess games, where the agent may not be aware of all the options at hand. Other complications arise when different epistemic levels concerning the ranking enter the picture. For example, the agent may believe that, of the three available options, either OP1 or OP2 is better than OP3 (on some more "objective" scale of preference), without having any belief as to which this option is. Such an agent who chooses OP3 comes out as "weakly rational" on the proposed definition. But this is justified only if the agent has some other beliefs concerning the ranking. E.g., he can believe that whatever option is better than OP3, OP3 is better than the remaining option; this can justify the choice of OP3 on risk-aversion grounds. Arguably, by factoring in risk-aversion, OP3 can be construed as the preferred option. Such problems and, in general, the constraints one should impose on the ranking go beyond the concerns of this paper.

The strategy for blocking the Strengthened Liar works also here. Mr. Z.'s statement does not describe any condition at all. Any condition, that is, that can figure in the rational agent's deliberations. Though Adam believes that $S1$ is inferior to $S2$, he cannot infer that the choice of $S1$ satisfies the bonus condition. There is nothing to satisfy. In the context of choosing this condition is not up for evaluation.

It appears, moreover, that this is the only strategy that lets Adam off the hook. If not eliminated, Mr. Z.'s statement strikes back at any rational choice; there is always the rejoinder: If this be rational then choosing otherwise would have satisfied the bonus condition. The rejoinder has no force only if any other choice is rational as well; which would imply that the bonus condition is not satisfiable. Yet, when *that* is taken into considerations the obvious rational choice is $S2$. So there is, after all, only one rational choice and we are cycling again in the loop.[5] Actually, $S2$ *is* the only rational choice. But this can be stable only if Z's statement is disqualified as having no standing in the deliberation.

Some suggested to me that the Irrational Man is simply a case of an inconsistently specified game (hence there is no paradox). The argument runs as follows. Since getting the $100 will maximize Adam's utility, the bonus condition, (BC), appears to imply:

(B′) Adam will maximize his utility iff he will act so as not to maximize his utility.

(We assume, of course, that the bonus is given only if (BC) is met.) And this is like stipulating that Adam wins the game iff he tries not to win, which yields the contradiction: Adam tries to win iff he tries not to win. But (BC) implies (B′) only if it implies that Adam can maximize his utility by gaining $100. And, as the above analysis shows, such is not the case. For a rationally choosing agent, (BC) does not provide any option of gaining the bonus. (BC) appears, but fails, to describe such an option, just as the Liar sentence appears, but fails, to make a (truth-valued) proposition. (BC) is not contradictory but inert, an assertion by Mr. Z. that Adam, *qua* rational agent, cannot use. Adam *can*, however, maximize his utility (and 'win' the game) by taking the $20. The game is therefore consistent.

Without going into the general question of what makes a paradox a paradox, let me suggest that the Liar qualifies for the title, because we cannot dismiss the sentence as false or self-contradictory. The reason for the paradox is that a very simple sentence cannot have truth-conditions. Which is a surprise. Similarly, the bonus condition in the Irrational Man is neither self-contradictory, nor does it render the situation incoherent. Mr. Z. made the statement and Adam is called to make something of it in the context of choosing. Z.'s statement does not (as I heard it claimed) render the situation chaotic, where 'all bets are off'. Adam still has his options and $100 are still better than

[5] For example, the following "solution" was suggested to me. Finding, after deliberation, that he does not know that one of $S1$ or $S2$ is better than the other, Adam invokes the rule that in such cases it is rational to toss a coin. So he tosses a coin, chooses accordingly, and loses the bonus. On this construal, coin tossing is a third option, call it $S3$. But then, following the line of reasoning just given, Adam can deduce that by choosing (rationally) $S3$ he will not get the bonus. Which, however, implies that $S2$ is preferable to $S3$.

$20, which are still better than $10 (that utilities are measured in dollars is of course unessential, you can write the scenario with any three prizes you prefer).

The essential difference between the semantic and the decision-theoretic cases is this. The Liar sentence says nothing that can be evaluated for truth-value. But the bonus condition in Z.'s statement has truth-conditions, and once Adam has chosen it can be evaluated. Has he chosen rationally (thinking that his choice is optimal) the condition is not satisfied. But it *is* satisfied if he has chosen irrationally, believing that his choice is not optimal (or not believing that it is optimal, if 'irrational is read as 'non-rational'.) And he could have done this. He could have chosen by whim, because of a feeling, a mood, or for no reason. The question how irrational choice is possible, what constitutes such a whim, impulse, temporary incoherence, weakness of will, or what have you, does not concern me here. I take it for granted that there will be cases which we shall characterize in this way (else 'rational' becomes a vacuous constraint). And if Adam chooses in this way he qualifies for the bonus, and will probably be surprised when he gets it. It is only from the perspective of *Adam* qua *deliberating rational agent* that the bonus condition must be excluded as meaningless. The case resembles somewhat Moore's paradox. Everyone, besides Moore, can assert: "There is a finch in the garden, but Moore does not believe it". But Moore cannot assert it without being incoherent.

2.2. The Truth Teller and the Rational Man

The anti-twin of the Liar is a statement asserting its own truth: 'The sentence I am uttering now is true' (when this is all that the speaker says). The Truth Teller does not lead to contradiction. But if the Liar has no truth-conditions on grounds of cyclicity, so does the Truth Teller. The decision theoretic parallel of the Truth Teller is the puzzle of the Rational Man. Presented with the two boxes, Adam is told that, if he acts rationally, Mr. Z. will give him a bonus of $100. So Adam chooses $S2$ (he takes the two boxes), expects the bonus and gets it. Adam is right in expecting the bonus; nonetheless the bonus cannot constitute any additional inducement for his action (and this is the puzzle). He should not try any harder. If opening the second box requires some effort, he should not expend any effort worth more than $10 (like hiring someone for $11 to crack it for him). If he does so, he acts non-rationally and forfeits the bonus. This shows that the bonus condition has no place at all in Adam's deliberation.

The Irrational Man, which is the most direct case of self-reference, is blocked by identifying certain information as meaningless in the context of choice. Thereupon the information is expunged. In the next paradox, which is more "realistic" and not so direct, the expunging is less obvious.

