Is Newcomb’s Problem a Problem?

Reconciling Intuitions with Prescriptions of Evidential Decision Theory

Daniel Listwa

Columbia College| Class of 2015

Submission for the Adam Leroy Jones Prize
“It is not that I claim to solve the problem, and do not want you to miss the joy of puzzling over an unsolved problem. It is that I want you to understand my thrashing about.”


At the center of a decades long debate, Newcomb’s Problem is said to present evidential decision theory with a significant challenge. Efforts at overcoming the problems have led to the formulation of an alternative to evidential decision theory, namely causal decision theory, but there have also been attempts made to prove this alternative decision theory unnecessary by showing how evidential decision theory does not encounter the problems thought to arise when it is applied to Newcomb’s Problem. The questions surrounding Newcomb’s Problem remain controversial and a lively discourse on the problem’s implications for decision theory continues.

The problem was first introduced into philosophical discourse by Robert Nozick [1969], but was created by physicist William Newcomb.¹ Nozick outlines the problem as follows: you are presented with two boxes, B1 and B2. B1 is transparent and you can see that it contains $1,000. B2 is opaque and contains either a million dollars or nothing. You are given a choice between two actions: (one-box) take only the contents of B2; and (two-box) take the contents of both boxes. Before you make your choice, you are told the following information. An extremely accurate predictor determines the content of B2. The predictor first makes his prediction (before you make your decision, say a week earlier) and then, if he predicts that you will two-box, he puts nothing in B2; if he predicts that you will one-box, he puts one million dollars in B2.

In this paper I aim to show how Newcomb’s problem motivates criticism of evidential decision theory, but ultimately fails to be a decisive challenge to this form of decision theory. By

analyzing David Lewis’s arguments in favor of two-box, I will show why there is reason to be suspicious of the one-box solution prescribed by a simple application of evidential decision theory. In response, I argue that Newcomb’s Problem as generally stated is underdetermined and that in specific formulations of the problem evidential decision theory also returns a solution of two-box. Further, I claim that in formulations of Newcomb’s problem in which evidential decision theory prescribes the one-box solution, intuitions break down to the point where the appeal of Lewis’s arguments no longer holds and it is unclear whether a return of either solution would give an intuition reason to question evidential decision theory. By briefly considering a variation of Newcomb’s Problem, I conclude that the argument against evidential decision theory fails to provide adequate justification to abandon or modify the theory.

I. TWO INTUITIVELY APPEALING ARGUMENTS

Nozick, in his original presentation of the problem, outlines two highly intuitive arguments with incompatible conclusions. The first argument favors one-box, and does so by appealing to a reasoning that can be formally represented by way of the evidential decision theory as it appears in, for example, Jeffrey [1990]. This argument holds that, given the predictor’s past record of predictive success, it is essentially certain that his prediction will be correct. Therefore, if you choose two-box, you will almost certainly take home only $1,000, while if you one-box, you will likely receive $1,000,000. We can conclude from this that the expected pay off of one-box is greater than that of two-box and, thus, one-box is the rational choice. To express this formally we calculate the expected value $V$ of each option using the type of conditionalization found in evidential decision theory. Consider the outcome matrix below.
Agent\Predictor | Predicts One-Box ($H_1$) | Predicts Two-Box ($H_1$)
---|---|---
One-Box | $1,000,000$ | $0$
Two-Box | $1,001,000$ | $1,000$

Taking $P(w|A)$ to be the agent’s subjective probability of outcome $w$ given action $A$, and taking $V(w&A)$ to be the value of outcome $w$ with $A$, we can calculate the expected value of an action $A$, $V(A)$. ³

$$V(A) = \sum_o V(w&A)P(w|A)$$

Taking Newcomb’s Problem’s statement regarding the reliability of the predictor to mean that, if the agent chooses either action one-box or two-box, she will be assign a probability of, say, .9 to the proposition that the predictor will correctly predict her action, we can assign the following conditional probabilities, using the shorthand indicated above for each outcome:

$P(H_1|\text{onebox}) = P(H_2|\text{twobox}) = .9$, $P(H_2|\text{onebox}) = P(H_1|\text{twobox}) = .1$. Using the above evidential decision theory-based formula we can then calculate the expected values of one-box and two-box, assuming for simplicity that value varies linearly with money, as

$$V(\text{onebox}) = V(H_1&\text{onebox})P(H_1|\text{onebox}) + V(H_2&\text{onebox})P(H_2|\text{onebox})$$

$$= 1,000,000(0.9) + 0(0.1) = 900,000$$

$$V(\text{twobox}) = V(H_1&\text{twobox})P(H_1|\text{twobox}) + V(H_2&\text{twobox})P(H_2|\text{twobox})$$

$$= 1,001,000(0.1) + 1,000(0.9) = 90,990$$

The above yields $V(\text{onebox}) = 900,000$ and $V(\text{twobox}) = 90,990$. Thus, $V(\text{onebox}) > V(\text{twobox})$, implying the rational choice is one-box.⁴

On the other hand, Nozick offers an argument that reasons that the predictor has already made his prediction and B2 either contains a million or has been left empty. The agent’s decision now will not change the contents of the second box. Therefore, there is either a million dollars in the box or there is not. In the case where there is a million in the box, then two-box yields $1,000,100$, while one-box yields $1,000,000$. In the case where the second box is empty, two-box yields $1,000$, while one-box yields $0$. Thus, in either case, two-box yields an additional $1,000$. By the principle of dominance, two-box is, therefore, the rational choice—contradicting the result of evidential decision theory.

