Why Take Both Boxes?

Jack Spencer and Ian Wells

In the classic Newcomb problem, there is a transparent box, an opaque box, and a predictor, known by the agent to be uncannily good:[1]

> *Classic Newcomb*. The agent has two options: she can take either only the opaque box or both boxes. The transparent box contains $1,000. The opaque box contains either $0 or $1,000,000, depending on a prediction made yesterday by the predictor. The opaque box contains $0 if the predictor predicted that the agent would take both boxes, and contains $1,000,000 if the predictor predicted that the agent would take only the opaque box. The agent knows all of this.

One-boxing is the claim that an agent facing *Classic Newcomb* is rationally required to take only the opaque box. *Two-boxing* is the claim that such an agent is rationally required to take both boxes.

In this paper we attempt to fortify the case for two-boxing. Fortification is needed, we think, because the standard argument for two-boxing—a causal dominance argument—fails. The crucial premise of the standard argument is a causal dominance principle, which, to a first approximation, amounts to an injunction against choosing causally dominated options. The standard argument fails because the principle is false. As we will see, it is sometimes rationally permissible to choose causally dominated options.

Happily for two-boxers, fortification is available. There is a successful argument for two-boxing, which goes not by way of a principle connecting causal dominance to rational choice, but rather by way of a principle connecting actual value maximization to rational choice.

[1] First discussed by Nozick (1969).

1

The actual value of an option (sometimes called the value, utility, or actual utility of the option) is the value of the outcome that would result if the agent were to choose the option. For example, imagine that there are several boxes, each containing a sum of money. The agent must choose one of the boxes. The outcome that would result if the agent were to choose a particular box is that she receives the sum of money contained therein. If money is all that matters, and more money is linearly better, then the actual value of choosing a box can be identified with the number of dollars contained therein.

The main task of decision theory is to identify the options, among those available to the agent, that the agent is rationally permitted to choose. The task is easy when the agent knows the actual values of her options, for then an option is rationally permissible to choose if and only if the option maximizes actual value.[2] The task is more interesting and more difficult when the agent does not know the actual values of her options.

2

Following Leonard Savage and Richard Jeffrey,[3] many decision theorists believe that an option is rationally permissible for an agent to choose if and only if the option maximizes *expected* value, where the expected value of an option is the agent's expectation of the actual value of the option.[4] There are many well-defined expected value quantities, and there is considerable disagreement about which of them, if any, is tied to rational choice. We will focus on two: causal expected value (hereafter *c-expected value*) and evidential expected value (hereafter *e-expected* value). Both can be defined in a common conceptual framework, which centers on the concept of a decision problem.

[2] Cf. Ramsey (1990 [1926], p. 70): "Let us begin by supposing that our subject has no doubts about anything, but certain opinions about all propositions. Then we can say that he will always choose the course of action which will lead in his opinion to the greatest sum of good."

[3] Savage (1954), Jeffrey (1965).

[4] See fn. 9.

A decision problem is characterized by a set of options, a set of possible outcomes, and a decision-making agent. The options $\mathcal{A} = \{A_1, A_2, ..., A_n\}$ are the things the agent is choosing between. We take options to be propositions that the agent can make true by choosing.[5] We assume that options are finite in number, mutually exclusive, and jointly exhaustive. The possible outcomes $\mathcal{O} = \{O_1, O_2, ..., O_m\}$ are the objects of non-instrumental desire. We take outcomes to be propositions that fully specify the desirable and undesirable consequences that might result from the choice. Like options, outcomes are assumed to be finite in number, mutually exclusive, and jointly exhaustive. We associate the agent both with a credence function C and with a valuation function V. The credence function, a probabilistically coherent function that maps propositions to the unit interval, represents the agent's beliefs. The agent's credence in P, $C(P)$, is the degree to which the agent believes that P. The valuation function, which maps outcomes to real numbers, represents the agent's desires.[6] The value of outcome O, $V(O)$, is the degree to which the agent finds O non-instrumentally desirable.

Given this conception of a decision problem, the e-expected value of an option $A \in \mathcal{A}$ can be written as a credence-weighted sum, wherein the relevant credences are conditional on the option in question:

$$eev(A) = \sum_{O} C(O \mid A)V(O).$$

The *rule of e-expected value* states that agents are always rationally required to choose so as to maximize e-expected value.