3. Cassandra's paradox

Cassandra, a prophet of doom, used to warn people against disastrous actions, but her warnings went unheeded. She was doomed to be disbelieved, by the same god who had given her the gift of foresight. And she knew it. (This is extremely likely: shortly after endowing her with foresight, Apollo spit in her mouth and she must have known what *that* meant.)

Imagine that, upon being asked for advice by some person, she warns the person against a certain action; but she also predicts that the person will not heed the warning. She makes thereby two predictions: that a certain action will have bad results, and that the person will take this action. If the person knows also of the second prediction, he may find himself in a dilemma. Suppose he weighs Cassandra's opinion according to her known performance. Her record being what it is, he inclines to accept her advice. In fact, the advice *would* be accepted, except for a strange complication. By accepting it he falsifies her second prediction, damaging thereby her performance record. It is quite possible that on the basis of the new record the advice will *not* be accepted. Yet, if it is rejected, Cassandra's record becomes even better than at the beginning; so the advice *should* be accepted. And so on ad infinitum.[6] As in the Irrational Man, the paradox can be represented in terms of inevitable regret. Whatever choice the person makes will seem, after being made, the wrong choice.

Let us modernize the story. Eve is asked by Clive – a friend but not a close one – to lend him, on short notice and without security, $1000, which he urgently needs to complete a deal. He tells her that he will receive in a couple of weeks a large sum and will return the loan with additional $500 interest. Eve postpones her answer and deliberates. Given her beliefs at that stage, she inclines to grant his request. Later, at a party, she meets Cassandra, known for psychological insights and successful predictions of people's behavior. Cassandra knows both Eve and Clive and she warns Eve against lending the money. Clive's promise, she says, will not, or is unlikely to, be kept (the details depend on the variant of the story). Cassandra goes on to predict that Eve, being the good soul she is, will lend the money nonetheless. Eve learns of the second prediction; either from a third person or from Cassandra herself (this can make a difference, which however is not relevant to the main point, cf. footnote 8).

There are several possible completions of the scenario, depending on Eve's decision procedure. Assume that Eve maximizes expected utility and that she has a method for combining probabilities given by experts, cf. [3]. Eve has a formula for assigning weights to experts according to their records. Let $L1$ and $L2$ be, respectively, the choices of lending and not lending the money. Let Eve's utility for $L2$ be 0. And let her utilities and subjective probabilities for the possible outcomes, given $L1$, be such that her expected utility for $L1$ is, initially, positive. Hence she should choose $L1$. When Cassandra enters the scene, she provides her own probabilities. Eve treats Cassandra as an expert, and the earlier stage of herself – as another expert. Cassandra, of course, is given a much higher weight, e.g., 0.8 (leaving 0.2 for the earlier Eve). All in all, the combined probabilities are such that the expected utility is negative: Eve

[6] I proposed this paradox (with a different cover story) in a talk given in a conference on the Foundations of Probability and Causality, Irvine, summer 1985. The published paper [3] was devoted to higher-order probabilities and did not include the paradox, nor did I publish it later.

should choose $L2$. If she chooses $L2$, Cassandra's prediction being falsified, her relative weight decreases; say, it is now 0.7. With the new weights, the combined probabilities yield positive expected utility: Eve should choose $L1$. If she chooses $L1$, thereby making true the prediction, Cassandra's weight increases, it is now higher than 0.8 (because an additional prediction has been fulfilled); the expected utility is negative: back to $L2$.[7]

There is also another, purely Bayesian, way of modeling this story, which I shall outline in Section 4.

The paradox arises in every decision theoretic setup in which an expert's opinion is evaluated according to the expert's record. In a different scenario Eve is a no-nonsense person who has no time for intricate Bayesian introspection. She uses a simple rule of thumb: Follow the expert's advice, an expert being anyone with at least 80% success rate in 5 or more predictions. If there are several experts she goes by the majority. If they are evenly divided she tosses a coin. If there are no experts around she goes by her feeling: what her heart tells her (she'll lend the money). Now assume that Cassandra has eight predictions in her record, seven successes and one failure, and see what you get.

Now the paradox arises because Eve's choice changes Cassandra's reliability as expert. That choice enters as follows: Cassandra is supposed to know, or have good estimate of, how one is likely to behave. Eve's choice confirms or infirms Cassandra's expertise by according or disaccording with her prediction. But while this is perfectly plausible when it comes to choices that serve to assess her expertise, it cannot apply to the choice in which that expertise enters as a reason. I cannot use an expert's advice as a guide to my choosing, and at the same time use my choosing as evidence for the expert's reliability. Seen thus, the backward looping arrow by which Eve's choice changes Cassandra's reliability is severed. That is, Cassandra's second prediction has no place in Eve's deliberations.

Eve therefore concludes that the second prediction cannot be used. Consequently, she chooses $L2$. There is no ground for regret, at least no rational ground. Eve should reflect that Cassandra's second prediction did not *state* anything that could enter as a factor into the deliberation. The failure of the prediction will enter into the future record, but this is another story.[8]

[7] Further details can, for example, be as follows. There are three possible outcomes to lending the money. $C1$: Clive keeps his promise. $C2$: Clive returns the loan without the interest. $C3$: The money is lost. Eve's utilities for these outcomes are 13, 7, and -16. Her initial subjective probabilities for them are 0..7, 0.2, 0.1. Cassandra's probabilities are: 0.2, 0.3, 0.5. Mixing with weights (0.2, 0.8) yields the probabilities: 0.3, 0.28, 0.42, for which the expected utility is negative. With weights (0.3, 0.7), the mixing yields: 0.35, 0.27, 0.38, and the expected utility is positive.

[8] All of which does not mean that Eve should ignore Cassandra's *making* the second prediction. If Cassandra herself tells it to Eve, the latter may find it strange that Cassandra should confuse her by giving out a piece of unusable "information". It may even go to detract from her standing ("what kind of person is she?"), thereby affecting Eve's choice. But there will be no paradox, because even if the prediction comes true, Cassandra's telling it to Eve remains strange. There are various ways in which Eve may interpret Cassandra's second act of prediction. But whatever the interpretation and whichever its effect, it cannot enter into the deliberations as a forecast.

