A (hopefully) comprehensive solution to the Newcomb's problem

Dimitri Lasserre (dimitrilasserre@hotmail.fr)
Centre Gilles-Gaston Granger https://orcid.org/0000-0002-8012-0285

Research Article

Keywords: Newcomb, Rationality, Causal decision theory

Posted Date: May 21st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-226163/v1

Noname manuscript No.
(will be inserted by the editor)

A (hopefully) comprehensive solution to the Newcomb's problem

Received: date / Accepted: date

Abstract The Newcomb's problem has been discussed in the philosophical literature for more than fifty years. So far quarrels mainly oppose causalism on the one hand, and evidentialism on the other hand. This paper wants to explain why causalists are right from a certain point of view, and why they should one-box — instead of keeping two-boxing. The Newcomb's problem can be thought in terms of possible worlds. I assume here that this problem does not happen in our world, but it occurs in a world where causality and rationality obey different rules. But, in any cases, when the predictor is reliable, the player must one-box.

Keywords Newcomb · Rationality · Causal decision theory

1 The problem

The problem we shall deal with here is the Newcomb's problem (NP) as it was originally proposed by Robert Nozick:

> You must choose between taking (and keeping the contents of) (i) an opaque box now facing you or (ii) that same opaque box and a transparent box next to it containing $1000. Yesterday, a being with an excellent track record of predicting human behaviour in this situation made a prediction about your choice. If it predicted that you would take only the opaque box ('one-boxing'), it placed $1M in the opaque box. If it predicted that you would take both ('two-boxing'), it put nothing in the opaque box (1969, p. 207)[1].

Since Nozick formulated this problem, several new versions or reformulations have risen in the literature. The point was often to make this problem

[1] See also Nozick (1993, p. 41) for another formulation of the same problem.

more intelligible, or easier to "solve". Thus NP became, in Jeffrey's paper, the "Fisher smoking case" (1983, p. 15); in the meantime, Kavka made it the "toxin puzzle" (1983); later, in Quattrone and Tversky's work, it was considered as a problem similar to what happens in "voting in large elections" (1986, pp. 58–57); and so on[2].

Also, some works try to compare NP to real life situations. Whether it is possible or not to consider NP as a real-life problem is a question openly asked by Bermúdez (2018, p. 19). In his paper Bermúdez notes that "the most frequently discussed examples of putative real-life NPs are medical in nature" (p. 23), and gives some examples from the literature[3]. Hence on the one hand are theoretical reformulations, and empirical reformulations on the other.

Both theoretical and empirical reformulations will be left out in this paper because they are neither necessary nor relevant to solve the Newcomb's problem. This idea is mainly defended with the following argument: if a solution to NP which does not need use reformulations exists, then reformulations are unnecessary and irrelevant to solve the problem. This solution exists: therefore, reformulations are unnecessary and irrelevant in order to solve it. Of course, one can object that the solution still lacks. The point of this paper is to present it and to prove that the solution is not missing.

Most of the debates about NP oppose two theories. The *Causal Decision Theory* (CDT) on the one hand generally considers that there is no causal link between the choice of the player and the prediction of the predictor, and as a consequence advocates the "two-boxing" strategy. As the final choice has no consequence upon the possible outcome the agent can get, it is rational to two-box. From this point of view, the two-boxing strategy always dominates the one-boxing strategy. Two early versions of this argument give clear justifications[4]. Gibbard and Harper explain that whatever is inside the opaque box (one million dollars or nothing), the payout will always be bigger by taking the two boxes; then the two-boxing strategy dominates the one-boxing one:

> Rational choice in Newcomb's situation, we maintain, depends on a comparison of what would happen if one took both boxes with what would happen if one took only the opaque box. What the agent knows for sure is this: if he took both boxes, he would get a thousand dollars more than he would if he took only the opaque box. That on our view makes it rational for someone who wants as much as he can get to take both boxes, and irrational to take only one box (1978, 155).

It seems clear, in this causalist perspective, that two-boxing is always the best strategy. Then, to the question "Why ain'cha rich?", frequently asked

[2] Some works also tend to make NP a prisoner's dilemma (e.g. Lewis (1979) and, more recently, Walker (2014); Bermúdez (2015)). This parallel, often disputed in the literature, leads to wonder if one can treat NP with the tools of game theory or not. These discussions will not be considered here, and NP will not be considered as a prisoner's dilemma.

[3] See, e.g., Skyrms, 1980; Horgan, 1981; Price; 1986.

[4] There are obviously more than two versions of this argument. The oldest one was provided by Stalnaker (1972). Nevertheless, the two versions proposed here gave birth to quite important discussions in the literature.

by one-boxers to two-boxers who failed in earning one million dollars, Lewis answers that:

> We two-boxers think that whether the million already awaits us or not, we have no choice between taking it and leaving it. We are convinced by counterfactual conditionals: If I took only one box, I would be poorer by a thousand than I will be after taking both [...] We reply that we never were given any choice about whether to have a million. When we made our choices, there were no millions to be had. The reason why we are not rich is that the riches were reserved for the irrational (1981, p. 377).

Lewis claims it is "irrational" to one-box because there is no way the final choice, the final action, *causes* the predictor's guess[5]. Thus only "irrational" one-boxers eventually get rich. Yet this argument has at least one weakness: if what is rational is acting so as to maximizing one's utility, isn't it *empirically* rational to "one-box" when one knows by evidence that most of those who one-boxed before actually got rich? – and that most of the two-boxers were left poor. Yes, this is an inductive reasoning; but the induction is rationally reinforced by the fact that the player knows the predictor almost always makes the right divination.