Let '$\Box\!\!\rightarrow$' be the non-backtracking counterfactual conditional. If we assume that, for each option, there is a fact of the matter about which outcome would result if the agent were to choose the option,[7] then the c-expected value of

[5] In taking options to be propositions, we follow Jeffrey (1965). The agent must know that she will choose an option if she tries to do so, for reasons discussed in, among other places Hedden (2015) and Pollock (2002).

[6] The valuation function is unique up to positive affine transformation.

[7] This assumption is tantamount to counterfactual excluded middle. For a discussion of

an option can be written as a credence-weighted sum, wherein the relevant credences are unconditional credences in counterfactual conditionals:[8]

$$cev(A) = \sum_O C(A \mathrel{\square\!\!\rightarrow} O)V(O).$$

The *rule of c-expected value* states that agents are always rationally required to choose so as to maximize c-expected value.[9]

In *Classic Newcomb*, the rule of e-expected value entails one-boxing,[10] and the rule of c-expected value entails two-boxing.[11] But appealing to the rule of e-

causal decision theory without counterfactual excluded middle, see, for example, Lewis (1981), Sobel (1994, p. 141-73) and Joyce (1999).

[8]See, for example, Gibbard and Harper (1978) and Stalnaker (1981).

[9]So long as every option has an actual value (see fn. 15), both e-expected value and c-expected value can be defined as expectations of actual value. An $av(A)$-level proposition has the form $[av(A) = v]$, and is true just if v is the actual value of A. The e-expected value of an option is the agent's conditional expectation of the actual value of the option and can be written $\sum_v vC([av(A) = v] \mid A)$. The c-expected value of an option is the agent's unconditional expectation of the actual value of the option and can be written $\sum_v vC([av(A) = v])$.

[10]Let A_{1B} be the option of taking only the opaque box. Let A_{2B} be the option of taking both boxes. Conditional on A_{1B}, the agent is highly confident that the opaque box contains \$1,000,000, so, equating dollars and units of value, $eev(A_{1B}) \approx 1,000,000$. Conditional on A_{2B}, the agent is highly confident that the opaque box contains \$0, so $eev(A_{2B}) \approx 1,000$. Since $eev(A_{1B}) > eev(A_{2B})$, the rule of e-expected value entails one-boxing.

[11]Let O_0, O_T, O_M, and O_{M+T} be the outcomes of receiving \$0, \$1,000, \$1,000,000, and \$1,001,000, respectively. The agent knows that either O_0 or O_M will result if she takes only the opaque box and that either O_T or O_{M+T} will result if she takes both boxes. Moreover, she knows that her choice has no causal bearing on what sum of money is contained in the opaque box, so $C([A_{1B} \mathrel{\square\!\!\rightarrow} O_0]) = C([A_{2B} \mathrel{\square\!\!\rightarrow} O_T])$ and $C([A_{1B} \mathrel{\square\!\!\rightarrow} O_M]) = C([A_{2B} \mathrel{\square\!\!\rightarrow} O_{M+T}])$. Hence, no matter what credence function she has, the c-expected value of taking both boxes is exactly 1,000 greater than the c-expected value of taking only the opaque box:

$$\begin{aligned}
cev(A_{2B}) &= \sum_O C([A_{2B} \mathrel{\square\!\!\rightarrow} O])V(O) \\
&= C([A_{2B} \mathrel{\square\!\!\rightarrow} O_T])V(O_T) + C([A_{2B} \mathrel{\square\!\!\rightarrow} O_{M+T}])V(O_{M+T}) \\
&= (1 - C([A_{2B} \mathrel{\square\!\!\rightarrow} O_{M+T}]))(1,000) + C([A_{2B} \mathrel{\square\!\!\rightarrow} O_{M+T}]))(1,001,000) \\
&= 1,000 + C([A_{2B} \mathrel{\square\!\!\rightarrow} O_{M+T}])(1,000,000) \\
&= 1,000 + C([A_{1B} \mathrel{\square\!\!\rightarrow} O_0])(0) + C([A_{1B} \mathrel{\square\!\!\rightarrow} O_M])(1,000,000) \\
&= 1,000 + \sum_O C([A_{1B} \mathrel{\square\!\!\rightarrow} O])V(O) = 1,000 + cev(A_{1B}).
\end{aligned}$$

expected value or the rule of c-expected value cannot settle the debate between one-boxers and two-boxers, for, as you might suspect, one-boxers typically reject the rule of c-expected value, and two-boxers typically reject the rule of e-expected value.[12] To move the debate forward, we need an independent argument, one that nowhere appeals to an expected value quantity.