3.1. *Jonah's puzzle*

The antithesis of Cassandra is Jonah, the prophet who knew that people will believe him. (It didn't make his life easier though; because they believed his prediction, people have acted so as to falsify it. But this Hebrew tangle is another story.) Suppose that, having advised me, by some sort of forecast, Jonah adds "And I know you'll take my advice". So by taking his advice I enhance his expert standing, which makes me even more assured that I did the right thing. I have an additional motive for taking his advice! The motive may even be crucial: I would not have taken his advice, were it not for his second prediction.

Cassandra and Jonah stand to each other like the Liar and the Truth Teller. Jonah's second prediction, like that of Cassandra, does not express anything that can figure in the considerations of the agent whose choice is being predicted. But, unlike the Liar and the Truth Teller, these predictions of forthcoming choices are ordinary statements, with ordinary truth-conditions. It is only in the court of the agent's deliberations that they have no standing.

3.2. *Foreknowledge of choices*

The rule exemplified in these stories is, I believe, true in general. Forecasts of the agent's choices cannot enter (as forecasts) into the agent's reasoning. This is directly related to a thesis of Levi [6, Ch. 4] that an agent cannot be construed as knowing in the course of deliberation what the choice will be. Knowing the choice constitutes the end of deliberation.[9] By the same token, the agent cannot rule out any of the feasible options; ruling it out means that the option is not feasible. Spohn [12, 13] claimed that one should not assign probabilities (determinate or indeterminate) to events describing one's choices. Spohn's view is supported by Levi [2] as a natural extension of the not-knowing-ahead-one's-choices thesis. In its full generality the thesis means that, whatever information one uses in one's deliberations, one cannot use any non-trivial information about the likeliness of what one will choose. The thesis, we shall see, follows from a general acyclicity principle that excludes vicious self-reference.

4. The acyclicity principle

4.1. *Preliminaries*

First, some clarifications concerning the concept of choosing. Say I am deliberating between A and B. My preferences and estimates being clear, it remains to compute a value. In computations I am reliable but slow. My friend C., who knows my estimates

[9] Levi's position is related to a line of thought due to Shackle [11]. The latter emphasized the open-ended nature of "deliberations", the impossibility in principle of circumscribing ahead the considerations that may enter therein.

and preferences and who is very good at computing, tells me: "I know you'll choose A". She has computed faster than I. Knowing *that*, I choose A without completing my slow computation. There is nothing incoherent here. I have not chosen to make her prediction come true. My choosing was already done: I chose the option determined by a certain mathematical condition. Then I chose to shortcut the implementation by "using" C. as a computing device. The same would apply had the choosing been a consequence of a logical deduction – in as much as the deduction comes under "computation".

It might appear that if I decide to choose by tossing a coin, then I, as anyone else, can assign probabilities to what I will choose. But this is a wrong construal of the situation. The choice has already been made, namely: to randomize by coin tossing. The rest is implementation. Suppose Eve decides to toss a coin and Cassandra predicts the outcome. Eve might change the coin suspecting that Cassandra knows that the coin is biased. Or she might go on, explaining Cassandra's prediction on other grounds. There is no paradox here, Cassandra's prediction is not about choosing but about a coin landing.

Moreover, and this is quite important, many choice situations that do not involve external randomization should be construed along this line. Such are cases where one goes by one's "feel", intuitive judgment, momentary impulse, and the like. The no-nonsense Eve decides in certain cases to go by her feelings; *that* is her choice. She implements it when she acts according to what she feels. Absent an external randomizing device, one can resort to internal randomizing: choice on impulse; one can do this when one's strategy against an opponent calls for randomization. (Here is a mental lottery that implements 'yes' with probability 1/3: Pick in your thoughts "randomly" two different numbers x, y between 1 and 100; compute the first two decimal digits of x/y and add them; answer yes if the result is divisible by 3. If this is not good enough, increase the range to 1000 and sum the first four digits of the ratio.) Some people, I am told, can recognize on immediate inspection the sex of small chicks. Employed to sort chicks, such a person should not be construed as an agent that decides each case. She, or the person employing her, have made only one decision: to trust to her recognition ability.

Known, or believed, past performance can enter into the deliberation ("I know from experience that I tend to judge right in these situations"). To be sure, very often the line between deliberation and unthinking intuition is hopelessly blurred. Yet the theoretician can draw a line, even arbitrarily, for the sake of a general coherent account.

Someone who estimates the probability of his own pending decision (accepting a job offer, moving to a new country), can be construed as one who has chosen to delegate the deciding authority to a partner that acts by feel, inclination, the pull of certain forces, and the like. Being intimate with that partner and having observed him for some time, the agent believes (perhaps mistakenly) that he can assess the partner's propensities.

Finally, a note of caution. By excluding choosing-on-impulse from the deliberative stage we do not exclude it from the class of purposeful doings. (We are still a long way from knee-jerk reaction where the person can no longer be regarded as an agent.) Yet, choosing-on-impulse can be shifted to the implementing stage and considered as "external" to the deliberation, in as much as the agent can reason *about* it.

4.2. C-sentences and the principle

Underlying the logic of truth are T-sentences of the form

(T) $\ulcorner A \urcorner$ is true if and only if p,

where A is some sentence and 'p' stands for its truth-conditions. Paradoxes may arise when A states, directly or indirectly (via a loop), something about its own truth. For various purposes, e.g., in setting up a recursive definition of truth for a formal language, the 'only if' part is omitted. The recursive clauses of a truth-definition have the form:

(T′) $\ulcorner A \urcorner$ is true if p.

The extension of 'true' is determined by iterating these clauses. The 'only if', for the appropriate right-hand sides, is then derived as a theorem.

By analogy, the logic of choice can be given by the following scheme. Call it a C-sentence:

(C) I make $\ulcorner A \urcorner$ true, for reasons R,

where A is a description of a state of affairs that the agent can bring about and 'R' stands for the agent's reasons for the choice. (C) means that R constitutes a *sufficient* reason. It is the analogue of (T′). We can formulate an analogue of (T) by adding 'and only for reasons'. In that case R amounts to a disjunction of groups of reasons, each of which is sufficient for the choice of A. But (C) is more natural in the context of a deliberating agent who is to choose from a list of options; a sufficient reason is all that the agent needs. Allowing myself a less scrupulous style, I shall use 'I choose A' as ellipsis for 'I choose to make $\ulcorner A \urcorner$ true', I shall often omit the corners and I shall use other self-explanatory phrasings.