This is where the *Evidential Decision Theory* (EDT) intervenes, "according to which, to quote Arif Ahmed, the rational act is whichever available one is the best evidence of what you want to happen" (Ahmed, 2018a, p. 8). The "rational" Lewisian two-boxer might be proud of her rationality, but the million dollars will certainly remain away from her. Then it is rational to wonder whether one-boxers are really less rational than two-boxers. Surely it is *theoretically* more rational to two-box, but it is *empirically* more rational to one-box, because one-boxing maximizes the player's utility. Thus if what is rational is maximizing one's utility, the rational means to use in any NP to reach this goal is generally to one-box.

Here, as Andreou noticed, two kinds of rationality must be distinguished: *theoretical* rationality on the one hand, *practical* rationality on the other (2018, p. 44). The aim of this paper is to show that it is not necessary to sacrifice rationality on the altar of practical decisions. Actually, this "practical" rationality, is not only practically, but also totally *theoretically* rational. At least this is what is to be proven in Sect. 2.

---

[5] This argument has found some echo in the literature. More lately, Wedgwood pointed out that:

> EDT endorses the perverse choice to act in such a way that you give yourself good news about which state of nature you are in, even though there is absolutely nothing that you can do to determine which of these states of nature you are in, and even though the alternative course of action is bound to be preferable whichever state of nature you are in. So it seems that without some further refinement, EDT cannot be the correct theory of rational choice (2013, p. 2647).

The same kind of idea, mixed with the argument of dominant strategy, can also be found in Arntzenius (2008, pp. 280-281).

Then Sect. 3 will discuss the notion of "causal independence", often used by two-boxers to justify their choice, and to claim that this choice is rational. In Sect. 4 the problem will be presented in terms of possible worlds. And the solutions will take place in these different possible worlds. Eventually Sect. 5 will open a discussion.

2 What a "rational" move for a NP player is

As many problems of the same sort, NP is a problem of utility maximization. What a NP player wants to do is maximizing her utility and, as in any other problem of this kind, rationality is a mean to achieve this goal. Being rational is not, for the player, an end by itself. The end is to get rich; the means, to be rational. As a consequence the "Why ain'cha rich?" (WAR) argument deserves better consideration. Using WAR is, in the end, asking this question: "if you, two-boxer, claim to be rational, why don't you choose acting in the way in which evidence tends to prove that it will maximize your utility?" and this other question: "how can two-boxers say they are rational if their rationality almost never lead them to maximize their utility?" Indeed it is hard to rationally defend the failure, in NP, of one's rationality while hiding oneself behind this same conception of rationality.

In the Lewisian argument, one-boxers are blamed for being "irrational". But a weakness of the classical two-boxing argument is to consider rationality more as an end than as a mean. When a player faces a NP she does not aim to be *absolutely rational*, she aims to become wealthy. Then she searches for the best decision in order to achieve that goal. The player's expectations are not relative to her rationality, but to her capacity of getting rich – or remaining poor.

The Lewisian argument presupposes that rationality is a norm, that there are *objectively* rational behaviors and *objectively* irrational behaviors. But in utility maximization problems there cannot be such normative rationality. What is rational is to act so as to maximize one's utility, and not to obey to certain behavioral norms which are presupposed to be rational. However, one-boxers, who get *evidentially* rich, act in a way which maximize their utility a lot more than two-boxers do. There is no way therefore that they act *irrationally*.

Nevertheless, the causalist argument seems to resist. A Lewisian voice may still whisper: "well, the way one-boxers found to get rich is by itself irrational." As a matter of fact, the final act of a two-boxer does not make her rich. This act, from EDT's point of view, does not *entail* the predictor to put one million dollars under the opaque box, and as a consequence this act is not the *cause* of the wealthiness of the one-boxer. The fact that the prediction and the final action of the player are *causally independent* is the last stand of Lewisian arguments. Of course, if this clause of *causal independence* were not true, then these arguments would also be wrong. In this case, WAR defenders

would be rational, and Lewisian rationalists would defend a theoretically false representation of rationality[6].

3 Causal (in)dependence

This section deals with the *clause of causal independence*. By *clause of causal independence* one must understand: in NP the action of the player is causally independent from the prediction of the predictor. Again, in other words: by choosing to take either one box or two, the player does not cause the million dollars to be, or not to be, inside the opaque box. This clause is a rarely debated commonplace in the literature[7]. Thus, one might take for granted that it is a necessary *fact* that player's and predictor's actions are causally independent to each other. But one question still needs to be asked: what if the clause of causal independence was, more than an obvious fact, nothing but an unnecessary or a misleading assumption? If it was the case, then it would be necessary to provide a new interpretation of NP, in which causal relationships would be newly explained. What I claim here is that not only the clause of causal independence is more an assumption than a fact, but also this assumption is necessarily false. More precisely, this assumption is *nonsensical in any possible world*.

In the line of the Kantian philosophy, and even more of the Schopenhauerian philosophy, I will consider that causality is, if not a "category of the human understanding"[8], the condition of any human representation. In Schopenhauer's words:

> Where the object begins, the subject ends. That they share a border is shown by the fact that the most essential – and therefore most general – forms of all objects (space, time and causality) can be discovered and fully comprehended starting out from the subject even in the absence of any cognition of the object itself. In Kantian terms, these forms lie in our consciousness *a priori* (2010 [1819], p. 26).

