Newcomb's Problem Revisited

By Terry Horgan

Newcomb's Problem, invented by the theoretical physicist William Newcomb, was first discussed in print by the late Harvard philosopher Robert Nozick, who formulated the problem as follows:

> Suppose a being in whose power to predict your choices you have enormous confidence. (One might tell a science-fiction story about a being from another planet, with an advanced technology and science, who you know to be friendly, and so on.) You know that this being has often correctly predicted your choices in the past (and has never, so far as you know, made an incorrect prediction about your choices), and furthermore you know that this being has often correctly predicted the choices of other people, many of whom are similar to you, in the particular situation to be described below. One might tell a longer story, but all this leads you to believe that almost certainly this being's prediction about your choice in the situation to be discussed will be correct.
>
> There are two boxes, (B1) and (B2). (B1) contains $1,000. (B2) contains either $1,000,000 ($M) or nothing…. You have a choice between two actions:
>
> (1) taking what is in both boxes
>
> (2) taking only what is in the second box.
>
> Furthermore, and you know this, the being knows that you know this, and so on:
>
> (I) If the being predicts you will take what is in both boxes, he does not put the $M in the second box.
>
> (II) If the being predicts you will take only what is in the second box, he does put the $M in the second box.
>
> The situation is as follows. First, the being makes its prediction. Then it puts the $M in the second box, or does not, depending upon what it has predicted. Then you make your choice. What do you do? (Nozick 1969, pp. 114–115)

Nozick pointed out that two widely accepted normative principles in decision theory make conflicting recommendations about the problem, thereby coming into conflict with one another. The traditional principle of *expected-utility maximization* recommends taking only the second box, because it is so likely that the being has correctly predicted your choice. But the *dominance principle* recommends taking

Terry Horgan is a professor of philosophy at the University of Arizona. He specializes in metaphysics, philosophy of mind, philosophy of psychology, epistemology, philosophy of language, and metaethics, and also is interested in philosophy of science, logic, and decision theory. He has published over 150 articles on these topics, has authored three books, and has been recognized extensively for his research.

both boxes, since (a) the current state of box 2 is causally independent of your choice and (b) you inevitably do better by $1,000 by taking both boxes. Nozick showed great respect for this problem. He also seemed to be inclined towards taking both boxes in the official version of the problem, but taking only the second box in the "limit case" version in which it is stipulated that the being infallibly predicts your choice.

Meanwhile, although there are many one-boxers among philosophers at large (and many two-boxers too), philosophers who work on the foundations of decision theory have been almost unanimous in advocating the two-box solution. Also, in the wake of Newcomb's problem there have been various proposals for re-defining the decision-theoretic notion of expected utility in such a way that the two-box choice turns out to have higher expected utility, under the revised definition, than the one-box choice. Traditional decision theory, the kind Nozick was addressing in his paper, has come to be called *evidential* decision theory; and the various versions of two-box-recommending decision theory are subsumed under the rubric of *causal* decision theory.

One version of traditional decision theory, in the form advocated by Jeffrey (1965), goes as follows. Let an agent know (i) that states $S_1,\ldots,S_n$ are mutually exclusive and jointly exhaustive possible states of nature, (ii) that acts $A_1,\ldots A_m$ are mutually exclusive and jointly available acts, and (iii) for each act A_i and state S_j, $O_{i,j}$ is the known outcome that would result from performing act A_i if state S_j obtained. For each such outcome, let $u(O_{i,j})$ be the *utility* of $O_{i,j}$ for the agent, as measured on an interval scale. Jeffrey defines expected utility this way:

$$V(A_i) = \Sigma_j \, prob(S_j \mid A_i) \cdot u(O_{i,j})$$

I.e., $V(A_i)$ is the weighted sum of the respective utilities of act A_i under the respective states $S_1,\ldots S_n$, with the utility of each outcome $O_{i,j}$ being weighted by the conditional probability of state S_j given act A_i.

There are various versions of causal decision theory. A version proposed by Gibbard and Harper (1978) defines expected utility not as $V(A_i)$ but instead this way:

$$U(A_i) = \Sigma_j \, [prob(A_i \mathbin{\square\!\!\rightarrow} S_j) \cdot u(O_{i,j})]$$

This is a different weighted sum of the respective utilities of act A_i under the respective states $S_1,\ldots S_n$; the utility of each outcome $O_{i,j}$ is now weighted not by the conditional probability of S_j given A_i, but instead by the probability of the counterfactual conditional $(A_i \mathbin{\square\!\!\rightarrow} S_j)$. As applied to Newcomb's problem, the idea is this: because each possible state of box 2 is causally independent of the agent's chosen act, $prob(A_i \mathbin{\square\!\!\rightarrow} S_j) = prob(S_j)$ for each available act and each possible state of box 2. Under this condition, two-boxing is guaranteed to have higher expected utility than one-boxing.