It is understood that if one chooses A then one actually makes A true. But we should clearly distinguish between making A true and *choosing* to make A true. The reasons R enter into the choosing process and it is they that determine whether some norm, such as rationality, is satisfied. Presumably, R is cast in some language, or system, which allows one to specify reasons for choice. At this stage we do not have a formalized version that can serve as a guide. So R is phrased in natural language, supplemented as the need may be, with some formal apparatus such as the setup of a probability space.

In giving reasons for choosing A, we do, as a rule refer to A. But we should not, on pain of creating a bad cycle, refer to the choosing itself. Provisionally, the acyclicity principle is:

(ac) R may refer to A, but it should not refer, directly or indirectly, to the choosing of A.

There is, however, an important proviso. My choosing of A can be for the purpose of *letting someone else know that I choose A*. Hence, it *can* be referred to in R inside a that-clause that expresses someone else's propositional attitude. For example, I go to the airport to meet a friend. There is no pressing need; she does not need my practical help (she might do better without it) and we can meet later in town. My main purpose is to express affection by showing her my choice. So I come to meet her in order that she know that I have come to meet her. This is the principle that underlies tokens of friendship, as well as hostility, and communicative acts in general.

Let us say that a choosing by X *signals* to Y if it affects Y's beliefs about that choosing in a way that is significant in X's reasoning. The proviso exempts self-reference that enters via signaling. It is of course pointless to a signal to oneself.[10]

One can choose A for the sake of A, that is, for its own intrinsic value. But, I claim, A cannot be chosen for the sake of choosing A. Consider what it would take to make up such a scenario. Suppose it is in my power to make A true. Mr. Z. (whom I completely trust, etc.) endows A with "intrinsic" value of \$100 by telling me:

If you make A true I shall give you \$100.

If the sum is worth my trouble I make A true. My reason is A itself (actually, the \$100 that go with it, but we can ignore the difference). Now Mr. Z. endows the *choosing* of A with a value of \$100, by telling me:

If you choose to make A true I shall give you \$100.

My reaction: I make A true, in a way that is likely to convince Z. that I fully intend it. If necessary, I make a visible effort. My choosing of A is supposed to signal to Z. But I do not choose A in order to choose A. It would take something like the following to force me to do this: Mr. Z. has a device that shows my real attitude (say he clamps it onto my head) and he can find whether my making A true is intentional or not. But even if the scenario is admissible, it is incoherent to require of me that I try to manipulate that device by producing in myself the state of choosing A. I cannot try to have an intention in order to have this intention.

If A is morally good, then to choose it for the right reason – *because* it is morally good – is to act morally. I can try to make A true, but I cannot try to make-for-thc-right-reasons A true. Either my reasons are right or they are not. I can bring about A, but I cannot bring about my reasons. Similarly in the case of avoiding. I try to avoid hurting my friend. (I try to make true: I do not hurt my friend.) If I hurt my friend

[10] An agent who does this should be construed as two persons that share the same payoff. Such scenarios can be used as literary devices to mark self-alienation. In Goddard's *Breathless* the anti-heroine turns her lover over to the police, in order to convince herself that she does not love him. There is also the more common phenomenon of choosing in a certain way in order to preserve one's self-image. This, as well, marks a certain cleavage in the agent's unity.

unintentionally, I will be less culpable, or – depending on the circumstances – not culpable at all. But it is incoherent to distinguish between trying not to hurt someone and trying not to hurt someone intentionally. One can try hard, or not so hard, but this is not the same distinction.

The same line applies to rationality. To choose rationally is to choose so as to secure an outcome that is more likely to rank higher on one's preference list. If I make A true for such and such reasons, my action is rational. But I cannot make my reasons to be of the "right" kind.[11] Now the story may not be as simple as that. We might need levels of reasons. Someone who finds it distasteful to be guided purely by mercenary considerations may be construed as invoking a second-level aesthetic evaluation in order to choose a first-level rule for choosing. And perhaps we should have, in general, levels of norms, so that one may apply a second-level norm in order to determine the desired first-level norm by which one chooses.[12] But all this does not undermine the validity of the acyclicity principle because a level applies to lower levels, not to itself. (The analogy with Russell's hierarchy of types, or with the hierarchy of metalanguages is obvious.)

It should be clear that "rationality" in this work is a minimal constraint, which can accommodate all kinds of preference. (It is only for the sake of simplicity that in the standard examples preference is measured by money.) A can be chosen because it is more interesting, challenging, exciting or what not. But then A is chosen just for these reasons (the interest, the challenge, the excitement, and so on). One does not chose it for the sake of choosing A; there is no cycle here. Some situations call for fine analysis. The act of choosing may itself carry some rewards, say, a feeling of being in control. But this presupposes that there is also a less "active" (do-nothing) option, and the more "active" A, is preferred because it involves doing. But then one chooses A, for the "doing" that goes with it, not for the sake of choosing A.

It has been suggested to me that the reason for choosing A can be one's belief that such a choice is truly free; e.g., one buys the less advertised item (though the more advertised ones may be better), because this is the way of manifesting freedom. Now in this work the metaphysics of free will does not enter; choice is, by the very logic of agency, "free". Still, I can make place for an agent who has such a belief. He, or she, employs a second-level norm that militates against "forced choosing", where "forced choosing" is choosing according to any of the ordinary criteria of utility, trendiness or what have you. By the samc token, onc can choosc in bizarre ways, in order to be original.

[11] The injunction 'be rational' is not a demand to have the right sort of reasons. Rather, one calls someone's attention to certain existing reasons that have been temporarily eclipsed. It is like saying 'come to your senses'. Similarly, one who says 'I decided to be rational' says only that considerations of a certain type have outweighed other considerations.

[12] This calls for modeling one's choice of A as consisting of: (i) choice of first-level norm, according to some second-level norm (ii) choice of A according to the first-level norm chosen in (i). All of which I ignore here.

One can also choose A, in order to impress someone else. Here indeed the choice of A enters as a reason; but this is a clear case of signaling. To choose A in order to impress oneself, or to evoke self-satisfaction ("I like to think of myself as the kind of person who prefers A") can be construed as targeting one's future stages, or along the lines of footnote 10.