Here, let us simply state that human beings represent themselves the world under causal relationships, and that they cannot imagine a world run by rules likely to violate causality. One does not have to assume that causality is a *real property* of the world, but only that, for any human being, no event can occur outside the law of causality; that is nothing can happen without a set of events, of elements, subsumed under the name of "cause". This condition is necessary because the understandability of any empirical problem depends

[6] For a recent defense of the WAR argument, see Ahmed (2018b). Unlike Ahmed however, for whom his arguments may be "very far from saying that WAR should move any consistent defender of Causal Decision Theory" (2018b, p. 71), I hope to convince the reader of this paper that WAR is one of the possible starting points of a necessary reconsideration of what is heard by "rational" in NP.

[7] Most of the articles, books or book chapters dealing with NP mention the clause of causal independence. In rarer places, this clause is however called into question.

[8] See the classic *Critique of the pure reason* (Kant, 1998 [1787]).

on it. One might object that NP occurs in a world where causality is not effective. Events in this world, as a consequence, are neither understandable nor explainable in causal terms. Therefore these events are impossible to understand to any human being from our world. In this peculiar possible world, where events cannot be causally explained, NP has absolutely no solution – *for* human beings[9]. This is why I will only deal with causally understandable and explainable solutions.

It appears that the only solutions of NP that can exist are based on explanations which rely on causality. Does it mean that we are back to CDT? Not exactly. CDT is based on two hidden or implicit assumptions, and these assumptions must be exposed and discussed. The first assumption – let us call it H_1 – has already been exposed: two-boxers take it for granted that what is *rational* to do is to two-box. The problem here is that H_1 is a normative judgement[10]; this statement presupposes a definition of what a rational behavior is, even though this behavior does not maximize the agent's utility. Then H_1 does not anymore rely on the definition of rationality standardly used by CDT defenders. How, indeed, can one claim that CDT advocates for "dominant strategies", that it "maximizes expected utility", and so on, while all evidences in NP show it simply does not? Then the only way to save "rationality" is to change its definition. H_1 defenders first use the standard definition of rationality: they compute expected utilities, explain what strategy dominates, *i.e.* enables to maximize one's utility. Secondly, when this strategy fails, they use a new definition of rationality, a *normative* definition, which does not depend on outcomes at all. As the outcome does not fit with the latter definition of rationality, CDT considers that NP rewards only "irrational" behaviors. But this conclusion results from a misleading reasoning, based on two different definitions of "rationality".

The second assumption (H_2) is maybe even more important here. It is a deeper assumption, shared in almost all discussions concerning NP, as much for EDT defenders as CDT defenders. H_2 states that NP occurs in our world, and not in other (imaginary) possible world. H_2 makes the clause of causal dependence be very strong. Indeed, if NP occurs in our world[11], it is impossible to understand how and why one-boxers get richer than two-boxers. If there is no way that any move from the NP one-boxer causes the million dollars to be inside the opaque box, then NP is simply impossible to understand, and probably impossible to solve. Of course one could not imagine that the

[9] One might go further and note, with Nagel (1974), that it is impossible to know what it is like to be anything else but a human being. And maybe a being of another kind might be able to represent herself the world without using causality. But these representations still lie incomprehensible to humans. Wittgenstein, on the other hand, offers another argument: talking about the nature of causality, or about any possibility of explaining the world in non-causal terms, is dealing with metaphysical stuff. Then these discussions are fruitless, because they are "nonsensical" (2002 [1921], 4.003, p. 22).

[10] In Putnam's words (2002) H_1 his a 'value judgement". It seems *better* to H_1 defenders to consider that rationality is defined independently from the outcome that results from an action or a behavior.

[11] In a world where a NP predictor does not exist.

one-boxer's final action causes the money to be in the opaque box in our world, neither that this action causes the prediction itself. But what if NP happened in another possible world, where the clause of causal dependence was not effective, or where it obeyed different laws, or where it had a different shape? If it was the case, *i.e.* if H_2 was false, NP would surely be possible to solve. But is H_2 false? Actually, it is.

NP can simply not take place in our world. Or, if it does, it means either that our world is built on some unsuspected structure, or that it is necessary to give up any conception of free-will. In the first case, any usual causal explanation must be suspect; and remain at least partly unsatisfying. In the second, (quasi) totally determinists assumptions must be taken into consideration. I shall add that H_2 carries assumptions not only about the world into which NP takes place, but also about the predictor's faculties, and about the decision process of both predictor and NP player. All these problems will be examined in Sect. 4.

4 The possible worlds where NP can take place: and the respectively corresponding solutions to NP

4.1 A world without causality

First one might consider that NP does not occur in a world where what we know as the "law of causality" exists ; that is where it is impossible to understand the world through causal perspectives. Let us call this assumption A_1.

It is hard to figure out how this world seems because human beings understand things, events, through the filter of causality. A world with no causality is actually a world impossible to understand thanks to causality. This world is not understandable for any human beings. Events in this world seem to randomly happen, and no law can be built to explain and describe their causal relationships – because they have no such relationships in human representations.

In this sort of world, not only NP, but every problem needing explanations, cannot be solved by human brains. If one makes A_1, then one must give up searching for a solution to NP. In a world where A_1 is true, moreover, NP might even be hardly thinkable.

4.2 A world with backward causation

Mackie (1977, p. 223) suggests that NP is ill-formed, because the solution of the problem requires to use either backward causation or inductive reasoning, which is inconsistent with "the requirements that the player should have". As for Mackie, backward causation is "trickery" ; and it might be so in our world. Yet, as soon as we consider NP to belong to another world, the "trick" only

consists in considering that backward causation can happen inside it, or is a property of it, or a property of the knowledge one can have about it, and so on. NP is, as Mackie notices, ill-formed. But it is not ill-formed because it needs some backward causal representations, but because it deals with a world which is not ours, while pretending it deals with our world.