Many two-boxers believe that they have an independent argument: namely, a causal dominance argument.[13]

3

A natural way to argue for two-boxing is by disjunctive syllogism. We can imagine running through the argument from the agent's point of view:

> The opaque box contains either \$0 or \$1,000,000. If it contains \$0, then both boxes together contain \$1,000, and hence I would make more money if I took both boxes. If it contains \$1,000,000, then both boxes together contain \$1,001,000, and hence I would make more money if I took both boxes. Either way, I would make more money if I took both boxes. So I should take both boxes.

This argument, although unregimented, seems compelling and nowhere invokes an expected value quantity.

A preliminary attempt to regiment the argument appeals to states and dominance. A set of propositions $\mathcal{S} = \{S_1, S_2, ..., S_n\}$ is a set of *states* if its members are mutually exclusive and jointly exhaustive, and each $S \in \mathcal{S}$ is compossible with each $A \in \mathcal{A}$. Let AS be the conjunction of option A and state S. If the options and states are sufficiently fine-grained (and let us choose them so that they are), then every AS necessitates a unique outcome. If AS necessitates O,

[12] Some have tried to reconcile the rule of e-expected value with two-boxing. See, for example, Eells (1982).

[13] See, for example, Joyce (1999, p. 152-54), Lewis (1981, p. 309-12), Skyrms (1984, p. 67) and Sobel (1984).

we set $V(AS)$ equal to $V(O)$. Option A_i dominates option A_j, then, if and only if there is a set of states $\mathcal{S}$ such that, for every $S \in \mathcal{S}$, $V(A_iS)$ exceeds $V(A_jS)$.

One might allege the following connection between dominance and rational choice:

> **Dominance**: If option A_i dominates option A_j, then it is not rationally permissible for the agent to choose A_j.

But it is common ground between one-boxers and two-boxers that Dominance is false. It is sometimes rationally permissible to choose dominated options, as cases like the following make vivid:[14]

> *The Extortionist.* A moviegoer parks her car in the lot. An extortionist, who the moviegoer has excellent reason to trust, says to the moviegoer, "If you pay me $10, I'll ensure that your windshield is unbroken when you return. But I'll smash your windshield if you don't pay me."

Let the set of states be $\{S_B, S_{\neg B}\}$, where S_B is the proposition that the windshield is broken when the moviegoer returns and $S_{\neg B}$ is the proposition that the windshield is not broken when the moviegoer returns. Let A_P be the option of paying the extortionist and let $A_{\neg P}$ be the option of not paying. $V(A_{\neg P}S_{\neg B}) > V(A_PS_{\neg B})$, since it would be better by the moviegoer's lights not to pay the extortionist and return to an unbroken windshield than to pay the extortionist and return to an unbroken windshield. $V(A_{\neg P}S_B) > V(A_PS_B)$, since it would better by the moviegoer's lights not to pay the extortionist and return to a broken windshield than to pay the extortionist and return to a broken windshield. Dominance therefore entails that the agent is rationally required to not pay the extortionist—which is absurd. The moviegoer is rationally required to pay the extortionist. Paying $10 is much better than paying $1,000 for a new windshield.

[14] An adaptation of an example from Joyce (1999, p. 114-19). Jeffrey (1965, p. 9-10) uses the example of nuclear disarmament.

A bit of reflection reveals why Dominance fails. Reasoning by Dominance is supposed to put the agent in a position to conclude a fact about the ordinal ranking of options vis-à-vis actual value. The disjunctive syllogism above, for example, is supposed to put the agent in a position to conclude that the actual value of taking both boxes exceeds the actual value of taking only the opaque box. But actual value does not respect dominance: the fact that A_i dominates A_j does not entail that the actual value of A_i exceeds the actual value of A_j. Since we are assuming that, for each option, there is a fact of the matter about which outcome would result if the agent were to choose the option, we can characterize actual value as a sum. Where T is an indicator function that assigns truths to one and falsehoods to zero, the actual value of an option, $av(A)$, can be written:

$$av(A) = \sum_{O} T(A \mathrel{\square\!\!\rightarrow} O)V(O).$$

Given our assumptions, there is exactly one $O \in \mathcal{O}$ for which $[A \mathrel{\square\!\!\rightarrow} O]$ is true. If $[A \mathrel{\square\!\!\rightarrow} O]$ is true, then O is the outcome that would result if the agent were to choose A, and $av(A)$ equals $V(O)$.[15] Let $S_@$ be the state that actually obtains. The fact that A_i dominates A_j entails that $V(A_iS_@)$ exceeds $V(A_jS_@)$. It might be tempting to identify the actual values of A_i and A_j with $V(A_iS_@)$ and $V(A_jS_@)$, respectively. But that temptation must be resisted. The actual value of A is equal to $V(AS)$ only if S would have obtained had the agent chosen A. If the agent chooses A_i, then the actual value of A_i is equal to $V(A_iS_@)$. But the actual value of an unchosen option A_j need not be equal to $V(A_jS_@)$.