In Horgan (1981) I argued in favor of one-boxism. After formulating both two-box reasoning and one-box reasoning using counterfactual conditionals, I pointed out that whereas the two-box argument presupposes what David Lewis (1979) called the "standard resolution" of the vagueness of counterfactuals, the

one-box argument instead presupposes what Lewis (1979) called the "backtracking resolution" of the vagueness of counterfactuals. I then argued that the backtracking resolution is appropriate for practical decision-making in Newcomb's problem, and that the standard resolution is not—and hence that practical rationality requires one-boxing.

I subsequently came to believe, however, that my argument in Horgan (1981) was essentially just a somewhat embellished version of the very one-boxer reasoning that two-boxers repudiate. I reluctantly came to agree with David Lewis (1981) that the debate between one-boxers and two-boxers is a hopeless stalemate—a view I embraced in Horgan (1985). I was still a one-boxer, though, and Lewis was a two-boxer—even though we were in agreement that the debate is a stalemate. Today I am still a one-boxer and I still regard the debate as a stalemate, but I do have some new things to say about the problem.

I. A new one-box argument

Here I will propose new reconstructions of what I take to be the basic intuitive reasoning that leads to the one-box conclusion in Newcomb's problem—first a formulation for the limit case in which it is stipulated that the chooser knows *for sure* that the predictor has correctly predicted what the chooser will do, and then a formulation for the official version in which it is stipulated that the chooser knows it to be *extremely probable* that the predictor has predicted the chooser's action correctly.

Let *P* be a standard matrix specification *P* of a decision problem: *P* specifies that (a) acts $A_1,\ldots,A_m$ are those open to the agent, and the agent knows this; (b) states $S_1,\ldots,S_n$ are mutually exclusive and jointly exhaustive possible states of the world, and the agent knows this; and (c) for each act A_i and state S_j, the agent knows that if she performed A_i and S_j obtained, then the outcome would be O_{ij}. I now introduce some notions that will prove useful below. Let the *complete act/outcome scenario-partition* C(*P*), for the decision problem *P*, be the unique set of scenarios comprising all and only the (mutually exclusive) scenarios of the form $(A_i \,\&\, S_j \,\&\, O_{ij})$ that arise from *P*. And let a *canonically selectional scenario-partition* (for short, a CS scenario-partition), for the decision problem *P*, be a set *S* such that (i) *S* is a subset of C(*P*), and (ii) for each act A_i in *P*, *S* contains exactly one A_i-involving scenario from C(*P*).

In addition, let *act-independent knowledge* (AIC knowledge), for a given decision problem *P*, be knowledge that is possessed by the chooser in *P* in a way that does not depend on any evidence that the chooser in *P* might possess concerning which act the chooser will perform.

Consider now the limit-case version of Newcomb's problem: by stipulation, the chooser knows that the predictor has actually predicted what the chooser will do. My proposed formulation of the reasoning in favor of choosing only box 2 is this:

Limit-case one-box argument:

L1. I have act-independent knowledge that I will act in the manner predicted.

L2. If I have act-independent knowledge that I will act in the manner predicted, then the only CS scenario-partition each of whose members is consistent

with my act-independent knowledge is the partition comprising the following two scenarios: (i) I choose only box 2 and obtain $1 million, and (ii) I choose both boxes and obtain $1,000.

Hence,

L3. The only CS scenario-partition each of whose members is consistent with my act-independent knowledge is the partition comprising the following two scenarios: (i) I choose only box 2 and obtain $1 million, and (ii) I choose both boxes and obtain $1,000.

L4. I prefer scenario (i) to scenario (ii).

L5. If there exists a CS scenario-partition *S*, and an act A_i open to me, such that (a) *S* is the only CS scenario-partition each of whose component scenarios is consistent with my act-independent knowledge, and (b) I prefer the A_i-involving scenario in *S* to every other scenario in *S*, then practical rationality requires me to perform A_i.