What should one think, then, if NP happened in a world where backward causation was possible, where causation did not work as it does in our world, or was ruled by different laws? In a thought experiment, Dummett imagines a situation in which a present action could seem to cause something to occur, even though this thing appears to be causally determined by past phenomena, independent from any moves that can be done right now. This experiment is about saying "Click!" before opening an enveloppe: when one says "Click!" then the enveloppe never contains a bill (1954, pp. 43–44). Dummett concludes:

> If I observe that saying "Click!" appears to be a sufficient condition for its not being a bill, then my saying "Click!" is good evidence for its not being a bill; and if it is asked what is the point of collecting evidence for what can be found out for certain, with little trouble, the answer is that this evidence is not merely collected but brought about. Nothing can alter the fact that if one were really to have strong grounds for believing in such a regularity as this, and no alternative (causal) explanation for it, then it could not but be rational to believe in it and to make use of it (1954, p. 44).

In a more recent paper, Liu and Price discuss Dummett's ideas about causality and about its inscription in a specific timeline:

> For Causalists it is no use trying to stipulate their way out of the problem, by saying in their description of the Newcomb Problem, "We assume that causation works in the ordinary way." For DEVI[12] will respond: "You're not entitled to assume that causation works in the normal way, in proposing examples that put this kind of pressure on the ordinary notion of causation. My point is that so long as you concede that the one-boxer will get rich, you've conceded that the case is in that contentious territory. At that point, my move is on the table, and you can't avoid it" (2018, p. 174).

When a given experiment always or quasi-always provide the same outcome, even though causality seems to be violated, it is not that awkward to assume that the rules of causation are at least partially unknown. What is awkward, however, is to consider that this experiment occurs in our world, where violation of causation has not been observed in everyday's life experiences. But as soon as one proposes a thought experiment that includes the violation of causation, then it is not weird anymore to consider that the world where the experiment happens allows other laws of causation – that is to say other representations of causation. And as a consequence it is not absurd anymore, in the case of a NP, to one-box in this possible world. Dummett's conclusions

[12] "DEVI" is for Dummett's Evidentialism.

about NP, even though they are not based on the exact same assumptions, also result in one-boxing:

> I thus have a choice between doing something that will, with a very high probability, result in my getting $1,000 and doing something that will, with a very high probability, result in my getting $10,000. Plainly, ... the rational thing for me to do is the second. After I have done it, the rules governing the assertion of counterfactual conditionals may entitle me to assert, "If I had taken both boxes, I should have got $11,000"; but that is only a remark about our use of counterfactual conditionals. Before I make my choice, I should be a fool to disregard the high probability of the statement "If I take both boxes, I shall get only $1,000." That is not merely a remark about our use of the word "probability," nor even about our use of the word "rational," but about what it is rational to do (1986, p. 375).

Dummett relies on an inductive reasoning here: "(almost) all of all the one-boxers have got richer than the two-boxers so far: so let's one-box". Yet it is still possible to transform this inductive process to a hypothetico-deductive one, while supposing that what enables this odd phenomenum to occur is an unusual structure of causation; what one might call "backward causation". If backward causation is the rule – in some special, defined and known cases – then there are no more oddities in NP. In this case, one-boxing is not only *practically* wise, but it is *theoretically* rational.

Backward causation, even though it violates the standards of causality, can be considered as a special kind of causality, with its own rules. One can imagine a world where present actions influence, because of the structure of this world or because of the way one tries to explain and understand it, past actions. Of course, backward causation is not a relevant way to explain phenomena when it is run itself by randomness. But in a world where the circumstances which make backward causation possible are clearly identified, there is no reason that this special form of causality could not explain empirical phenomena, and could not be used to predict some of them. In this world one-boxing is the solution of NP.

4.3 A world with mere causality

The solution of NP will be exposed in this section. Not that what was presented in 4.1. and 4.2. were not solutions to NP, but the very specific cases of a non-causal world and of a world were backward causation is possible are not relevant in a world where causality works as it does in our world.

4.3.1 Only mere causality

One thing remains to be told about backward causation: one can still imagine that NP may occur in a world like ours, and may reveal at the same time

anomalies concerning the causal structure of our representations. In this case, as Horgan recommends, it might be relevant to act "*as if* one's present choice could causally influence the being's prior prediction, but my argument does not presuppose backward causation" (1981, pp. 340-341). Here one actually need not believe in *true* backward causation, but only need take into consideration the NP's oddities before acting – which would lead one to act "as if" the world was sometimes ruled by backward causation. But saying that one must act "as if" backward causation was possible while considering it is not *is not* a solution to NP: the reason why the player would have had better one-box is not rationally established.

What is important here, more than believing in backward causation or other phenomena of this kind, is to seriously consider, with Viger et al., that the content of the opaque box is not independent of the player's choice (2019, p. 420). At least this assumption of *causal dependence* must be investigated. NP is before all a problem of causality. And the solution to NP depends on how causation mechanisms are considered. This idea is, with good reason, defended by McKay:

> The right way to approach the Newcomb problem is to attempt to work out the underlying causal structure, just as the causalists prescribe. You must decide whether or not there is a concealed causal connection. And for either possible answer on the underlying causal structure, the causalist prescribes the right choice. If you still think there must be no causal connection since the action of the predictor really is in the past, you should two-box. Alternatively, if you think there probably is some cheating going on undetected by you, then you think there probably is a causal connection, and you should one-box. That is why there is no set answer to the Newcomb problem. There are two possible answers and always will be, because the right choice depends on extra information a of the predictor not given in standard descriptions of the case (2004, pp. 188–189).