If the above arguments are valid, then by the same logic we should rule out any cyclic reasoning in which the choosing of A enters into the reason R, except when the choosing is relevant as the object of a propositional attitude of another person. Essentially, this is a coherence requirement; the paradoxes only help us to uncover it. Sometimes the reference to the choosing of A is indirect or disguised. But if it is essential to R, i.e., if R cannot be recast so as to bypass it, then it counts as cyclic. Furthermore, the rule generalizes to choosing from a set of available options:

(AC) The reasons for choosing A can refer to each of the available options, but they cannot refer in an essential way to the *choosing* from these options, except through considerations of signaling.

Signaling now covers the choosing of any one of the options, which enters as the object of someone else's propositional attitude, in a way that is relevant to the agent's reason. For example, one can choose an option for the reason that choosing the alternative can be interpreted as a hostile act. Note that the agent can refer freely to choosing by other people, to his past acts of choosing, or to his future ones.

Since the reasons for making A true are instrumental in the choice, any non-trivial estimate of the probability, or the likeliness, of A being chosen by the agent, constitutes a reference to the choosing. The reference is implicit but it is there. AC therefore implies that one should not, in the course of deliberation, estimate the likeliness that one will choose A. Spohn's injunction against assigning probabilities to choices follows in this way. Levi's position against foreknowledge of what one will choose, or not choose, follows as well. This generalizes further:

(AC^*) One should not use conditional probabilities (or likeliness estimates) of choices, which are obtained by conditionalizing on some event (or parameter) upon which the choice, in the agent's judgment, has no bearing.

That the choice has no bearing means that it is considered irrelevant to the event (or parameter) in question. Such events (or parameters) can be subject to probabilistic estimates outside the choice context. Using which estimates, we may, by conditionalizing on these events, get estimates for the choices themselves. The estimates will be non-trivial in as much as the conditionalization does some real job. Else the conditionalization is, in principle, eliminable.

Let us call *choice events* events of the form: A is made true (by the agent), and let us call *given-world events* events on which choice events have no bearing in the agent's judgment. Then the class of all given-world events is closed under Boolean operations (finitary, or infinitary, depending on the modeling), and the agent can have likeliness estimates, or subjective probabilities, defined over that algebra, which are unrelated to his or her forthcoming choices. AC* amounts to a symmetric constraint, which prohibits any non-trivial probabilistic (or likeliness) connections between choice events and given-world events; such connections are banned because they enable non-trivial conditionalizations of the former on the latter.

Usually, "having no bearing" is determined by causality considerations. Thus, an event in the past (a proposition whose truth-value is determined previous to the deliberation) is considered a given-world event. But in extreme situations, such as Newcomb's scenario (to be presently discussed), causality in itself may not suffice to determine the events on which the choice has a bearing.

In the Cassandra scenario, Eve's choice does not affect Cassandra's expertise; it affects only what Eve *will know* of this expertise (the choice, which will verify or falsify the prediction, will be part of the record). Had she thought that Cassandra's expertise is actually affected by what she chooses, she could not coherently use her as an expert. In the loop-forming reasoning that leads to the paradox Cassandra's expertise and Eve's choice are not independent. The reasoning is therefore blocked by AC*.

Now a Bayesian setup opens possibilities for subjective probabilities that are defined over a comprehensive classes of events, including all choice events. The assignment of probabilities to choice events is, however, harmless, as long as the values that are assigned to the choices do not affect the deliberation. Most often they do not. Spohn observed that when we maximize expected utility, we need only the conditional probabilities $P(\ |A)$, where A ranges over the choice events. Moreover, if $A_1, A_2, \ldots, A_n$ is a division of the space into mutually exclusive choice events, then the probabilities assigned to the A_i's do not matter (as long as they are positive, which I assume). For we can redistribute the probabilities, via Jeffrey's conditionalization, without affecting any of the $P(\ |A_i)$. This shows that explicit violations of AC* are not likely. But it does not bar implicit violations, because a dependency between some choice event and a given-world event may inhere in the prior probability (such dependencies are unaffected by Jeffrey's conditionalization). The Bayesian modeling of the Cassandra scenario, to be presently discussed, is a case where it is easy to go astray via implicit dependencies. As a condition on the prior probability, AC* amounts to the following constraint:

(AC$^{\#}$) Choice events and given-world events are probabilistically independent.

In a way, AC$^{\#}$ is self-evident: If a choice event, say X, has no bearing on Y, then obviously we should have: $P(Y|X) = P(Y)$. But the point of AC$^{\#}$ is that the agent should decide the events upon which the choices have no bearing and then impose AC$^{\#}$ as necessary condition on his or her prior probability. Viewed thus, AC$^{\#}$ can be quite a substantial constraint. The Bayesian Cassandra is a case in point. And as Newocomb's problem will show, the determination of the given-world events can be quite tricky. It may be also of interest to note that in either of the case $P(Y|X) > P(Y)$ or $P(Y|X) < P(Y)$ we get some non-trivial bound for $P(X)$, in terms of $P(Y)$, where $P(Y)$, recall, is unrelated to choices. The bounds result from the following elementary inequalities:[13]

$$P(Y) - P(Y|-X) \leqslant P(X) \leqslant 1 - (P(Y) - P(Y|X))$$

Hence, if $P(Y|X) > P(Y)$, then $P(Y|-X) < P(Y)$ and the first inequality yields a non-trivial lower bound for $P(X)$. And if $P(Y|X) < P(Y)$, the second inequality yields a non-trivial upper bound. In each case, we violate AC*.

4.3. *The bayesian Cassandra*

The basic intuition underlying Cassandra's expertise can be roughly expressed by an inductive rule:

(In) The more successful Cassandra's predictions have been, the more it is likely that her present prediction will succeed.