Nozick offers a variation on the original problem, the “Well-intentioned Friend” variation, which strengthens the intuitions behind the dominance argument.⁵ Imagine that the opaque box has a transparent window on the back. Behind the boxes, the agent’s trusted friend is sitting and can see the content of both boxes. The friend can give the agent a recommendation on how to act. If the opaque box is empty, the friend will recommend two-box, so as to at least receive $1,000$. If the million is in the opaque box, then the friend will see it and recommend two-box, so as to ensure the agent receives the extra $1,000$. Clearly, the advice will be two-box regardless of whether the million is there or not. The friend is well-intentioned and better informed than the agent, so it would be irrational to ignore the advice. Push the variation farther and imagine both boxes are transparent and the agent can see the contents of the opaque and transparent boxes—the results are the same. The agent knows what the friend is going to recommend without needing to see into the box, since it will be two-box regardless, suggesting

⁴ Note that this result holds as long as the agent assigns a probability of at least .50050 that the predictor is correct given the agent’s action.

that the presence of the friend is actually redundant. Even without the friend, i.e. in the original variation, the agent can imagine the recommendation of such a friend and thus will choose two-box.

II. PRISONER’S DILEMMA AND THE TWO-BOX SOLUTION

To determine whether the Newcomb Problem really does lead to a challenge to evidential decision theory, it is worthwhile to question whether the dominance principle that underlies the argument in favor of two-box can be validly applied in this situation. David Lewis [1979] provides a compelling argument in favor of the application of the dominance principle and the two-box solution by restating the Newcomb Problem to be essentially identical to the Prisoners Dilemma. Lewis claims that a specification of the mechanism by which the prediction is made is inessential to the statement of Newcomb’s Problem. Given this, he suggests that one potential predictive process is simulation, through the use of a replica of the agent. This allows for a formulation in which the million is only put in the box if the replica chooses one-box. Different types of replicas could be used for the simulation, each with varying degrees of reliability. An exact duplicate of the agent, atom for atom, may make for an extremely reliable prediction; however, as already noted, the problematic nature of the Newcomb Problem arises as long as the conditional subjective probability that the prediction is correct given the agent’s action is at least .5005—which would imply a highly imperfect replica. Lewis suggests that another person would serve as a sufficiently reliable replica given this constraint.

Leaving aside the “window-dressings,” it is clear that the Prisoner’s Dilemma can be made to be seen as a particular formulation of Newcomb’s Problem. To make this explicit, we can change the choice of one-box to “not confessing” and two-box to “confessing,” which allows

---

us to create an outcome matrix that respects the preference ordering of the usual statements of both the Prisoner’s Dilemma and Newcomb’s Problem. This is illustrated in the diagram below, which gives the sentencing of the agent, given the possible actions of the “replica,” assuming fewer years in jail is preferred to more years:

<table>
<thead>
<tr>
<th>Agent\“Replica”</th>
<th>Not Confess</th>
<th>Confess</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Confess</td>
<td>1 year in jail</td>
<td>10 years in jail</td>
</tr>
<tr>
<td>Confess</td>
<td>0 years in jail</td>
<td>5 years in jail</td>
</tr>
</tbody>
</table>

It is fairly uncontroversial that in the case of the Prisoner’s Dilemma, the application of the dominance principle is appropriate. If the other prisoner does not confess, my best response is to confess. If the other does confesses, my best response is still to confess. No matter how similar to me the other prisoner is, I am better off confessing, and, thus, no matter what the Predictor predicts, two-box is the better choice. The connection to the Prisoner’s Dilemma succeeds in drawing out a crucial distinction. Although the Predictor’s prediction is probabilistically dependent on the agent’s action, the prediction is causally independent of whether the agent actually chooses one-box or two-box, since the prediction is made prior to time when the agent actually chooses.\(^7\) For this reason, the dominance argument appears highly convincing, giving good reason to support two-box as the rational choice.