McKay assumes that causalists are always right, and this assumption is right from a certain point of view. As soon as one assumes that NP occurs in a world where human beings cannot represent themselves the world outside causality, a solution to NP has to be explained in causal terms. If not, this solution would remain impossible to understand and even to explain. If causalists are right, it is not because they have good intuitions, or because they chose the good assumptions. No, if they are right, it is because *they have no other choice than being right*. Why do they have no choice? Because in a possible world where causality has the same structure as in ours human beings can simply not think empirical phenomena outside causality. As a consequence anyone who wants to solve NP needs not only assume causality but also has to take causality as a *necessary* assumption. And not only this assumption is necessary to solve the problem, it is necessary in order to *think* and *present* it.

Causation mechanisms can be understood and examined in only one way: this way is constrained by causality itself, that is what causality is, and what

it can be. Causality presupposes a representation of the world anchored in time. It also presupposes that explanations about empirical phenomena are written in a timeline, follow a chronology. Causes precede effects. What is caused by something is necessarily caused by a past event[13]. Therefore if a NP player believes she must one-box, she has to *causally* justify her choice. This means she must not consider that NP is questionnable outside causality, or can be solved by invoking backward causation. Thus the question is: in a world where representations are ruled by mere causality, is there any good reason to one-box?

4.3.2 A common cause

A common and very relevant idea in the literature suggests that a good reason to one-box in a causal world is the existence of a *common cause* upstream of the predictor's guess and the action of the player. This idea was first proposed by Reichenbach (1956). Stern calls it the "principle of common cause (PCC)":

> PCC: If variables F and G are correlated, then either F (directly or indirectly) causes G, G (directly or indirectly) causes F, or F and G are (direct or indirect) effects of a common cause (Stern, 2018, p. 204).

Yet PCC must be discussed. It is well known that a correlation is not sufficient to establish a causal link; and it is also well known that establishing a causal link is not an easy task. But in NP the correlation between the prediction and the player's move can either be explained by chance or by a common cause. Here the common cause really looks like the best explanation. If the predictor always made right guesses by chance this would be a "miracle"[14]. And even though a miraculous event of this sort might happen, if one can solve NP by using explainable tools and assumptions, then it seems better to leave behind the "miraculous" explanation. Thus I do not consider PCC because it is attractive, or because it might be a fruitful assumption, but because it is the most rational assumption.

The idea of a common cause was discussed by Eells (1982, chap. 6–8), and also by Spohn, for whom "It may be unclear what the common cause is, but it must exist" (2012, p. 101). Then Spohn, as Stern (2018) does besides, presents graphs that illustrate the structure of causal mechanisms including a common cause in NP.

On the other hand Viger et al. give a description of the causal structure in NP that justify both the necessity of a common cause and the one-boxing strategy:

[13] A teleological philosopher might object there are final causes, aiming future events, which determine present actions. But obviously these causes are empirically determined by present psychological and some other contingent factors, and these factors cause future actions.

[14] This abductive reasoning is quite similar to Putnam's "miracle argument" (see Putnam, 1975, p. 73; Tiercelin, 2011, pp. 232–233).

> We argue that psychological coherence requires certain backtracking counterfactuals to be true in the scenario stipulated in the Newcomb Problem, which entail that choosing one box is the rational decision. In effect, a person's psychological make-up is a common cause of what she will choose and what is in the opaque box, mediated via the predictor. Treating decisions as isolated, independent events can, in certain circumstances such as the reflexive context of the Newcomb Problem, lead to paradoxes about what is rational (2019, p. 408).

This argument shows the limit of the traditional approaches of NP, of the usual struggle between EDT and CDT. Neither EDT nor CDT ask the clause of causal independence, which is necessarily false. No one can actually explain the success of the predictor if this clause is true; no one can either explain why one-boxers are more successful than two-boxers if the prediction and the action of the player are causally independent – except if one add extra hypotheses about the circumstances of the experience, or the possible world where NP stands. This argument is strong, also – and so is any argument considering a common cause –, because it reconciles one-boxing with causalism.

4.3.3 What the common cause is not

Here the problem seems almost solved. As a matter of fact, it has not been totally solved yet. The questions a NP player asks are: (1) "How can I get rich?" and (2) "Is there a way to be sure to win the million dollars?" The answer to (1) is simple. Based on past observations one can easily say that one-boxing should warrant wealthiness. But (2) is a more challenging question, because evidentialism is causally irrational. From this point of view EDT does not appear to be the solution of NP. EDT actually cannot solve NP because it does not explain causal oddities; it does not give a deductive reason why the player should one-box. One-boxing, as for EDT, is only justified by induction; but the inductive reasoning that motivates one-boxing does not explain why it is right to one-box. EDT is based on an intuition: the intuition that one-boxing "works". But no evidentialist *knows why* this strategy works.

Here it appears that Lewis is somehow right when he says that the money goes, in NP, to irrational people. But he is not right because it is irrational to one-box. Lewis is right because one-boxers do not one-box for the good reasons. As long as one consider that NP happens in a world of mere causality, it is absolutely rational to one-box – and respectively irrational to two-box –, but it is relatively irrational to one-box, when one one-boxes for bad reasons (like intuition, belief, inductive reasoning, and so on).

Obviously the good reason to one-box consists in considering that there is a common cause, that there is something that both causes the prediction and the action of the player. Generally the literature considers this common cause is the *intention* of the player. But, for some reasons, this assumption is insufficient and some problems remain. These problems lead to this argument:

1. Intention might change;

2. If people are free, then their intention is inaccessible to the predictor before NP experiment;
3. If intention might change and if people are free, then there is no way the predictor is omniscient or quasi-omniscient;

∴ As a consequence intention *is not* the common cause by itself. Or the common cause cannot be reduced to intention.