Hence,

L6. Practical rationality requires me to choose only box 2.

Consider next the standard version of Newcomb's problem: by stipulation, the chooser knows that it is extremely probable that the predictor has actually predicted what the chooser will do. My proposed formulation of one-box reasoning for this generalized version is this:

Generalized one-box argument:

G1. I have act-independent knowledge that it is extremely probable that I will act in the manner predicted.

G2. If I have act-independent knowledge that it is extremely probable that I will act in the manner predicted, then the only CS scenario-partition each of whose members is consistent with what I currently act-independently know to be extremely probable is the partition comprising the following two scenarios: (i) I choose only box 2 and obtain $1 million, and (ii) I choose both boxes and obtain $1,000.

Hence,

G3. The only CS scenario-partition each of whose members is consistent with what I act-independently know to be extremely probable is the partition comprising the following two scenarios: (i) I choose only box 2 and obtain $1 million, and (ii) I choose both boxes and obtain $1,000.

G4. I strongly prefer scenario (i) to scenario (ii).

G5. If there exist a CS scenario-partition *S*, and an act A_i open to me, such that (a) *S* is the only CS scenario-partition each of whose component scenarios is consistent with what I act-independently know to be extremely probable, and (b) I strongly prefer the A_i-involving scenario in *S* to every other scenario in *S*, then practical rationality requires me to perform A_i.

Hence,

6. Practical rationality requires me to choose only box 2.

As I said at the outset, these formulations of one-box reasoning now seem to me to do well at reconstructing the fundamental line of thought that

underlies the pre-theoretic intuition that practical rationality requires taking only box 2. This seems so to me even though it has required some careful and deliberate reflection on my part to craft these formulations, and to articulate the key notions they employ—viz., the notion of a CS scenario-partition, and the notion of act-independent knowledge. Intuitive appreciation of the fundamental rationale for one-boxing, like other kinds of intuitive judgment, may well rest in part upon considerations that one need not be readily able to articulate explicitly. (The same goes, for instance, for intuitive appreciation of the applicability or non-applicability of a given general concept—e.g., the concept of knowledge or the concept of water—to some actual or hypothetical concrete scenario—e.g., a Gettier scenario, or a Twin Earth scenario.) Highly pertinent to the intuitive appeal of the one-box choice is the fact that the chooser *knows* how good the predictor is vis-à-vis the chooser herself in her current situation—and, moreover, the chooser knows this independently of any evidence she might possess concerning which action she will perform. My proposed formulations attempt to make explicit just how these facts are pertinent.

The two arguments here formulated neither assert nor presuppose that the agent's choice in Newcomb's problem will causally influence the state of box 2. That is a good thing, since clear-headed one-box reasoning should be entirely consistent with the fact—known by the agent—that there is no such causal influence.

The two arguments eschew the use of act-to-state or act-to-outcome conditional statements—either counterfactual conditionals or material conditionals. That is a good thing too, in my view. The fundamental rationale for the one-box position seems to me now not to depend upon such conditionals. This alters—and I think clarifies—the dialectical structure of the dispute between one-boxers and two-boxers. Contrary to what I maintained in Horgan (1981), the dispute is not fundamentally about whether one should use standard-resolution counterfactuals or instead should use backtracking counterfactuals when doing practical deliberation concerning Newcomb's problem.

Two normative principles figure in the arguments above—principles L5 and G5. Both are intuitively powerful. Indeed, I maintain that both are partly *constitutive* of the notion of practical rationality. I do not believe, however, that this fact leads to a clean victory for one-boxism over two-boxism. On the contrary, not only do I continue to believe that Newcomb's problem is a stalemate (a view I have held since Horgan 1985), but also I now think that Newcomb's problem is what I call a "deep antinomy of practical reason." (See Section 2 below.)

Normative principle G5 employs the notion of something's being *extremely probable*, and the notion of one thing's being *strongly preferred* to another. Both notions are qualitative, not quantitative. Something can be extremely probable without having any quantitative probability at all, either known or unknown. Likewise, one thing can be strongly preferred to another without either of the two things having any quantitative utility at all, either known or unknown (i.e., without either of the two things having desirabilities for the agent that conform to some interval-scale or ratio-scale measure that is unique up to linear transformations). This is for the best, in my view, for two interconnected reasons. First, I maintain that it is only in rare and special circumstances that real-

life decision problems are such that the potential states of nature have quantitative probabilities and the outcomes of the act/state pairs have quantitative utilities. Second, normative standards governing pragmatic rationality often apply to real-life decision problems that lack quantitative probabilities of states and/or lack quantitative utilities of outcomes. (The general version of Newcomb's problem is a case in point: although the agent perhaps has quantitative utilities that are linear with the monetary values of the potential outcomes, the scenarios in the complete act/outcome scenario-partition possess only *qualitative* degrees of likelihood—some states being extremely probable, others being extremely improbable.)