III. CAUSAL DECISION THEORY

Given the apparent failure of evidential decision theory to recommend the intuitively rational choice of two-box, several theorists have attempted to formulate an alternative form of decision theory, generally referred to as causal decision theory (CDT), which explicitly makes

---

\(^7\) The prediction need not actually occur before the choice as long as the correct type of causal independence is respected.
reference to concepts discussed above, such as causal independence. While numerous fully
articulated, and largely equivalent, versions of CDT have been offered, including one found in
Lewis [1981], I here, for simplicity, make use of the sketch of such a theory from Gibbard and
Harper [1978], which appeals to intuitions regarding the probabilities of subjunctive
conditionals. The essential feature of CDT is that the agent’s judgments regarding the causal
consequences his or her actions will have on the possible outcomes are used in weighting the
expected utilities. Taking $A \square \rightarrow w$ to represent the conditional “if $A$ were chosen then $w$ would
be true,” we can represent the expected utility of $A$ as:

$$U(A) = \sum_o U(w \& A)P(A \square \rightarrow w)$$

CDT can be applied to Newcomb’s Problem in the following way. The agent assigns
probabilities to the conditionals $\text{onebox} \square \rightarrow H_1$, $\text{onebox} \square \rightarrow H_2$, $\text{twobox} \square \rightarrow H_1$, and
$\text{twobox} \square \rightarrow H_2$. Since it is the case that “if one-box is chosen, then the Predictor predicted one-
box” and “if two-box is chosen, then the Predictor predicted one-box” are both true iff the
Predictor predicted one-box, it is clear that $P(\text{onebox} \square \rightarrow H_1) = P(\text{twobox} \square \rightarrow H_1) = P(H_1)$,
which reflects the idea that the action chosen does not causally influence the likelihood that the
money is in B2, since it is either already there or not. Following Gibbard and Harper, we assign
$P(\text{onebox} \square \rightarrow H_1) = P(\text{twobox} \square \rightarrow H_1) = \mu$, where $\mu$ is the subjective unconditional
probability that the predictor predicted one-box.\textsuperscript{12} By the same reasoning, it follows

\[ P(\text{onebox}\square \rightarrow H_2) = P(\text{twobox}\square \rightarrow H_2) = 1 - \mu. \]

Using the above, we can calculate the expected utility of one-box and two-box, respectively as

\[ U(\text{onebox}) = U(\text{onebox}\& H_1)P(\text{onebox}\square \rightarrow H_1) + U(\text{onebox}\& H_2)P(\text{onebox}\square \rightarrow H_2) = 1,000,000\mu \]

\[ U(\text{twobox}) = U(\text{twobox}\& H_1)P(\text{twobox}\square \rightarrow H_1) + U(\text{twobox}\& H_2)P(\text{twobox}\square \rightarrow H_2) = 1,000,000\mu + 1,000 \]

Therefore, regardless of the value assignment of \(\mu > 0\), two-box yields a higher expected utility, in agreement with the dominance argument.

While this result is appealing, there is reason to be cautious of CDT. Causal decision theory relies on a much more complex metaphysical apparatus than evidential decision theory. While not a reason in and of itself, the fact that CDT requires these stronger assumptions elicits some measure of caution. Isaac Levi [2000] offers a more specific objection, arguing that CDT makes an unreasonable demand by having agents assign probabilities to their own actions.\textsuperscript{13} In calculating the expected utility, we invoke \(\mu\), the unconditional subjective probability that the Predictor predicted one-box. The value assignment of \(\mu\) depends on the agent’s degree of belief that the Predictor predicts that she will choose one-box. Since the Predictor is so reliable, the agent’s own beliefs about which option she will choose provides evidence on how the Predictor will predict. Thus, the agent’s determination of what action she will take depends in some part on the probabilities she assigns to the outcome of her deliberative process. There is a clear sense in which the idea of assigning probabilities to one’s own actions appears incoherent. The concept of

\textsuperscript{12} Ibid., 173.
subjective probability makes reference to fair betting-rates. Suppose that a bet costs some price to enter and consists of a payoff of one utility unit if some event E occurs. There is a price in utility units, \( p \), at which the agent is prepared to take either side of the bet—buy it or sell it—this is the fair betting-rate. The agent’s degree of belief is operationally defined as equal to \( p \). In order for the agent to have definite probability assignments, under this definition, the fair price is assumed to be unique. This definition breaks down, however, when the event E is under the agent’s control, say, if E is the event that she raises her hand. At any \( p \) (of less than one utility unit), the agent stands to make a guaranteed gain since, if she has bet against E, she will cause E to not occur, but she will cause E to occur, if she bet in favor of E. It is thus unclear whether it makes sense to assign probabilities in the way CDT involves in calculating the expected utility of one- and two-box, casting doubt on the theory.