This point is important because it justifies why the world where NP happens is not our world, but a possible world, with mere causality, where a predictor can exist. In this world the notion of freewill is dubious; and so is the notion of intention.

4.3.4 What the common cause is in a world of mere causality

How can the predictor do such accurate predictions? What necessary conditions are needed to enable these predictions to exist? First the predictor has to know, before the player opens only one or the two boxes, whether she will one or two-box. This means that when she acts, the player has (almost) absolutely no choice. Or, if she chooses *when* she one-boxes or two-boxes, the player's choice's determinations are already settled and known by the predictor. In this case the "choice" is simply the consequence of determined causes from which the player cannot escape. Thus the world where NP occurs does not allowed free will, especially if the predictor is omniscient[15]. The NP predictor and free will are two mutually exclusive hypotheses. If one assumes free will, one cannot assume an omniscient (or quasi-omniscient) predictor, and *vice versa*.

Secondly the predictor knows all the determinations that will lead the player to make her decision. These determinations have no specific boundaries, but the predictor needs access them to make her prediction. These determinations remain unknown to the player – who need not know them anyway –; yet they determine her final action and, or, decision. The predictor, on the other hand, knows them; and this is why he manages to make always right predictions.

Eventually, even though this might seem anecdotic, in NP mere causality world, the predictor has to have an access to the relevant information. Here one can assume this access is made possible by telepathy or by any technology. This point is not that important because it is obvious that the predictor needs a way to know what she needs to know in order to make right predictions. Hence proposing NP presupposes making ontological assumptions about the "real" powers of the predictor and the structure of the world where NP occurs, assmuptions without whom NP cannot make any sense.

Therefore the *common cause* is both the determinations that lead the NP player to act in a certain way, and the determinist structure of the world where NP occurs. Actually determinism is the first condition, because it allows

[15] I will not discuss here the very hard question of free will. What one must notice, however, is that free will is simply impossible in a world where NP is possible.

absolute determinations to be known. This common cause is much wider than the simple idea of "intention". Intention, moreover, does not mean much in a determinist world; it can only slightly exist in people's mind. What causes the predictor's omniscience is the very specific determinist structure of the NP world of mere causality, and the very special power of a predictor who can access information about people's decisions and actions.

4.3.5 Who wins at the Newcomb's problem game?

The answer might look simple: one-boxers win. But why do they win? They actually do not win *because* they do what constitute their best evidence that they will realize their aim. They win because the possible world where NP happens is a determinist world with a very specific causal structure. Thus one-boxers, in this world, are causalists. And, in this world, CDT recommends one-boxing.

This means that CDT *cannot* recommend two-boxing. What is wrong with causalists, though, is that they do not wonder how the predictor can have such unusual and extraordinary powers. Then they do not investigate the conditions that make these powers possible. From this issue results the clause of causal independence, which is necessarily wrong in a world where NP is possible. This clause gives birth to interpretations that rely on a misleading understanding of the causal structure of the world where NP is possible. For example, the clause of causal independence leads Picavet to claim:

> Causality from present to past does not exist in a our world; neither does it in the situation Newcomb imagined. The problem simply creates the illusion that the player's decision can influence the prediction. But only the predictor's beliefs determine her action. Future events, however, have no influence; because they precisely belong to the future. Recommending one-bowing is relying on a very weird notion of causality, in which real influence of an event on another event does not really matter (1996, p. 245)[16].

Picavet is right when he says that backward causality is impossible (as far as our knowledge goes) in our world. Then he should draw this consequence: NP is therefore impossible in our world. If causalists consider that the prediction only depends on the predictor's beliefs, then they should wonder how her beliefs are always, or almost always, right; in other words, *what causes her beliefs to be always right*. But they do not ask this question, and this is why they recommend two-boxing; even though they *should not*. Yet, as the beliefs of the predictor have to be caused by something, in a world where causality travels from the past to the present, and to the future, CDT should always lead, in a NP, to two-boxing.

Yet the one-boxing causalist argument requires a little more explanations. First of all, the solution is not exactly the same depending on whether the

[16] My translation from French.

predictor is omniscient or not. Then there are two different solutions to NP (which are actually almost the same, but their justification is a little different), depending on the predictor's omniscience.

Solution 1: The predictor is omniscient. If the predictor is omniscient, then she has a probability $p_O = 1$ to make right guesses. In this case the probability of making a right guess is always 100% because the predictor knows what the player will do whatever the player chooses to do. Therefore the player has to one-box if she wants to get rich.

This argument can be very easily transcribed in mathematics terms. For the one-boxing strategy, where EU is the expected utility of the player and p_O is the probability of an omniscient predictor to make a right guess, we have:

$$\begin{aligned} EU_{1box} &= p_O \cdot 1000000 + (1 - p_O) \cdot 1000 \\ &= 1 \times 1000000 + 0 \times 1000 \\ &= 1000000 \end{aligned}$$

Then, if the player two-boxes:

$$\begin{aligned} EU_{2box} &= p_O \cdot 1000000 + (1 - p_O) \cdot 1000 \\ &= 0 \times 1000000 + 1 \times 1000 \\ &= 1000 \end{aligned}$$

Thus:

$$EU_{1box} > EU_{2box}$$

Therefore it is rational for the player to one-box.

Solution 2: The predictor is not omniscient. The standard and well-known solution uses expected utilities so as to decide which strategy maximises EU. The strategy depends on the probabilities of success of the predictor p_1 and p_2 which respectively correspond to the one-boxing and two-boxing strategies. Here expected utilities will be left aside because they do not fit with NP as it has been defined in this paper. NP deals with a deterministic situation. If the predictor gets wrong more when players one-box than when they two-box or vice-versa, then in both cases she is not fully reliable. There is actually no reason to believe that the source of the predictor's errors depends necessarily on the player's strategy. It depends on contingences likely to create lack of information in the predictor's mind.