To illustrate, return to *The Extortionist*, and suppose that the extortionist is trustworthy. The moviegoer irrationally chooses to not pay the extortionist and

[15] If counterfactual excluded middle fails, unchosen options might fail to have actual values. When $[A \mathrel{\square\!\!\rightarrow} O]$ is true, the chance of O conditional on A, i.e. $CH(O \mid A)$, is one, so we could broaden the notion of actual value by setting $av(A)$ equal to $\sum_O CH(O \mid A)V(O)$. But it is unclear whether the broadened notion of actual value can do the meta-ethical work done by the narrower notion.

returns to a broken windshield. S_B is true, and the actual value of not paying is equal to $V(A_{\neg P}S_B)$. $V(A_{\neg P}S_B)$ exceeds $V(A_P S_B)$, of course, since not paying dominates paying. But the actual value of paying is not $V(A_P S_B)$; rather, it is $V(A_P S_{\neg B})$, since the outcome that would result if the agent were to pay the extortionist is that she would have \$10 fewer and an unbroken windshield. Moreover, $V(A_P S_{\neg B})$ *exceeds* $V(A_{\neg P}S_B)$.

While actual value does not respect *dominance*, it does respect causal dominance. A state is causally act-independent for an agent if and only if the agent knows that she has no causal influence over whether the state obtains. (More formally, S is causally act-independent for an agent if and only if the agent knows, for each $A \in \mathcal{A}$, $S \leftrightarrow [A \mathrel{\square\!\!\rightarrow} S]$.) If there is a set of causally act-independent states $\mathcal{S}$ such that, for every $S \in \mathcal{S}$, $V(A_i S)$ exceeds $V(A_j S)$, then A_i *causally dominates* A_j.[16] The alleged connection between causal dominance and rational choice is structurally identical to the alleged connection between dominance and rational choice:

> **Causal Dominance**: If option A_i causally dominates option A_j, then it is not rationally permissible for the agent to choose A_j.

But Causal Dominance is more plausible than Dominance. Causal Dominance avoids the absurd recommendation, in *The Extortionist*, that the moviegoer rationally ought to not pay the extortionist.[17]

Causal Dominance is weaker than Dominance but still strong enough to entail two-boxing. Let the set of states be $\{S_{\$0}, S_{\$M}\}$, where $S_{\$0}$ is the proposition that the opaque box contains \$0 and $S_{\$M}$ is the proposition that the opaque box contains \$1,000,000. Since $V(A_{2B}S_{\$0}) = 1,000 > 0 = V(A_{1B}S_{\$0})$, and $V(A_{2B}S_{\$M}) = 1,001,000 > 1,000,000 = V(A_{1B}S_{\$M})$, A_{2B} dominates A_{1B}. Moreover, the agent knows that she has no causal influence over the amount of

[16] If A_i causally dominates A_j, then $V(A_i S_@) > V(A_j S_@)$. Since $S_@$ is causally act-independent, $[A_i \mathrel{\square\!\!\rightarrow} A_i S_@]$ and $[A_j \mathrel{\square\!\!\rightarrow} A_j S_@]$ both are true, so $av(A_i) = V(A_i S_@)$ and $av(A_j) = V(A_j S_@)$. Hence, $av(A_i) > av(A_j)$.

[17] The moviegoer knows that she exerts causal influence over the future state of her windshield, so neither S_B nor $S_{\neg B}$ is causally act-independent.

money in the opaque box, so $S_{\$0}$ and $S_{\$M}$ are causally act-independent states. Hence, taking both boxes causally dominates taking only the opaque box, a fact exploited in the *Causal Dominance Argument* for two-boxing:

(P1) If option A_i causally dominates option A_j, then it is not rationally permissible to choose A_j.

(P2) In *Classic Newcomb*, taking both boxes causally dominates taking only the opaque box.

(C) Therefore, an agent facing *Classic Newcomb* is rationally required to take both boxes.