The inductive rule is implemented in well-known ways by choosing a suitable prior probability. To do this we must be able to treat sentences that are uttered in making predictions as objects. We need also the predicate Pred(x) (Cassandra predicted x) as well as the truth predicates True(x) and False(x). Cassandra's record is given as a conjunction of sentences of the forms:

$$\mathtt{Pred}(s) \wedge \mathtt{True}(s) \qquad \mathtt{Pred}(s) \wedge \mathtt{False}(s)$$

Let B represent Eve's background knowledge. Then B contains (as a conjunct) Cassandra's record, as well as sentences stating the well known semantic relations between truth, falsity and negation. B also contains conjuncts describing Cassandra's first and second prediction, as well as the T-sentences for these predictions. The second prediction is $L1$ (Eve will lend the money). Hence B contains as conjuncts the sentences:

$$\mathtt{Pred}(\ulcorner L1 \urcorner), \quad \mathtt{True}(\ulcorner L1 \urcorner) \leftrightarrow L1$$

In evaluating the options Eve uses conditional probabilities of the forms

$$P(Y|L1 \wedge B) \quad P(Y|L2 \wedge B)$$

In order to avoid the paradox, the success or failure of the second prediction should not affect these conditional probabilities. The upshot is (I will not enter into the details) an independence constraint on the Bayesian prior:

$$P(L1|B) = P(L1)$$

(which implies also $P(L2|B) = P(L2)$). The effect of the constraint is that the success of Cassandra's second prediction is probabilistically independent of her record. This is

[13] To derive the inequalities observe that $P(X) = P(Y \cap X) + P(-Y \cap X) \leqslant P(Y|X) \cdot P(X) + P(-Y) \leqslant P(Y|X) + 1 - P(Y)$ (since $P(X) \leqslant 1$). This gives the second inequality. Apply this, with X substituted by $-X$ and $P(-X)$ rewritten as $1 - P(X)$, and you get the first inequality.

a far from trivial constraint; roughly speaking, it excludes the second prediction from the inductive rule.

The application of AC is much simpler in the case of the Irrational Man. *S*2 (the two-box option) is mandated by a straightforward reason: $20 is better than $10. The "counter-reason" is:

(c) If Adam chooses *S*2 for the straightforward reason, then his choice is rational. Hence, he forfeits the bonus, which he could have received by choosing *S*1.

Evidently, (c) refers to the choosing of *S*2 and the reference is essential. Therefore (c) is ruled out by AC. This leaves no other reasons in the field, except the straightforward reason. So Adam chooses *S*2. (Presented with the Irrational Man, one's natural reaction is to go for the "irrational" move *S*1. The first straightforward move is left unnoticed. But it is right here that AC blocks the paradox.)

In the Irrational Man, Mr. Z. is assumed to be a perfect omniscient reasoner. In other circumstances Adam may rationally try to outsmart Z. Suppose Z. has decided to reward irrationality, but did not tell it to Adam who has learnt of it from someone else. And suppose that Adam believes that Z. does not know that Adam knows. Then (c) can be rephrased as a legitimate case of signaling: Adam signals (deceptively) to Z. that by choosing *S*1 he is behaving irrationally. Deceptive signaling is, of course, useless if you deal with an omniscient reasoner.

5. Prisoner's dilemma

The first player chooses a row, the second – a column; the payoffs are given in the resulting square, the first number is the payoff of the first player, the second – of the second. ('*C*' and '*D*' stand for 'cooperation' and 'defection').

	C	*D*
C	10, 10	0, 11
D	11, 0	1, 1

Matrix 1

On the accepted view a rational player defects (chooses *D*). The argument: (I) one's action cannot affect what the other does, and (II) *D* is dominant: whatever the other does, one is better off by choosing *D*. If both players are rational they end in (D, D), though each would have done better in (C, C).

Some implicit assumptions should be clarified at the outset. It is presupposed that other considerations – such as friendliness, honor (one should not rat, one does not return defection for cooperation), or the common good – do not enter. Or, alternatively, that they have been included in the payoffs. E.g., the first player's payoff for outcome (D, C) includes the negative contribution of any qualms he will have upon finding that he has defected while the other has not. And *this* presupposes that such non-monetary factors can be evaluated by bounded amounts of utils (utility units). The alternative to this need not consist in assigning such factors an infinite value. One can be more flexible than that. The non-monetary factors can be given changing values that depend on the initial monetary-valued matrix. For example, a view that reflects Kantian considerations can be implemented thus. Given an initial matrix M, compute another matrix $K(M)$ and use $M + K(M)$ as basis for decision. The function $K(\,)$ represents a Kantian norm, say: the general good that all free-reasoning beings can consistently aspire to. On this view Matrix 1 can never constitute the final story. There is nothing 'irrational' in the view. Roughly, it avoids (D,D) by appealing to a different scale of values.

Assuming that Matrix 1 *is* the final story and that (I) holds, D is indeed the rational choice. There have been attempts to avoid the uncomfortable (D, D) without paying the price of value adjustment. Hofstadter [5] argued that a rational player who assumes that the other is rational too, should choose C on force of the following reasoning: The situation being symmetric, rationality dictates to each player the same choice (it is as if both were solving the same mathematical problem). Moreover each player knows this. Therefore a player knows that the possible outcomes are (C, C) and (D, D). When this restriction is imposed on the outcomes, we get a game in which the obvious choice is C. One might object that solving a problem by assuming that there is a unique solution is not guaranteed to yield a solution, unless that assumption is established independently; and one can doubt whether rationality implies a unique solution. But the main objection is that, even if we grant a unique solution, the reason for choosing C is cyclic. Spelt out in full it is:

(R1) Since I and player 2 are rational, either I choose C and player 2 chooses C, or I choose D and player 2 chooses D; the outcome is therefore (C, C) or (D, D) and the first is preferable. I should choose C.

Here the cycles are obvious: the references to my choosing C and my choosing D are essential. They cannot be replaced by references to C and to D; the rationality of the players is needed in order to imply how they choose and there is no other way to conclude that the outcome is (C, C) or (D,D) . It should come at no surprise that, by continuing to circle around, the player can contradict himself: Since player 2 is rational, he will – by the argument just given – choose C, but knowing this, D is preferable, since (D, C) is better than (C, C) . I should choose D.

Another cyclic argument that mandates C assumes that the players are sufficiently similar, say they are identical twins.

(R2) Since I and player 2 are twins, we reason alike; if I choose C he (very likely) chooses C for the same reasons; similarly for D. Therefore (C,C) or (D,D) is very likely. Given this, I should choose C.

The reasoning can be cast in the form of expected utilities, by interpreting "very likely" in terms of conditional probabilities. Again, one's choosing is referred to in an essential way; the similarity of twins is invoked in order to infer the sameness of their reasoning. And again, a self-refuting circle is obtained: Since player 2 is similar to myself, he (very likely) follows the reason I have just given for choosing C. He (very likely) chooses C; but then I am better off choosing D, so I should choose D.