IV. The “Tickle Defense” and Common-Cause Newcomb Problems

Due to these concerns regarding the coherency of CDT, one may favor a solution which reconciles evidential decision theory with Newcomb’s Problem rather than replaces it. Ellery Eells [1984] attempts such reconciliation by demonstrating that evidential decision theory also arrives at the conclusion that two-box is the rational choice.\(^{14}\) The “Tickle Defense,” as the argument is called, avoids the problems of the “naïve” application of evidential decision theory, by introducing a more sophisticated method of applying the theory, involving screening off certain evidence. The defense is arrived at through consideration of a variant of Newcomb’s Problem, called the “Fisher smoking hypothesis.” Suppose that an agent must decide between smoking (S) and not smoking (\( \sim S \)). The agent is concerned about the significant negative value of getting lung cancer (C) and the relatively small pleasure received from smoking (S). The agent is aware that there is a strong correlation between smoking and lung cancer, but fully believes

---

that the reason for this correlation is that there is some genetic common cause (G) that causes both cancer and the desire to smoke, and, thus, that smoking does not cause cancer nor vice versa, but rather that the gene affects beliefs and desires in such a way as to make it more likely that the individual will choose to smoke. This is illustrated schematically in fig. 1. Since smoking does not have any negative effect, but does give some utility, clearly the dominance principle dictates that one should smoke. In contrast, a naïve application of evidential utility theory would note that cancer is more probable if the agent smokes, and thus prescribe not smoking—a clearly irrational choice.

![Diagram of Fisher Smoking Hypothesis](image)

**Figure 1. Fisher Smoking Hypothesis**

To show that a sophisticated application of evidential decision theory does not prescribe this irrational choice, it is supposed that the agent knows her subjective probabilities and utility assignments. A common formulation of the “Tickle Defense” supposes that there is some sort of tickle that gives the agent information regarding her degrees of beliefs and preferences, hence the name of the argument; however, this is unnecessary since evidential decision theory assumes in the case of a perfectly rational agent that she is aware of her probability and utility assignments. Therefore, we let R be the proposition that the agent has the subjective probabilities and utility assignments that she has, and recognize that, trivially, this implies \( P(R) = 1 \). From the above it is clear that the agent believes that the gene can only affect her decision whether to smoke by affecting her beliefs and desires, as embodied in R. This implies the following

\[
P(S|R \cap G) = P(S|R \cap \sim G)
\]
The agent is aware that the gene has no causal effect on her decision, other than to affect R. Thus, given that she knows R, i.e. $P(R) = 1$, the equation becomes

$$P(S|G) = P(S|\sim G)$$

Which, by the symmetry of probabilistic independence, is equivalent to

$$P(G|S) = P(G|\sim S)$$

The equation reflects the evidential irrelevancy of the act of smoking to the subjective probability of having the bad gene. Essentially, what has been done is that the information regarding the agent’s beliefs and desires allowed her to screen off the probabilistic dependence between having the gene and choosing to smoke. This allows for the calculation of the expected value of choosing to smoke without having to take into account the statistical correlation between cancer and smoking. The result is that evidential decision theory will yield the correct result, i.e. to smoke. Clearly, since this argument also can be applied to the essentially identical story of Newcomb’s Problem as it is traditionally told, the “Tickle Defense” reconciles evidential decision theory with the prescription of the dominance principle and the common intuition.

As Eells himself admits, this result relies on the agent believing that some common cause does exist. The stipulation that there is a common cause, however, does not always appear in the formulation of the problem. In these cases, Eells says the agent should postulate a common cause, even if she cannot say what the particular causative factor is. This suggestion is convincing in the stories describing Newcomb’s Problem considered so far, so much so that Lewis [1981] concedes that the “Tickle Defense” is conclusive, at least in the case of a perfectly rational agent (though he think the case of a non-perfectly rational agent remains problematic to evidential decision theory).\footnote{Lewis, “Causal Decision Theory,” 10n7.} It is reasonable to presume the existence of some common cause in many formulations of Newcomb’s Problem. For example, if the Predictor uses some sort of
psychological analysis, then the psychological state of the agent at the time of the prediction was a cause of both the prediction and the action the agent takes. In the Prisoner’s Dilemma formulation, we can take, as Eells suggests, some shared social or genetic condition or perhaps even shared rationality between the agent and the replica as the common causative factor.\(^\text{16}\) In these cases, we can alter the “window-dressings” of the particular formulation in order to show them all to be essentially identical. For this reason, I suggest referring to these formulations of Newcomb’s Problem as Common-Cause Newcomb Problems. In these common cause scenarios, the agent justifiably believes there to be a common cause and, thus, the “Tickle Defense” provides the method to arrive at the intuitively correct answer through evidential decision theory; however, this type of scenario does not exhaust the possible formulations of Newcomb’s Problem.