The Causal Dominance Argument is the aforementioned standard argument for two-boxing.[18]

Note the intimate relation between Causal Dominance and the rule of c-expected value. Given a set of causally act-independent states $\mathcal{S}$, the c-expected value of an option can be characterized as a function of the agent's unconditional credences in the members of $\mathcal{S}$:

$$cev(A) = \sum_{O} C(A \mathrel{\square\!\!\rightarrow} O)V(O) = \sum_{S} C(S)V(AS).$$

As the last sum in the equation makes clear, a causally dominated option cannot maximize c-expected value. The rule of c-expected value therefore entails Causal Dominance.

4

We believe that Causal Dominance is false, and hence that the Causal Dominance Argument is unsound. We will offer two counterexamples to the rule

[18] Some prefer an informational variant; see, for example, Pollock (2010, p. 57-82). Not every two-boxer relies on dominance reasoning. See, for example, Levi (1975).

of c-expected value, and then transform them into counterexamples to Causal Dominance.

The first counterexample is non-ideal. An ideal agent is both introspective—she knows all of the facts about her own beliefs and desires—and logically omniscient. A non-ideal agent is introspective but not logically omniscient. An ideal counterexample features an ideal agent, and a non-ideal counterexample, like the following, features a non-ideal agent:[19]

> *The Fire*. The fire alarm rings and the agent, a firefighter, hurries onto the truck. On the ride over she deliberates. She has three options: she can enter the building through the left door, the middle door, or the right door. Since she does not know the exact distribution of residents in the building, she does not know which option will result in the most rescues. Based on her credences about the distribution of residents, she calculates the c-expected value of each option and writes the value on a notecard. After exiting the truck and attaching the water hose, she races toward the building. She reaches into her pocket, but the notecard is gone! Time is of the essence. She knows that all of the residents will die in the time it would take her to recalculate the c-expected values. Her credences about the distribution of residents are unchanged, so she knows that her current c-expected values are what they were when she calculated them. But she cannot fully remember the results of her calculations. She remembers that the c-expected value of entering through the middle door is 9. Of the other two options, she remembers that one has a c-expected value of 0 and that the other has a c-expected value of 10, but she cannot remember which c-expected value goes with which option. (In fact, entering through the right door has a c-expected value of 10, as the lost notecard attests.)

[19] *The Fire* is an elaboration of a case discussed by Kagan (MS). The fact that non-ideal agents are not always able to access expected value is also discussed in, among other places, Feldman (2006) and Weirich (2004, ch. 5).

We say that the agent facing *The Fire* is rationally required to enter through the middle door, even though it is true, by hypothesis, that the option that uniquely maximizes c-expected value is entering through the right door.

The second counterexample is ideal.[20]

> *The Frustrater*. There is an envelope and two opaque boxes, A and B. The agent has three options: she can take box A, box B, or the envelope. (The three options may be labeled A_A, A_B, and A_E, respectively.) The envelope contains $40. The two boxes together contain $100. How the money is distributed between the boxes depends on a prediction made yesterday by the Frustrater, a reliable predictor who seeks to frustrate. If the Frustrater predicted that the agent would take box A, box B contains $100. If the Frustrater predicted that the agent would take box B, box A contains $100. If the Frustrater predicted that the agent would take the envelope, each box contains $50. The agent knows all of this.

We say that an ideal agent facing *The Frustrater* is rationally required to choose the envelope. But the options that maximize c-expected value are A_A and/or A_B, depending on the agent's credences.[21] (*Proof*: No matter what credence function the agent has, $cev(A_E) = 40$ and $cev(A_A) + cev(A_B) = 100$. Two numbers smaller than 40 cannot sum to 100.)[22]

[20] This example is inspired by other purported counterexamples to the rule of c-expected value: Bostrom (2001), Egan (2007) and especially Ahmed (2014b).

[21] We assume that the agent facing *The Frustrater* cannot play a mixed strategy. Perhaps the agent is unable to randomize her choice, or perhaps it is simply unwise to play a mixed strategy, since the Frustrater is very good at detecting whether an agent is playing a mixed strategy and punishes the agent severely for doing so.

[22] If we transform *The Frustrater* into a sequence of choices—first a choice between A_E and eliminating A_E, and then, if A_E is eliminated, a choice between A_A and A_B—the rule of c-expected value as applied to the sequence recommends A_E. We note three things. First, this is a different decision problem. *The Frustrater* remains a counterexample to the rule of c-expected value. Second, it may not be rationally permissible for the agent to choose between A_E and eliminating A_E—perhaps because the Frustrater punishes agents who do so. Third, not all of the counterexamples to the rule of c-expected value can be transformed into a sequence of choices, cf. Egan (2007). Thanks to Caspar Hare and Bernhard Salow for discussion on this point.