Now it is possible to reconstitute (R2) as non-cyclic, by arguing from some other facts, without referring to reasons, that the players are likely to do the same:

(R3) Given facts F, it is likely that whatever option I realize, the other player realizes the symmetric option. Hence [with 'likely' appropriately specified] the expected utility of C is higher. I should choose C.

(R3) amounts to giving up, under the force of F, the independence assumption (I). What kind of facts can play this role? Lewis [8] suggested that the similarity of players is, by itself, sufficient. Lewis rejects the expected utility reasoning. On his view, the right grounds for preferring an option involve a counterfactual: D is better than C because had I chosen D my payoff would have been greater, the assumption being that in the counterfactual situation the other player chooses as before. His argument is proposed with the aim of prescribing counterfactual reasoning, rather than expected utilities, also in Newcomb's scenario. Let us take a closer look.

Let C_1 and D_1 be the events of the first player's cooperation and defection, and C_2, D_2 – those of the second player. Let c be each player's payoff when both cooperate, and d – if both defect. (So if player 1 defects and the other cooperates, player 1 gets $c+d$ and the other gets 0). By elementary algebra, the first player's expected utility is greater for C_1 than for D_1 if and only if

$$P(C_2|C_1)+P(D_2|D_1)>1+d/c$$

In Matrix 1 $d/c=0.1$. In a more extreme situation, where the first player's payoffs for (C,C) and for (D,C) are 1000 and 1001, $d/c=0.001$. Let $r=1/2[P(C_2|C_1)+P(D_2|D_1)]$, then the above inequality is equivalent to $2r>1+d/c$, i.e., $r>0.5005$. Lewis regards r as a measure of predictability from one player's behavior to the other's, and he argues that under mild similarity assumptions $r>0.5005$ is quite plausible.

This however is wrong. The point at issue is not the players' similarity, but the dependence of certain events. r is a combined measure of the dependence of C_2 on C_1 and of D_2 on D_1. You can see this by observing that the average of $P(C_2|C_1)-P(C_2)$ and $P(D_2|D_1)-P(D_2)$ is $r-0.5$, and that $P(Y|X)-P(Y)$ measures in a way the dependence of Y on X (it is 0 just when the events are independent). It would make no difference if the players were duplicates of each other, we would still expect $r=0.5$. If two identical coins are tossed (say in separate rooms) we expect the tosses to be independent. We expect – since the coins are identical – that in many rounds of simultaneous tosses the frequencies of heads will be very close. But we do not expect any correlation between the outcomes. It can however be shown that $2r-1$ is the correlation between the random variables heads_1 (heads for coin 1) and heads_2 (heads for coin 2); which is also the correlation between tails_1 and tails_2. Therefore, Lewis's assumption implies a correlation of 0.001. It would take very strong evidence to establish a positive correlation as a serious possibility. For this would imply that some influence is leaking from one coin to the other, or that some unknown factor influences in the same way each pair of tosses. Given our initial beliefs, even an appreciable matching will be written off as a chance effect. Imagine a similar scenario with repeated rounds of Prisoner's Dilemma. (In order to avoid the signaling that takes place in the many-round Prisoner's Dilemma, the outcomes are not revealed until the end.) In order to establish the correlation between the random variables C_1 and C_2 (or between D_1 and D_2) many sequences of repeated rounds are needed. It would be extraordinary if the correlation persists through all this. Extraordinary enough to convince some people to choose C. But now we are in Newcomb's land.

6. Newcomb's problem: causality versus agency

Newcomb is a game of a single player against 'nature'. The player has two choices $B1$ and $B2$. The payoff depends on whether a certain event, E, unknown to the player at the time of choice, has taken place.

	E	not-E
$B1$	1,000	0
$B2$	1,001	1

Matrix 2

The player, say Eve, does not have any causal influence on E. From what she knows E and not-E are past events. She is however told that it is very likely that if E took place, she will choose $B1$, and if not-E took place she will choose $B2$. The evidence confirms this. Repeated trials show her that, by far and large, E had happened when she chose $B1$, not-E – when she chose $B2$. Also, when she witnesses other people play – at which times she can observe E or not-E at the time of their happening – the same pattern emerges.

In the original scenario you are offered the choice of either (i) taking a closed box which may or may not contain \$1 000 000 or (ii) taking both the closed box and a second, open one, containing \$1000. You are told that if you will take only one box you will find in it \$1 000 000; but if you will take two, you will find the first empty. There is also an explanation: a superior being had predicted your choice and, according to the prediction, decided what to put in the box.

Whatever its explanatory value, the cover story of the perfect predictor blurs the picture. It introduces an obvious but redundant cycle: A prediction of one's choice is offered as a possible reason for making, or not making, the choice. Yet the paradox does not depend on the perfect-predictor assumption. Eve finds herself swayed merely by the force of accumulated evidence: the on going trials in which Eve participates as player or as witness. (Such 'realistic' scenarios of repeated trials were emphasized by Bar-Hillel and Margalit, [1].) It remains to be seen whether being influenced by this evidence amounts to cyclical reasoning. Eve can follow two courses of action:

(N1) Take one box for the reason: Given the evidence, if I take one box (make $B1$ true), I am likely to find there a very large sum; but if I take two I am likely to find the first empty, and the payoff from the second is comparatively paltry. The reasoning can be cast in terms of expected utilities, where $P(E|B1)$ and $P(\text{not-}E|B2)$ are sufficiently high.

(N2) Take two boxes for the reason: Given the evidence, my doing does not influence in any way what the box already contains. Whatever is there, I do better by choosing $B2$.

The reasoning can be cast in terms of the dominance principle, phrased as a counterfactual: Had I chosen $B2$, I would have gotten more.

One thing Eve cannot do within the context of deliberation is to regard her choosing as caused by events beyond her control. This will lead either to cycles or, in the extreme case, there would be no coherent C-sentence. You cannot follow 'for reasons' by naming causes. If she regards part of her doing as being caused, this part becomes for her an external event.[14] The problem is often posed as one of free will. Actually it is a problem of logic: the logic of agency. I can be told that I have been hypnotized to choose this, that, or whatever I choose. For all I know it might be true. But to entertain the possibility *qua* agent, is for me incoherent.