In decision problems where the states and outcomes, respectively, *do* have known quantitative probabilities and known quantitative utilities, another normative principle becomes applicable that is a quantitative analogue of the qualitative principle G5—viz., the principle of expected-utility maximization in pre-causal decision theory, where expected utility is defined the traditional way via conditional probabilities of states given acts: $V(A_i) = \Sigma_j [prob(S_j \mid A_i) \cdot u(O_{ij})]$. This principle too, I maintain, is partly constitutive of the notion of practical rationality—even though it only becomes applicable in decision problems where the states and outcomes have known quantitative probabilities and outcomes. That happens, for example, in versions of Newcomb's problem in which some specific, sufficiently high, quantitative probability is specified for the proposition that the being has correctly predicted the agent's action. But any such quantitative version of Newcomb's problem, I think, again constitutes a deep antinomy of practical reason, as discussed in the next section.

II. Newcomb's problem as a deep antinomy of practical reason

Robert Nozick begins his seminal paper on Newcomb's problem with this epigraph, a quotation from Kant's *Critique of Pure Reason*:

> Both it and its opposite must involve no mere artificial illusion such as at once vanishes upon detection, but a natural and unavoidable illusion, which even after it has ceased to beguile still continues to delude though not to deceive us, and which though thus capable of being rendered harmless can never be eradicated.
>
> Immanuel Kant, *Critique of Pure Reason*, A422, B450

Kant is here describing the antinomies of pure reason, as he construes them. For him they are illusions—albeit unavoidable ones—because they allegedly arise from the illicit tendency to try to reason about noumenal reality.

The term 'antinomy' literally means the mutual incompatibility, real or apparent, of two laws. We can distinguish three distinct kinds of potential antinomy, each of which fits this generic characterization. Let an antinomy of *type 1* have the features Kant has in mind: it is an unavoidable illusion, and it stems from the illicit tendency to try applying to noumenal reality certain categories of pure reason that cannot legitimately be so deployed. Let an antinomy of *type 2* have the features explicitly cited in the passage that Nozick uses as his epigraph—whether or not one embraces any of Kant's doctrines about the putative noumenal/phenomenal divide and about the

putative unknowability of the noumenal realm, and whether or not one construes the unavoidable illusion as arising from an illicit attempt to reason about noumenal reality. (Type 1 antinomies are thus a sub-species of type 2 antinomies.) Let an antinomy of *type 3* be a real— not merely apparent, not illusory—incompatibility between two or more normative principles, each of which is partly constitutive of some particular concept. I will call antinomies of type 3, if such there be, *deep* antinomies; this label underscores their non-illusory nature.

Nozick embraces two-boxism in his paper, which commits him to the contention that it is sometimes a requirement of pragmatic rationality to choose an act that fails to possess maximal expected utility (given the standard definition of expected utility at the time Nozick was writing, prior to the advent of causal decision theory). He embraces two-boxism on the grounds that if one available act is dominant in a matrix formulation of a decision problem, and the states in the matrix are causally independent of the acts, then practical rationality requires performing the dominant action. (This principle dictates taking two boxes in Newcomb's problem, even though taking one box maximizes expected utility as it was then understood.) He also maintains, though, that the advocate of two boxing owes an explanation why two-boxing is not *clearly* the rationally required act in Newcomb's problem, given that there are (he alleges) other decision problems where the pertinent dominance principle is clearly applicable and (traditional) expected-utility maximization is clearly mistaken. Putative cases of the latter kind include the hypothetical decision problem in which one desires to take up smoking, and one knows both (a) that smoking has no tendency to cause lung cancer, and (b) that there is a heritable gene whose presence in people who desire to take up smoking has a strong tendency to cause them to take it up, and whose absence in such people has a strong tendency to cause them to refrain from taking it up. It is clear, allegedly, that here practical rationality dictates taking up smoking, even though refraining from smoking is the act that maximizes (traditional) expected utility. Concerning the difference between such putatively clear cases and Newcomb's problem, Nozick writes:

> What then is the difference that makes some cases clear and Newcomb's example unclear, yet does not make a difference to how the cases should be decided? Given my account of what the crucial factors are (influence, and so on), my answer to this question will have to claim that the clear cases are clear cases of no influence…, and that in Newcomb's example there is the *illusion* of influence. The task is to explain in a sufficiently forceful way what gives rise to this illusion so that, even as we experience it, we will not be deceived by it. (Nozick 1969, p. 136)

He offers us a story about why/how the illusion tends to arise, and he intimates (without ever quite saying explicitly) that this illusion strongly tends to persist even once it is recognized to be an illusion—all in close alignment with Kant's remarks in the epigraph passage. In short, Nozick treats Newcomb's problem as

an antinomy of type 2, resulting from the illusion that one's choice will causally influence the state of box 2.

I applaud Nozick's thought that one-box intuitions should be treated with serious philosophical respect. It is unfortunate, I think, that so much of the recent philosophical literature on the foundations of decision theory repudiates one-box intuitions out of hand and treats the two-box choice as obviously and unproblematically the only rationally appropriate choice. Moreover, I need not deny that there is a strong tendency—at least in some people—to experience an illusion of influence in Newcomb's problem. Nor need I deny that this tendency can contribute to the intuitive appeal of one-boxism.

Nonetheless, I deny that that the psychological pull of one-boxism rests merely, or primarily, on a putative illusion of influence. On the contrary, I maintain that the fundamental rationale for the one-box choice is provided by the two normative principles I set forth in Section 1: principle L5 (applicable to the limit-case version of Newcomb's problem, in which the agent knows for sure that the being has correctly predicted what the agent will choose), and principle G5 (applicable to the original version, in which the agent knows that it is extremely probable that the predictor has correctly predicted the agent's choice). Likewise, for versions of Newcomb's problem in which some specific quantitative probability is specified concerning the predictor's having predicted correctly in the present case (and in which it is stipulated or assumed that the agent has quantitative interval-scale or ratio-scale utilities that are linear with the monetary values of the outcomes of the act/state pairs), the applicable normative principle—a quantitative analogue of the qualitative principle G5—is the principle requiring the agent to choose an act that maximizes expected utility as traditionally defined, i.e., the quantity $V(A_i) = \Sigma_j [prob(S_j \mid A_i) \cdot u(O_{ij})]$.

None of these three normative principles assumes or presupposes that the agent can influence the state of box 2. On the contrary, the principles are intuitively very plausible in and of themselves, even for cases (like Newcomb's problem) where the agent is—or anyway *should* be—fully cognizant that the available acts cannot have any causal influence on which of the pertinent states of nature obtains. (The same is true, I maintain, for suitably "cleaned up" versions of cases like the one in which lung cancer is known to be caused not by smoking but by a gene that also causes a strong tendency to take up smoking—as I argued in Horgan (1981). Cleaning up the lung cancer case, for instance, requires stipulating that a *felt desire* to take up smoking, no matter how intense, does not provide any significant evidence that one will get lung cancer—even though actually *taking up* smoking supposedly does provide strong evidence for that claim.)

Also intuitively very plausible, I readily acknowledge, are several principles that recommend two-boxing in Newcomb's problem. Let an act A_i, in a matrix formulation of a decision problem, be *qualitatively dominant* in that problem just in case (i) for each state S_j in the problem, the outcome of A_i under S_j is at least as preferable to the agent as the outcome of any other act under S_j, and (ii) for some state S_k in the problem, the outcome of A_i under S_k is more preferable to the agent than the outcome of any other act under state S_k. Likewise, if the outcomes of the act/state pairs have utilities for the agent on an interval scale or a ratio scale, then let act A_i be *quantitatively dominant* in the given problem just in

case A_i satisfies the usual definition of dominance in decision theory—viz., (a) for each state S_j, the outcome of A_i under S_j has a utility that is at least as high as the utility of the outcome of any other act under S_j, and (b) for some state S_k, the outcome of A_i under S_j has a utility that is higher than the utility of the outcome of any other act under S_k. The following two principles are both extremely plausible:

> *Qualitative dominance given causal independence*: If an act A_i in a decision problem qualitatively dominates all the other acts, and the states are causally independent of the acts, then practical rationality requires performing act A_i.
>
> *Quantitative dominance given causal independence*: If an act A_i in a decision problem quantitatively dominates all the other acts, and the states are causally independent of the acts, then practical rationality requires performing act A_i.