What is important for the player is to earn the million dollars, not to maximise her utility. She wants the opaque box to be full of cash. Thus the additional thousand dollars do not really matter for the player. If she wants to get one thousand dollars, she can choose two-boxing and reach her aim. But,

of course, a NP player does not expect to get only one thousand dollars; she expects to become rich.

If the predictor is not omniscient, it means she does not have all the information to make the right guess everytime. Sometimes she ignores the player's determinations and either mistakes or lets chance decide. Intuitively it is very easy to consider that if the predictor makes only 50% right guesses, or less, then it means that she never does better than chance. In this case the predictor does not prove to have special powers; she does not prove to actually be a predictor — she might pretend, but she is not. As a consequence, the best strategy, in this situation, *always is* two-boxing.

This result can be obtained easily with a very simple mathematic demonstration. With p_{NO} the probability of success of a non omniscient predictor and $p(S_{Trust} = G)$ and $p(S_{NoTrust} = G)$ the probability to win the million dollars when the player's strategy relies on trusting in the predictor, and respectively on not trusting in her, we have :

$$\begin{aligned} p(S_{Trust} = G) &> p(S_{NoTrust} = G) \\ p_{NO} &> 1 - p_{NO} \\ p_{NO} &> 1/2 \end{aligned}$$

From a deterministic point of view trusting in the predictor means one-boxing, because there is no good reason to two-box if the predictor guesses right. Therefore the player must one-box if the predictor guesses right more than 50% of the time. As long as the predictor beats chance, the player must trust her and one-box.

5 Discussion

In recent developments, Ahmed (2019) compared NP to Frankfurt cases[17]. This paper constitues an answer to a usual objection addressed to the WAR argument, that is to say the "opportunity defense" (Ahmed, 2019, p. 3392)[18]. Ahmed's paper is a tentative to reformulate NP in order to unravel the causality relations which are relevant in a NP situation. Causation is the main problem in NP, and what I hopefully showed is that causal mechanisms in NP can be understood with no need to imagine new situations, different but comparable to the original NP situation.

The toxin problem (Kavka, 1983), which consists in a reformulation of NP, is also solved by the solution proposed here. If the predictor is omniscient, then something *causes* the fact that the player drinks the toxin; what leads her to drink leads the predictor to place the million dollars on the player's

[17] Ahmed holds these cases from works made by Frankfurt. They are based on what Frankfurt calls the "Principle of alternate possibilities (PAP)" which is: "A person is responsible for doing something only if he could have done otherwise" (1969, p. 829).

[18] Here Ahmed precisely confronts Joyce's (1999) and Wells' (2017) arguments.

bank account. The player does not drink the toxin *in order to* get the million dollars. She drinks it because she trusts in the predictor's powers. The same thing happens in a traditional NP: the player does not one-box in order to cause the million dollars to be under the opaque box, but only because she trusts the predictor. This confidence is the "common cause"[19].

The solution proposed in this paper might be comprehensive because it presents the Newcomb's problem as it must be presented, that is to say in terms of possible worlds. NP does not occur in the exact same world as ours. It does not because a person such as the NP predictor does not exist in our world. Then the question is: what conditions make this predictor and her powers possible? NP can be solved only after answering this question. And hopefully this is what has been done here.

One of the main point is that one has to wonder what "being rational" means in the very special case of NP, and what it actually means in general. Many quarrels that oppose EDT and CDT are founded on disagreements about what rationality is, about how it must be defined. But opposing theoretical and empirical rationality is misleading. *What is* rational depends in NP on the player's expectations. If the player expects to get rich she must do what is likely to lead her to get rich. And what is likely to lead the player to become rich belongs to the empirical field. Yet this does not mean that this rationality cannot be theorized. Thus rationality must be understood from both an empirical and a theoretical point of view. Rationality is not an absolute norm. It must not resist to reality. Rationality is a helpful notion so as to understand our representations of reality. This general proposition is true in any special case. And the Newcomb's problem is one of them.

However two objections might rise. First one might object that the assumptions concerning determinism and free-will are far too radical. Actually, even though this would not change the results obtained here, they might. If NP happens in a world of mere causality where the predictor is omniscient, it seems necessary to consider that determinism is a real property of this world; or at least a property of our understanding in this world. But in cases where the predictor is not omniscient, the possible world where NP occurs might not have to be completely deterministic. The predictor must only access some information about the player's decision: this obviously says nothing about determinism and free-will.

The second objection is more empirical. When Nozick published his paper, \$1000 were surely worth more than they do today. It would be certainly interesting to propose experiments where subjects would face a situation in which the transparent box does not contain \$1000, but maybe \$10,000 or even more. Here again, *rationality* and *rational behaviors* might deviate from the trajectory that has been drawn in this paper.

[19] Obviously the cause of this trust is certainly the desire, for the player, to become rich.

References

Ahmed, A. (2018a). Introduction. In A. Ahmed (Ed.), *Newcomb's Problem* (pp. 1–18). Cambridge: Cambridge University Press.

Ahmed, A. (2018b). The "Why ain'cha rich?" argument. In A. Ahmed (Ed.), *Newcomb's Problem* (pp. 55–72). Cambridge: Cambridge University Press.

Ahmed, A. (2019). Frankfurt cases and the Newcomb Problem. *Philosophical Studies*, *177*, 3391–3408.