V. UNDERDETERMINED FORMULATION AND STEADY-GUESS NEWCOMB PROBLEMS

The statement of the original problem supposed that the Predictor is highly reliable, meaning that the probability of his prediction being correct (C) is very high, i.e. \(P(C) \approx 1\). Though when originally calculating expected value we assumed that this statement regarding the Predictor’s reliability indicated that if the agent chooses either action one-box or two-box, she will be assign a probability close to 1 to the fact that the predictor will correctly predict her action, this is not a precise interpretation of the claim. A more exact reading of the supposition is that \(P(C) = P(\text{onebox} \cap H_1 \cup \text{twobox} \cap H_2)\) is close to 1. Since \(P(C) = P(\text{onebox} \cap H_1 \cup \text{twobox} \cap H_2) = P(\text{onebox}|H_1)P(H_1) + P(\text{twobox}|H_2)P(H_2)\), this supposition implies that at least one of the two probabilities \(P(\text{onebox}|H_1)\) or \(P(\text{twobox}|H_2)\), must be high—though not necessarily both. We could assume, for example, that both of these are high, say .9, implying

\(^{16}\) Eells, "Metatickles and the dynamics of deliberation," 93n9.
that, given that the predictor predicts one-box (two-box) there is a .9 probability that the agent will choose one-box (two-box). It does not follow, however, that given that that \( P(\text{onebox}|H_1) \) and \( P(\text{twobox}|H_2) \) are high that \( P(H_1|\text{onebox}) \) and \( P(H_2|\text{twobox}) \), or, equivalently, \( P(C|\text{onebox}) \) and \( P(C|\text{twobox}) \), are also high, which is what we had assumed in initially calculating the evidential expected value of the options. That this is the case can be illustrated through three example cases of frequency data results of an imagined experiment, offered by Levi [1975].

Case 1:

<table>
<thead>
<tr>
<th>Agent\Predictor</th>
<th>Predicts One-Box (H₁)</th>
<th>Predicts Two-Box (H₂)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chooses one-Box</td>
<td>900,000</td>
<td>10</td>
</tr>
<tr>
<td>Chooses two-Box</td>
<td>100,000</td>
<td>90</td>
</tr>
</tbody>
</table>

Case 2:

<table>
<thead>
<tr>
<th>Agent\Predictor</th>
<th>Predicts One-Box (H₁)</th>
<th>Predicts Two-Box (H₂)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chooses one-Box</td>
<td>495,045</td>
<td>55,005</td>
</tr>
<tr>
<td>Chooses two-Box</td>
<td>55,005</td>
<td>495,045</td>
</tr>
</tbody>
</table>

Case 3:

<table>
<thead>
<tr>
<th>Agent\Predictor</th>
<th>Predicts One-Box (H₁)</th>
<th>Predicts Two-Box (H₂)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chooses one-Box</td>
<td>90</td>
<td>100,000</td>
</tr>
<tr>
<td>Chooses two-Box</td>
<td>10</td>
<td>900,000</td>
</tr>
</tbody>
</table>

---

Following Levi, we can use the frequency data from the above tables to calculate the relevant probabilities (assuming that the agent sets her subjective probabilities equal to the relative frequency).

Example

\[
P(\text{onebox}|H_1) = P(H_1|\text{onebox}) \quad P(H_2|\text{twobox}) \quad P(H_1) \quad P(\text{twobox}|H_2)
\]

| Case | \(P(\text{onebox}|H_1)\) | \(P(H_1|\text{onebox})\) | \(P(H_2|\text{twobox})\) | \(P(H_1)\) | \(P(\text{twobox}|H_2)\) |
|------|-----------------|-----------------|-----------------|--------|-----------------|
| Case 1 | 0.9             | 0.9999888       | 0.0008991       | 0.9999000 |
| Case 2 | 0.9             | 0.9000000       | 0.9000000       | 0.900000  |
| Case 3 | 0.9             | 0.0008991       | 0.9999888       | 0.0000999 |

As given by Levi, each of these cases corresponds to a highly reliable Predictor. In cases 1 and 3, the Predictor makes that same prediction on almost every trial, \(H_1\) in case 1 and \(H_2\) in case 3. Further, in these two cases \(P(C|\text{onebox}) \neq P(C|\text{twobox})\), implying that the correctness of the predictions is probabilistically dependent on the agent’s choice. In these formulations of Newcomb’s Problem, we can apply evidential decision theory in the “naïve” way used earlier. First we consider case 1.

\[
V_1(\text{onebox}) = V(H_1&\text{onebox})P_1(H_1|\text{onebox}) + V(H_2&\text{onebox})P_1(H_2|\text{onebox})
\]

\[
= 1,000,000(0.9999888) + 0(1 - 0.9999888) \approx 999989
\]

\[
V_1(\text{twobox}) = V(H_1&\text{twobox})P_1(H_1|\text{twobox}) + V(H_2&\text{twobox})P_1(H_2|\text{twobox})
\]

\[
= 1,001,000(1 - 0.0008991) + 1,000(0.0008991) \approx 1000101
\]

Therefore, \(V_1(\text{twobox}) > V_1(\text{onebox})\), implying that the prescription of evidential decision theory is two-box. A similar result is found in case 3.