With a few alterations, both *The Fire* and *The Frustrater* can be transformed into counterexamples to Causal Dominance. Start with a variation on *The Fire*:

> *The Dominating Fire*. Everything is the same as in *The Fire*, except that, unbeknownst to the agent, the option of entering through the right door causally dominates the other two options.

From the standpoint of rationality, *The Dominating Fire* is no different than *The Fire*. A non-ideal agent might not be in a position to know which options causally dominate which others. (We can imagine that the $V(AS)$'s are stored in the agent's brain, in the form of a payoff matrix, and that it takes the agent a non-trivial amount of time to survey the matrix.) If an agent is not in a position to know that an option is causally dominated, then the fact that the option is causally dominated is not relevant to what the agent rationally ought to choose. Therefore, as in *The Fire*, an agent facing *The Dominating Fire* is rationally required to enter through the middle door, even though entering through the middle door is causally dominated by entering through the right door.

Causal Dominance is an elimination principle, which marks options as rationally impermissible to choose. But it entails the following selection principle:

> **Causal Dominance Selection**: If option A_i causally dominates all other options, then the agent is rationally required to choose A_i.

The Dominating Fire is a counterexample not just to Causal Dominance, but also to Causal Dominance Selection.

There are no ideal counterexamples to Causal Dominance Selection, a fact that we will return to, and explain, later. But there are ideal counterexamples to Causal Dominance:

> *The Semi-Frustrater*. There are two buttons, a white button and a black button. The agent has four options: she can press either button with either hand. (The four options may be labeled $A_{RH:W}$, $A_{LH:W}$, $A_{RH:B}$, and $A_{LH:B}$.) The white button connects to the

white box, the black button connects to the black box, and the agent will receive the contents of whatever box is connected to the button she presses. One of the boxes contains \$0 and the other contains \$100. Which box contains which sum depends on a prediction made yesterday by the Semi-Frustrater. The Semi-Frustrater seeks to frustrate. If the Semi-Frustrater predicted that the agent would press the black button, the white box contains \$100. If the Semi-Frustrater predicted that the agent would press the white button, the black box contains \$100. There are two left-right asymmetries. First, the agent will receive an extra \$5 if she presses a button right-handedly. Second, because the Semi-Frustrater bases her prediction on a scan of merely half of the agent's brain, the Semi-Frustrater is a 90% reliable predictor of right-handed button pressings but only a 50% reliable predictor of left-handed button pressings. The agent knows all of this.

We say that *The Semi-Frustrater*, like *The Frustrater*, is an ideal counterexample to the rule of c-expected value. In our view, the agent is rationally required to choose $A_{LH:W}$ or $A_{LH:B}$, and rationally permitted to choose either, even though the options that maximize c-expected value are, depending on the agent's credences, $A_{RH:W}$ and/or $A_{RH:B}$.[23] What is more surprising is that we have an ideal counterexample to Causal Dominance. The claim that an (ideal) agent is never rationally permitted to choose a (strictly) causally dominated option is a staple of game theory, where it appears in textbooks as the injunction against playing strategies that can be iteratively eliminated by (strict) causal domination,[24] and is regarded as sacrosanct by many expert decision theorists.[25] But

[23]Either S_W, the white box contains \$100, or S_B, the black box contains \$100. Since the agent knows that her choice has no causal influence over the contents of the boxes, $\{S_W, S_B\}$ is a set of causally act-independent states. Equating dollars and units of value, $V(S_W A_{RH:W}) = 105 = 5 + V(S_W A_{LH:W})$; $V(S_B A_{RH:W}) = 5 = 5 + V(S_B A_{LH:W})$; $V(S_W A_{RH:B}) = 5 = 5 + V(S_W A_{LH:B})$; and $V(S_B A_{RH:B}) = 105 = 5 + V(S_B A_{LH:B})$. So $cev(A_{RH:W})$ maximizes if $C(S_W) \geq 0.5$, and $cev(A_{RH:B})$ maximizes if $C(S_B) \geq 0.5$.

[24]See, for example, Fudenberg and Tirole (1991, ch. 2) and Myerson (1991, s. 3.1).

[25]Briggs (2015, p. 836): "The following is an independently compelling claim about ratio-

$A_{RH:W}$ causally dominates $A_{LH:W}$, and $A_{RH:B}$ causally dominates $A_{LH:B}$, so an ideal agent facing *The Semi-Frustrater* is rationally required to choose a causally dominated option.