Andreou, C. (2018). Newcomb's problem, rationality and restraint. In A. Ahmed (Ed.), *Newcomb's Problem* (pp. 42–54). Cambridge: Cambridge University Press.

Arntzenius, F. (2008). No regrets, or: Edith Piaf revamps decision theory. *Erkenntnis*, *68*, 277–297.

Bermúdez, J. L. (2015). Strategic vs. parametric choice in Newcomb's Problem and the Prisoner's Dilemma: reply to Walker. *Philosophia*, *43*, 787–794.

Bermúdez, J. L. (2018). Does Newcomb's problems actually exist? In A. Ahmed (Ed.), *Newcomb's Problem* (pp. 19–41). Cambridge: Cambridge University Press.

Dummett, M. A. E. (1954). Can an effect precede its cause? *Proceedings of the Aristotelian Society, Supplementary Volume*, *28*, 27–44.

Dummett, M. A. E. (1986). Causal loops. In A. Flood, M. Lockwood (Ed.), *The Nature of Time*. Reprinted in Dummett, M. A. E., *The Seas of Language* (pp. 349–375). Oxford: Oxford University Press, 1993.

Frankfurt, H. (1969). Alternate possibilities and moral responsibility. *Journal of Philosophy*, *66*, 29–839.

Gibbard, A., & Harper, W. (1978). Counterfactuals and two kinds of expected utility. In A. Hooker, J. J. Leach, E. F. McClennen (Ed.), *Foundations and applications of decision theory* (pp. 125–162). Dordrecht: D. Reidel.

Horgan, T. (1981). Counterfactuals and Newcomb's problem. *Journal of Philosophy*, *78*(6), 331–356.

Jeffrey, R. C. (1983). *The logic of decision*. 2nd ed. Chicago: University of Chicago Press.

Joyce, J. M. (1999). *Foundations of causal decision theory*. Cambridge: Cambridge University Press.

Kant, I. (1998) [1787]. *Critique of the pure reason*. P. Guyer, A. W. Wood (Ed.). Cambridge: Cambridge University Press.

Kavka, G. (1983). The toxin puzzle. *Analysis*, *44*(1), 33–36.

Lewis, D. K. (1979). Prisoners' dilemma is a Newcomb problem. *Philosophy and Public Affairs*, *8*(3), 235–240.

Lewis, D. K. (1981). Why ain'cha rich? *Noûs*, *64*(2), 377–380.

Mackie, J. L. (1977). Newcomb's Paradox and the direction of causation. *Cana-*

dian Journal of Philosophy, *7*, 213–25.

McKay, P. (2004). Newcomb's problem: the causalists get rich. *Analysis*, *64*(2), 187–189.

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, *83*(4), 435–450.

Nozick, R. (1969). Newcomb's problem and two principles of choice. In N. Rescher (Ed.), *Essays in Honor of Carl G. Hempel* (pp. 144–146). Dordrecht: D. Reidel.

Nozick, R. (1993). *The Nature of Rationality*. Princeton: Princeton University Press.

Picavet, E. (1996). *Choix rationnel et vie publique. Pensée formelle et raison pratique*. Paris: Presses Universitaires de France.

Price, H. (1986). Against causal decision theory. *Synthese*, *67*, 195–212.

Price, H., & Liu, Y. (2018). "Click!" bait for causalists. In A. Ahmed (Ed.), *Newcomb's Problem* (pp. 160–179). Cambridge: Cambridge University Press.

Putnam, H. (1975). *Mathematics, matter and method*. Cambridge: Cambridge University Press.

Putnam, H. (2002). *The collapse of the fact/value dichotomy, And other essays*. Cambridge: Harvard University Press.

Quattrone, G. A., & Tversky, A. (1986). Self-deception and the voter's illusion. In J. Elster (Ed.), *The Multiple Self* (pp. 35–58). Cambridge: Cambridge University Press.

Reichenbach, H. (1956). *The Direction of Time*. Berkeley: University of California Press.

Schopenhauer, A. (2010) [1819]. *The world as Will and representation*, Vol. 1. J. Norman, A. Welchman, C. Janaway (Ed.). Cambridge: Cambridge University Press.

Skyrms, B. (1980). *Causal Necessity: A Pragmatic Investigation of the Necessity of Laws*. New Haven: Yale University Press.

Spohn, W. (2012). Reversing 30 years of discussion: why causal decision theorists should one-box. *Synthese*, *187*, 95–122.

Stalnaker, R. C. (1972). Letter to David Lewis. Reprinted in W. L. Harper, R. Stalnaker and G. Pearce (Ed.), *Ifs: Conditionals, Beliefs, Chance and Time* (pp. 151–152). Dordrecht: Springer, 1981.

Stern, R. (2018). Diagnosing Newcomb's problem with causal graphs. In A. Ahmed (Ed.), *Newcomb's Problem* (pp. 201–220). Cambridge: Cambridge University Press.

Tiercelin, C. (2011). *Le ciment des choses*. Paris: Ithaque.

Viger, C., Hoefer, C., & Viger, D. (2019). The philosopher's paradox: How to make a coherent decision in the Newcomb Problem. *Theoria*, *34*(3), 407–421.

Walker, M. T. (2014). The real reason why the Prisoner's dilemma is not a Newcomb problem. *Philosophia*, *42*, 841–859.

Wedgwood, R. (2013). Gandalf's solution to the Newcomb problem. *Synthese*, *190*,

2643–2675.

Wells, I. (2017). Equal opportunity and Newcomb's problem. *Mind*, *128*, pp. 429–457.

Wittgenstein, L. (2002) [1921]. *Tractatus logico-philosophicus*. D. F. Pears, B. F. McGuinness (trans.). 2nd ed. Routledge Classics.