\[
V_3(\text{onebox}) = V(H_1&\text{onebox})P_3(H_1|\text{onebox}) + V(H_2&\text{onebox})P_3(H_2|\text{onebox})
\]

\[
= 1,000,000(1 - 0.0008991) + 0(0.0008991) \approx 999101
\]
Again, implying that evidential decision theory agrees with the dominance principle-based intuition that two-box is the rational choice. This result is quite general. In fact, it holds as long as $P(H_1|onebox) + P(H_2|twobox) < 1.001$. This implies that the result will hold for any case in which the Predictor tends to overwhelmingly often make the same prediction and that the prediction corresponds to the choice the agents almost always makes. This includes the case in which the predictions are often correct simply because the agents in the trials almost always choose the same way and that just happens to correspond to the way the Predictor almost always predicts. Note that this does not necessarily imply that the Predictor is always choosing the same way because he has noticed that the agents tend to always choose the same way, since that situation would be open to interpretation as a Common-Cause Newcomb Problem. Suppose no common causes exists and thus that the prediction is causally independent of the agent’s choice. Such a situation is not particularly far-fetched. For example, the Predictor may simply be a computer that is programmed to randomly make a prediction of $H_2$ 99.99% of the time, while almost all the agents tested are proponents of the two-box choice. Note that in such an example, $P(C)$ would still be quite high, indicating a reliable Predictor; however it would not be true that $P(onebox|H_1)$ would be close to 1. Due to the fact that these scenarios, such as case 1 and 3, involve the Predictor almost-consistently making the same guess, I call these Steady-Guess Newcomb Problems.

This follows from the following calculations: $V(twobox) > V(onebox)$ if

$V(H_1&twobox)P(H_1|twobox) + V(H_2&twobox)P(H_2|twobox) > V(H_1&onebox)P(H_1|onebox) + V(H_2&onebox)P(H_2|onebox) → 1,001,000(1 − P(H_2|twobox)) + 1,000P(H_2|twobox) > 1,000,000P(H_1|onebox) → P(H_1|onebox) + P(H_2|twobox) < 1.001$
VI. **ORACLE NEWCOMB PROBLEMS AND FAILED INTUITIONS**

Case 2 differs from the other examples Levi provides in that \( P(C|\text{onebox}) = P(C|\text{twobox}) = P(C) \), implying that the correctness of the prediction does not depend on the agent’s choice. Unlike in the other two cases and in the Steady-Guess scenarios generally, the agent has a high degree of belief that Predictor’s prediction will be correct, regardless of the choice she makes. We suppose that no common cause exists, so that the prediction and choice are causally independent as well. This implies that the agent cannot screen off evidence as she would, according to the “Tickle Defense,” if there were some common cause. Therefore, the agent again applies the evidential decision theory in the “naïve” fashion.

\[
V_2(\text{onebox}) = V(H_1|\text{onebox})P_2(H_1|\text{onebox}) + V(H_2|\text{onebox})P_2(H_2|\text{onebox}) \\
= 1,000,000(0.9) + 0(0.1) = 900,000
\]

\[
V_2(\text{twobox}) = V(H_1|\text{twobox})P_2(H_1|\text{twobox}) + V(H_2|\text{twobox})P_2(H_2|\text{twobox}) \\
= 1,001,000(0.1) + 1,000(0.9) = 90,990
\]

As we calculated earlier, this yields \( V_2(\text{onebox}) \geq V_2(\text{twobox}) \), implying that evidential decision theory prescribes one-box as the rational choice. When we first considered this result, we dismissed it as erroneous because it did not conform to the prescription derived from the dominance principle. The intuitions behind the dominance principle were supported by drawing upon the equivalencies to the Prisoner’s Dilemma. In this case, however, the comparison to the Prisoner’s Dilemma is no longer appropriate because there is no common cause. Whereas the Common-Cause and Fixed-Guess scenarios where realistic enough to allow us to confidently apply intuitions, in this case, our intuitions may be in error.

Seidenfeld [1984] argues in favor of the one-box by invoking a Newcomb’s Problem in which there is no common cause and the Predictor is infallible, by which he means
$P(C|\text{onebox}) = P(C|\text{twobox}) = 1$.\textsuperscript{19} This turns the problem into one of choice under certainty. No risk is involved and the outcomes of getting either a total of $0$ or $1,001,000$ are null-events. In this case, Seidenfeld claims any standard account of choice, including the dominance principle, would recommend one-box. Seidenfeld then considers if the Predictor is not infallible but still highly accurate, such that $P(C|\text{onebox}) = P(C|\text{twobox}) = 1 - \varepsilon$, where $\varepsilon$ is small (note that setting $\varepsilon = .1$ is case 2). He argues there is not a good enough argument for why one should switch principles of rationality when $\varepsilon$ changes from zero to a minute quantity, so if one-box is the rational choice when the Predictor is infallible, then it should also be the rational choice when the predictor is nearly infallible.