The following are non-controversial: (i) If Eve knows (or believes) that E, she should choose $B2$. (ii) If Eve knows (or believes) that not-E, she should choose $B2$. It now seems to follow that she should choose $B2$ no matter what. This is another way of expressing the intuition underlying (N2). It is very strong if we imagine Eve's family and friends among the spectators, all of whom know what the first box contains (from their angle the box is transparent) and all of them wish her to choose $B2$. They do not have to say it, Eve knows that they know and she knows how they would advice her. So? So she chooses $B2$ and finds the box empty. Eve's not knowing whether it is E

[14] Oedipus reacted rationally to the oracle. He regarded his doom as an outside force and tried to outdo it. Sincc he had no idea how this force operates, it would have been equally rational for him to continue his ordinary life, taking the usual precautions against incest. But had he continued merely by observing himself marrying his mother, he would have ceased to be an agent.

or not-E functions like a physical constraint inherent in the scenario. It is impossible for third party observers to pass information to her; near as they seem they are on a different planet.

There is another way of arriving at $B2$. Eve might refuse to draw any inductive inference, no matter how overwhelming the evidence. She treats all repeated trials, her's and other's, as irrelevant. Her probability is such that the evidence and the event of finding the box empty are independent under each of the conditional probabilities $P(\ |B1)$, $P(\ |B2)$. The reason for discounting the evidence is no other than lack of causal arrow from choices to E. Again we appeal to causality, but now $B2$ is mandated by an expected utility argument. Recall that $\mathrm{AC}^{\#}$ prohibits probabilistic dependencies when the choice is judged to have no bearing on the event. Indeed if choosing $B1$, or $B2$, is judged to have no bearing on whether she finds the box empty, then the prior probability underlying (N1) violates $\mathrm{AC}^{\#}$, and (N1) is vitiated on grounds of cyclicity. But can Eve say, in face of all the evidence, that her choice has no bearing on what she will find in the box?

Newcomb's scenario brings to the extreme a conflict between our conception of causality, interpreted in counterfactual terms, and our conception of an agent, as one who can – to the extent of having a choice – affect the outcome. [15]

The overwhelming intuition in favor of (N1) is that you can make $B1$ true and that people who do so are getting rich. (That aspect was nicely brought out by Bar-Hillel and Margalit who advocate (N1)). To the question "Why ain'cha rich?" Lewis [9] replied that it was not his fault. In Newcomb's land nature discriminates against people who subscribe to Lewis-type rationality: whenever the choice is offered to them their box is empty. I do not find the reply convincing. Nature can discriminate against high intelligence, or against sanity (games of "chicken" favor those who put less value on their lives). But these are things one can do nothing about. Whereas in Newcomb's case one *can* do something about it: Take, whenever the choice is offered, one box only. (Note that Lewis's kind of reply can be given by anyone who refuses, on a priori grounds, to learn from evidence. Suppose that a coin lands tails in 90 out of 100 tosses. Knowing this, one might still bet with odds 1 : 1 on heads, arguing that this is mandated by the coin's symmetry. Persisting in this, he offers "discrimination by nature" as an explanation for his declining wealth.) This is the irresistible internal logic of agency. On this logic, finding and not finding, the box empty are not given-world events. Therefore E, as well, is not a given-world event, notwithstanding its being in the past.

[15] The possible-world device is simply a way of formalizing the causal intuition; hence it does not contribute any additional motive in favor of (N2). On Lewis's theory X counterfactually implies Y ($X \,\square\!\!\rightarrow Y$) if there is a close enough possible world in which X is true and, in every close enough world, if X is true so is Y. Now suppose that E obtains and Eve chooses $B1$. We have still to decide whether a world in which E obtains and Eve chooses $B2$, is closer than a world in which not-E obtains and Eve chooses $B2$. And the only way to decide this is by invoking causality. For if we assume that $B2$ causes not-E, the second world would presumably be closer.

So in the tug of war between causality and agency, agency wins; even when our action achieves its effect in ways we cannot understand, ways that go against the grain of entrenched thinking. Imagine a world where Newcomb's game is realized. Let the payoffs be \$10 000 and \$10 010. This would allow for many tens of thousands of rounds, without creating nationwide inflation (which would undermine the payoffs). Some people you know get rich in this way. Let me make a prediction about myself and others. I, and many I know, will switch very soon to choosing $B1$. I suspect that David Lewis will switch too. If he does not, he will not be a martyr to rationality but to a metaphysical theory of counterfactuals.

References

[1] M. Bar-Hillel, A. Margalit, Newcomb's paradox revisited, Br. J. Phil. Sci. 23 (1972) 295–304.

[2] H. Gaifman, Paradoxes of infinity and self-applications I, Erkenntniss 20 (1983) 131–155.

[3] H. Gaifman, A theory of higher order probabilities, in: B. Skyrms, W. Harper (Eds.), Causation, Chance, and Credence, Kluwer Academic Publishers, Dordrecht, 1988, pp. 191–219.

[4] H. Gaifman, Pointers to truth, J. Philos. LXXXIX (5) (1992) 223–261.

[5] D. Hofstadter, Mathematical themes, Scientific American (1983).

[6] I. Levi, Hard Choices, Cambridge University Press, Cambridge, 1986.

[7] I. Levi, Consequentialism and sequential choice, in: M. Bachrach, S. Hurley (Eds.), Foundations of Decision Theory, Basil and Blackwell, Oxford, 1991.

[8] D. Lewis, Prisoners' dilemma is a newcomb problem, Philos. Public Affairs 8 (3) (1979).

[9] D. Lewis, Why Ain'cha Rich? Nous 15 (1981) 377–380.

[10] B. Russell, Mathematical logic as based on the theory of types, Amer. J. Math. (1908). Reprinted in: R. Marsh (Ed.), Logic and Knowledge, Unwin Hyman, London, 1989.

[11] G. Shackle, Time in Economics, North-Holland, Amsterdam, 1958.

[12] W. Spohn, Where Luce and Krantz do really generalize savage's decision model, Econometrica 40 (1977) 113–134.

[13] W. Spohn, Grundlagen der Entsscheidungtheorie, Copenhagen, Scriptor, 1978.

[14] B. van Frassen, Presumption, implication, and self-reference, J. Philos. 65 (1968) 136–152.