Also extremely plausible, for decision problems in which the agent has pertinent quantitative probabilities in addition to interval-scale or ratio-scale utilities for the outcomes of the act/state pairs, is the normative principle requiring the agent to perform an act that has the maximal causal-decision-theoretic expected utility, U—where U is to be explicated by one or another version of causal decision theory. (Perhaps, for instance, U can be defined in Gibbard and Harper's way, thus: $U(A_i) = \Sigma_j [prob(A_i \,\square\!\!\rightarrow S_j) \cdot u(O_{ij})]$, with the pertinent counterfactuals receiving a non-backtracking reading.) The principle of qualitative dominance given causal independence recommends taking two boxes in all versions of Newcomb's problem; the principle of quantitative dominance given causal dependence does so for all versions in which it is also stipulated or assumed that the agent has interval-scale or ratio-scale utilities that are linear with monetary outcomes; and the principle of U-maximization does so for all versions in which this latter assumption is supplemented with specific quantitative unconditional probabilities, for the agent, of the propositions "Box 2 contains $1 million" and "Box 2 contains nothing." (Under the non-backtracking reading of the counterfactuals, $prob(A_i \,\square\!\!\rightarrow S_j) = prob(S_j)$ for each A_i and S_j, supposedly whenever the states are causally independent of the acts.)

What explains the striking fact that, on one hand, the three normative principles mentioned two paragraphs ago are all intuitively highly plausible even when one holds in abeyance any illusion of causal influence, while, on the other hand, the three normative principles mentioned in the preceding paragraph also are intuitively highly plausible? The proper explanation, I submit, is that *each of these principles is partly constitutive of the notion of pragmatic rationality*. This means that Nozick was right to intimate that Newcomb's problem is an antinomy. But it also means, contrary to Nozick, that it is not an antinomy of type 2; the conflict in what the competing normative principles require does not arise from an illusion (and hence does not arise because the intuitive plausibility of the principles that dictate one-boxing is caused by an illusion of causal influence). Rather, it is a type 3 antinomy—a *deep* antinomy, in which distinct normative principles that really are each partly constitutive of pragmatic rationality come into direct conflict with one another. *That's* why Newcomb's problem is so maddeningly paradoxical!

What I am offering here, in support of the hypothesis that Newcomb's problem is a deep antinomy, is an abductive argument. This hypothesis, I claim, explains well some phenomena that call out for explanation—and provides a better explanation than do any alternative hypotheses. I have just mentioned one such phenomenon: the fact that all the above-mentioned normative principles are so strongly plausible intuitively, despite yielding conflicting normative recommendations in some decision problems including Newcomb's problem.

Another related phenomenon also explained well by the deep-antinomy hypothesis is the fact that there is a roughly equal split, among people who are first confronted with Newcomb's problem, between those who initially opt for one-boxing and those who initially opt for two-boxing. That would be expected, if indeed there are normative principles partly constitutive of practical rationality that dictate one-boxing and there are other normative principles, also partly constitutive of practical rationality, that instead dictate two-boxing. Given the deep-antinomy hypothesis, both groups are deploying their conceptual competence with the notion of practical rationality making their initial judgments about Newcomb's problem, even though the two groups are making conflicting judgments. All else being equal, if a proffered explanation of a widely shared pattern of intuitive judgments about how a concept applies to a thought-experimental scenario attributes those shared judgments to conceptual competence, it is a better explanation than one that instead treats the shared judgments as all resulting from some sort of conceptual performance-error.

A third phenomenon explained well by the deep-antinomy hypothesis is the fact that some people who espouse two-boxing in the official version of Newcomb's problem (in which the chooser knows only that it is extremely likely that the predictor has correctly predicted what the chooser will do) nonetheless find themselves espousing one-boxing in the limit-case version (or at least very strongly inclined that way), while also finding themselves puzzled about why one should think there is any important difference between the two versions. Strikingly, one such person was Nozick himself, which means that he apparently was not an unequivocal two-boxer. Near the end of his seminal paper, he says the following:

> If the fact that it is almost certain that the predictor will be correct is crucial to Newcomb's example, this suggests that we consider the case where it *is* certain, where you know the prediction is correct (though you do not know what the prediction is). Here one naturally argues: I know that if I take both, I will get $1000. I know that if I take only what is in the second, I get $ *M*. So, of course, I will take only what is in the second. And does a proponent of taking what is in both boxes in Newcomb's example (e.g., me) really wish to argue that it is the probability, however minute, of the predictor's being mistaken which makes the difference? Does he really wish to argue that if he knows someone using the predictor's theory will be wrong once in every 20 billion cases, he will take what is in both boxes? Could the difference between one in n, and none in n, for arbitrarily large finite n, make this difference? And how exactly does the fact that the predictor is certain to have been correct dissolve the force of the dominance argument? (Nozick 1969, pp. 140–141)

Nozick leaves the questions in this passage unaddressed, while also strongly intimating that he himself finds it obvious that one should choose only box 2 in the limit-case version. The deep-antinomy hypothesis explains well the sentiments and the puzzlement expressed in the passage, as follows. When Nozick says

> Here one naturally argues: I know that if I take both, I will get $1000. I know that if I take only what is in the second, I get $ *M*. So, of course, I will take only what is in the second.

he is revealing an appreciation of the fact the normative principle L5 is partly constitutive of pragmatic rationality. (Opting for taking both boxes even in the limit case is a very hard bullet to bite.) When he acknowledges that he himself is a proponent of taking both boxes in the original version of Newcomb's problem, he is revealing an appreciation for the fact that the principles of dominance given causal independence are also partly constitutive of pragmatic rationality (although he does not take note of the distinction I have drawn between qualitative and quantitative dominance). When he asks

> [D]oes a proponent of taking what is in both boxes in Newcomb's example (e.g., me) really wish to argue that it is the probability, however minute, of the predictor's being mistaken which makes the difference?

he is revealing an uncomfortable near-appreciation of the fact that normative principle G5 is partly constitutive of pragmatic rationality, alongside L5. And when he asks

> And how exactly does the fact that the predictor is certain to have been correct dissolve the force of the dominance argument?

he is revealing an uncomfortable near-appreciation of the fact that the principles of dominance remain partly constitutive of pragmatic rationality even with respect to the limit-case version of Newcomb's problem. All this, taken together, constitutes a near-recognition of the admittedly disturbing truth: Newcomb's problem, in both the original version and the limit-case version, is a *deep* antinomy of practical reason.

Well, what *should* the agent choose, in either version of Newcomb's problem, given that different normative principles—each partly constitutive of pragmatic rationality—yield conflicting prescriptions? And what exactly does this question even *mean*, given that unhappy situation? Perhaps one can do no better than appeal to whichever constitutive normative principles happen to exert a stronger psychological pull upon oneself: depending on how the psychological tug-of-war works out in one's own case, be a consistent one-boxer, or be a consistent two-boxer, or (like Nozick, evidently) be a two-boxer concerning Newcomb's original problem and a one-boxer concerning the limit-case version.

Speaking for myself, consistent one-boxing wins the psychological tug-of-war. Here is why. Regret is virtually inevitable in this decision situation: either I will take only the second box and then end up regretting having passed

up $1,000 that I knew all along was there for the taking in addition to the contents (if any) of the second box, or I will take both boxes and then (very probably) end up regretting that I am the kind of person about whom the being has predicted will take both boxes. Since I strongly prefer the first kind of regret to the second, I will take only box 2, collect my $1 million, and then regret that I did not take both.

References

Gibbard, A. and Harper, W. (1978). Counterfactuals and Two Kinds of Expected Utility. In C. A. Hooker *et. al.* (eds.), *Foundations and Applications of Decision Theory, Volume 1.* (Dordrecht: D. Reidel Publishing Company), 125–162.

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

Horgan, T. (1985). Newcomb's Problem: A Stalemate. In R. Campbell and L. Sowden (eds.), *Paradoxes of Rationality and Cooperation* (Vancouver: University of British Columbia Press), 223–234.

Jeffrey, R. (1965). *The Logic of Decision* (New York: McGraw-Hill).

Lewis, D. (1979). Counterfactual Dependence and Time's Arrow. *Nous* 13, 455–476.

Lewis, D. (1981). Causal Decision Theory. *Australasian Journal of Philosophy* 59, 5–30.

Nozick, R. (1969). Newcomb's Problem and Two Principles of Choice." In N. Rescher *et. al.* (eds.), *Essays in Honor of Carl G. Hempel* (Dordrecht: D. Reidel Publishing Company), 114–146.