Any situation we could imagine in which the conditions considered above hold, i.e. $P(C|\text{onebox}) = P(C|\text{twobox})$ is close to or equal to one and there is no common cause, would be rather strange and exotic. For example, we could tell a story in which the Predictor uses some sort of perfect oracle, such as magic coin. For this reason, I call these scenarios Oracle Newcomb Problems. Imagine there is a perfect correlation between the coin landing heads and the agent’s decision of two-box, but there is no explanation for why this is. This situation, unlike some Common-Cause and Steady-Guess Newcomb Problems, cannot be realizable in reality given established science unless by a tremendous coincidence, which would not be expected to continue. Since our intuitions are arguably a product of our experiences, we have reason to doubt our intuitions in these reality-bending scenarios. Consider again the “Well-intentioned Friend” variation. Though the friend would always recommend two-box, the agent would have to reject the value of the recommendation—perhaps by assuming that his eyes are deceiving him—since otherwise the resulting scenario would be one in which the Predictor consistently and accurately

predicts two-box, which is a case of Steady Guess rather than an Oracle Newcomb Problem. While we could possibly respond that in no such incredible scenario could one form subjective probabilities, this does not imply that assigning $P(C|\text{onebox}) = P(C|\text{twobox})$ close to or equal to one and deriving a prescription of one-box leads to a conceptual conflict. James Cargile [1975] argues that “If you ask a scientist to speculate on what would happen if the law of gravity completely failed, you can hardly complain that his answers fly in the face of scientific theory.” By the same token, we should not be discouraged that in an Oracle situation, evidential decision theory provides a prescription that does not conform to intuitions.

VII. A “PSEUDO-NEWCOMB’S PROBLEM”

The argument against evidential decision theory claims that, while CDT and evidential decision theory generally agree, in cases where they do not agree, e.g. Newcomb’s Problem, CDT provides the intuitively correct answer, while evidential decision theory yields the intuitively wrong choice. As the discussion so far as shown, in many formulations of Newcomb’s Problem, the two decision theories do not actually disagree in their prescriptions. In the case where they do disagree, e.g. Oracle Newcomb Problems, intuitions fail to give a clear recommendation as to which response, evidential decision theory’s prescription of one-box or CDT’s two-box, is to be preferred. Since Newcomb’s Problem hits an intuitional roadblock, it is valuable to consider an alternative situation in which CDT and evidential decision theory potentially diverge, but is not as problematic in terms of what is intuitively correct. Levi [1982] offers such a problem through his “pseudo-Newcomb’s problem,” which is closely related to the Oracle scenario. Suppose that there are two urns. In the first urn (U1), there are 90 empty

---

boxes and 10 boxes containing a million dollars each. In the other urn (U2), there are 90 boxes containing a million dollars and 10 empty boxes. One opaque box has already been selected randomly from each urn. The agent is now presented with the choice of either (1-box) the opaque box from U2 on its own or (2-box) the opaque box from U1 and a transparent box containing one thousand dollars. While this situation is clearly different from the *Oracle Newcomb Problem*, the payoffs and subjective probabilities are essentially identical from the perspective of evidential decision theory and the prescribed choice, as intuition agrees, is 1-box. The prescription from CDT, however, is less clear.

While CDT can be made to yield 1-box, by applying the arguments made regarding casual influence in the original Newcomb’s Problem it can also yield a prescription of 2-box. Like in the original problem, the content of the opaque boxes was fixed well before the agent makes her decision; therefore, the agent has no casual influence over whether there is a million dollars in the box. Levi points out that CDT, specifically the form given in Gibbard and Harper [1978], holds that the probability of a state of nature outside of the agent’s control should be entered into the expected utility calculations unconditional on the acts. Therefore, if we take \( M \) to be the proposition that the selected box contains a million dollars, Levi’s argument implies \( P(1\text{box} \rightarrow M) = P(2\text{box} \rightarrow M) = P(M) = \mu \). Like in the case of the original Newcomb’s problem, the implication is that \( U(1\text{box}) = 1,000,000\mu \) and \( U(2\text{box}) = 1,000,000\mu + 1,000 \). Therefore, regardless of the value assignment of \( \mu \), this calculation prescribes 2-box. Lewis [1983] counters the claim that CDT can be made to prescribe 2-box in the way Levi outlines, by arguing that \( M \) incorrectly identifies the relevant state of nature and thus leads to a misrepresentation of the casual relationship between outcome and act.\(^{22}\) The fact, however, that CDT is so sensitive to the

identification of relevant states provides further reason to be suspicious of the theory. While evidential decision theory allows one to systematically generate consistent prescriptions, CDT leads to potential inconsistencies. While a discussion of the implications this has on the application of CDT to Newcomb’s Problem is beyond the scope of this paper, the result, along with the other objections to CDT, gives reason to doubt that CDT’s prescription of two-box in the Oracle scenario provides reason to challenge evidential decision theory for prescribing one-box.