5

Although the Causal Dominance Argument is unsound, a successful, independent argument for two-boxing is in the nearby vicinity. The successful argument relies on a meta-ethical principle connecting actual value maximization to rational choice.

There are two 'ought's of decision-making, an objective 'ought' and a rational 'ought'. Decision theory, being consequentialist in nature, takes both 'ought's to be reducible to value quantity maximization.

The objective 'ought' reduces to actual value maximization. Agents are always objectively required to choose so as to maximize actual value.

The objective 'ought' is not our main concern, however. Our main concern is the rational 'ought', which can, and often does, come apart from the objective 'ought'. For example:

> *Boxes like Miners.* There are three opaque boxes: the left box, the middle box, and the right box. The agent must choose exactly one box. The agent knows that the middle box contains \$9. Of the other two boxes, the agent knows that one contains \$0 and that the other contains \$10, but does not know which box contains which sum. (In fact, the right box contains \$10.)

nality: if it is knowable a priori that strategy a yields a better result than strategy b, then it is pragmatically irrational to choose strategy b when strategy a is available." Pettigrew (2015, p. 806): "the so-called Dominance Principle, which says that an option is irrational if there is an alternative that is guaranteed to be better than it, and if there is nothing that is guaranteed to be better than that alternative [...] is an uncontroversial principle of decision theory." Also see, for example, Buchak (2015), Briggs (2010), Gibbard and Harper (1978), Lewis (1981), Joyce (1999), Nozick (1969), Sobel (1994) and Skyrms (1984). In epistemic decision theory, too, the claim that (strictly) dominated options are *ipso facto* irrational is relied upon heavily. See, for example, Joyce (1998).

An agent facing *Boxes like Miners* is, though objectively required to choose the right box, rationally required to choose the middle box.

At the meta-ethical level, the most important difference between the objective 'ought' and the rational 'ought' is a difference of guidance. The objective 'ought' is not always capable of guiding the agent's choice. Actual value is the value quantity the maximization of which makes options objectively permissible for the agent to choose, but agents are not always capable of being guided by actual value. A necessary condition on being capable of being guided by actual value is being in a position to know of some option that it maximizes actual value, and agents often are in no such position. An agent facing *Boxes like Miners*, for example, is in no such position.

The rational 'ought', by contrast, is always capable of guiding the agent's choice. An agent is always capable of being guided by the value quantity the maximization of which makes options rationally permissible for the agent to choose.

It is here that we break with the meta-ethical orthodoxy. The orthodoxy has it that a value quantity is choice-guiding if and only if the facts about which options maximize the value quantity supervene on the facts about the agent's beliefs and desires.[26] Actual value fails this supervenience test. The actual value of an option is a function of the truth-values of certain counterfactual claims, and such truth-values float free of the agent's psychology. By contrast, e-expected value and c-expected value pass the supervenience test. Both are functions of the agent's beliefs and desires.

In our view, the orthodoxy is mistaken twice over. First, it is a mistake to try to divide value quantities into those that are choice-guiding and those that are not. Whether a value quantity is capable of guiding an agent's choice is settled, in our view, occasion by occasion, not once and for all. Second, it is a mistake to identify choice-guidance with supervenience on the agent's beliefs and desires. On some occasions an agent is capable of being guided by a value quantity that

[26]Or, more generally, supervene on the agent's internal mental states. See, for example, Conee and Feldman (2004).

does not supervene on her beliefs and desires, and on some occasions an agent is incapable of being guided by a value quantity that supervenes on her beliefs and desires.

We claim that a value quantity is capable of guiding an agent's choice on an occasion if and only if the agent has stable access to the value quantity on that occasion. Stable access is defined in terms of being in a position to know. An agent is *in a position to know* a proposition if and only if there is no obstacle blocking her from knowing the proposition.[27] An agent has *access* to a value quantity Q if and only if there is an option $A \in \mathcal{A}$ such that the agent is in a position to know of A that it maximizes Q. An agent has *stable access* to Q if and only if there is an option $A \in \mathcal{A}$ such that (i) the agent is in a position to know of A that it maximizes Q, and (ii) conditional on A, the agent still is in a position to know of A that it maximizes Q.[28] If an agent has stable access to Q, then the agent is *stably* in a position to know of some option A that it maximizes Q.[29]

[27] Cf. Williamson (2000, p. 95).

[28] By "conditional on A," we have the following in mind. Take the agent's credence function and conditionalize it on A. Then ask whether the agent still is in a position to know that P, relative to her updated credence function. If she is, then she is stably in a position to know that P. If not, not.