VIII. Common-Cause and Oracle Scenarios in Contrast

In comparing Common-Cause Newcomb Problems with Oracle Newcomb Problems, I suspect some would be suspicious of the prescription of different actions when the scenarios are ostensibly identical in all but the fact that Common-Cause problems make reference to a causal relationship that does not exist in the latter variations. Though an exhaustive discussion would go beyond the scope of this paper, I will briefly comment that Common-Cause Newcomb Problems differ from Oracle Newcomb Problems structurally in a significant way as a result of the causal relationships. In the Oracle case, the oracle’s prediction can be made at any time, even simultaneously with the agent’s decision, without altering the problem in any meaningful way. In contrast, the Common-Cause predictor divides the problem into an initial stage prior to the prediction being made and a second stage after the prediction is made but before the agent chooses an action.23

The arguments for two-box always consider what the agent should do once she is at the second stage. If we assume that the agent, in the first stage, has some amount of influence over her state, which will affect the predictor’s prediction, then it is clear that the choice of one- or

two-box in the second stage does not fully describe the agent’s available strategies, since it does not specify what action she should take in the first stage. It is conceivable that a more complete statement of the agent’s available strategies would involve some sort of commitment to one-box, which is established in the first stage and realized in the second stage. Since the *Oracle* problem does not involve division into further stages, the player’s strategies can be considered fully described. Given this, both the *Common-Cause* and *Oracle* scenarios would have the agent ultimately choosing one-box and likely winning $1,000,000.

Further, if we were to conceive of Newcomb’s Problem as a sort of game between the Predictor and the agent, it should be clear that the nature of the agent’s opponent is quite different in the case of the *Common-Cause* scenario as opposed to the *Oracle* scenario. While the exact nature of the Predictor depends on the story that is told, in the *Common-Cause* case there is some sort of reasoning or computing on the part of the Predictor that leads him from the input provided by the common cause to the ultimate prediction. With an oracle, on the other hand, there is little more than a chance mechanism, albeit a magical one, being employed. This difference in itself is suggestive that a more sophisticated method of assigning conditional subjective probabilities, as the “Tickle Defense” dictates, is justified in the *Common-Cause* scenario.

IX. CONCLUSION

Newcomb’s Problem has been the source of a fracturing of decision theory into camps defending two competing theories, evidential decision theory and causal decision theory. The motivation for this divide has been the apparently separate responses to Newcomb’s Problem.

---

24 Whether such an option is available would depend on what strategies are delineated as within the agent’s menu. One conceivable method of committing to one-box might involve paying a “hired hand” to play the game in the agent’s stead, with the condition that the agent receives the reward and the hired hand will only be paid if he chooses one-box.
given by each theory. In particular, evidential decision theory is depicted as prescribing the seemingly-irrational choice of one-box, while CDT prescribes two-box. The challenge, however, to evidential decision theory is misguided in that it relies on an incomplete specification of Newcomb’s Problem. As my designations of *Common-Cause, Steady-Guess,* and *Oracle Newcomb Problems* evoke, the statement of the problem as given by Nozick fails to provide the necessary information on the nature of the Predictor’s accuracy to be able apply decision theory correctly. In particular, the statement that the Predictor is nearly or entirely infallible informs us that $P(C) = P(\text{onebox } \cap H_1 \cup \text{twobox } \cap H_2)$ is equal or close to 1, but does not tell us sufficient information regarding the conditional probabilities and the causal relationship between the prediction and the agent’s choice. While I do not claim that my three designations of particular formulations of Newcomb’s problem are unambiguous or exhaustive, they provide a framework for considering evidential decision theory’s prescriptions.

From this framework, it is clear that the argument against evidential decision theory is far from decisive. In *Common-Cause* scenarios, the “Tickle Defense” provides a way of reconciling evidential decision theory with the two-box solution, eliminating the motivation to abandon the theory in these cases. In other cases, such as the *Steady-Guess Newcomb Problems,* there is not a need for reconciliation as a simple application of the theory yields two-box. The third situation, the *Oracle* problem, forces us to confront a situation in which our intuitions break down. Unlike in the “realistic” formulations, *Oracle Newcomb Problems* diverge so far from reality that we are no longer justified claiming that evidential decision theory is in error by prescribing one-box. Thus, the *Oracle* case provides a situation in which CDT and evidential decision theory disagree and yet we cannot use our intuitions to settle the dispute. In this case, I argue that a roughly analogous “pseudo-Newcomb’s Problem,” gives us reason to be suspicious of CDT’s
prescriptions, thus giving reason to be more confident in evidential decision theory’s results.

The arguments given above are far from universally acceptable and the intuitions I have assumed are not shared by all. For those who share my intuitions regarding the rationality of the two-box solution but think abandonment or modification of evidential decision theory overly rash, there is good reason to believe these positions on Newcomb’s Problems are not incompatible.
BIBLIOGRAPHY