[29] Although the difference between stable access and ratifiability is not important for the purposes of this paper, here it is anyway. Stable access is a relation between an agent and a value quantity. We believe that the fact that an option maximizes a value quantity can be relevant to what an agent rationally ought to choose only if the agent has stable access to the value quantity. Ratifiability is a property of options. An option A_i is ratifiable if and only if $\sum_O C([A_i \mathrel{\square\!\!\rightarrow} O] \mid A_i)V(O) \geq \sum_O C([A_j \mathrel{\square\!\!\rightarrow} O] \mid A_i)V(O)$, for any $A_j \in \mathcal{A}$. Those who believe that ratifiability plays a role in decision theory—for example, Harper (1986), Jeffrey (1965) and Sobel (1994)—make one of two claims: either that it is never rationally permissible to choose a non-ratifiable option, a claim refuted by *The Frustrater*, or that ratifiable options are infinitely more choiceworthy than are non-ratifiable options, a claim refuted by the following decision problem, based on an example due to Skyrms (1984, p. 85-6): There are three options, A_A, A_B, and A_C. The three states, S_A, S_B, and S_C, corresponding to the reliable predictor predicting A_A, A_B and A_C, respectively. $V(A_AS_A) = 1$, $V(A_BS_A) = 0$, $V(A_CS_A) = 0$, $V(A_AS_B) = 0$, $V(A_BS_B) = 9$, $V(A_CS_B) = 10$, $V(A_AS_C) = 0$, $V(A_BS_C) = 10$ and $V(A_CS_C) = 9$. If ratifiable options are infinitely more choiceworthy than non-ratifiable options, then an agent is rationally required to choose A_A, no matter what credence function she has. We say that it is rationally impermissible for the agent to choose A_A, unless the agent is antecedently nearly certain that she will. For more on ratifiability and stability, see, among others, Arntzenius (2008), Egan (2007), Jeffrey (1965), Hare and Hedden (2015), Joyce (2012),

6

If an agent is incapable of being guided by a value quantity, then rationality does not require her to choose so as to maximize that value quantity. In our view, this is the fact that explains why the rule of c-expected value admits of counterexamples. Agents are not always capable of being guided by c-expected value—that is, agents do not always have stable access to c-expected value. An agent facing *The Fire* or *The Dominating Fire*, for example, does not have access to c-expected value, since the external time constraints, together with the agent's limited powers of deduction, form an obstacle blocking her from knowing that entering through the right door maximizes c-expected value.[30] An agent facing *The Frustrater* or *The Semi-Frustrater* has access but lacks stable access to c-expected value, since there is no option available in either decision problem that maximizes c-expected value conditional on itself.[31] We claim that the rule of c-expected value admits of counterexamples only when agents lack stable access to c-expected value. In other words, we accept the *restricted rule of c-expected value*: that agents who have stable access to c-expected value are rationally required to choose so as to maximize c-expected value.

Since we accept the restricted rule of c-expected value, we think that there is a sound argument from c-expected value to two-boxing. A competent agent facing *Classic Newcomb* has stable access to c-expected value because (i) she is in a position to know that A_{2B} (uniquely) maximizes c-expected value, and (ii) conditional on A_{2B}, she still is in a position to know that A_{2B} (uniquely) maximizes c-expected value. Hence, by the restricted rule of c-expected value,

---

Rabinowicz (1988), Weirich (1988), Weirich (2004) and Spencer and Wells (MS).

[30] *Question*: What value quantity is an agent facing *The Fire* rationally required to maximize? *Answer*: What we might call c-expected$_2$ value. A $cev(A)$-level proposition is of the form $[cev(A) = v]$. Just as the c-expected value of an option is a credence-weighted average of the agent's hypotheses about the actual value of the option (see fn. 9), the c-expected$_2$ value of an option is a credence-weighted average of the agent's hypotheses about the c-expected value of the option: $cev_2(A) = \sum_v vC([cev(A) = v])$. More generally, for any $n > 1$, $cev_n(A) = \sum_v vC([cev_{n-1}(A) = v])$.

[31] *Question*: What value quantity is an agent facing *The Frustrater* rationally required to maximize? *Answer*: See Spencer and Wells (MS), in which we develop a theory of rational choice in the face of decision instability.

she is rationally required to take both boxes.

But, as noted above, arguing from c-expected value to two-boxing fails to move the debate forward. What we need is an independent argument for two-boxing.

7

We think that the best independent argument for two-boxing goes through the restricted rule of actual value.

According to the *rule of actual value*, agents are always rationally required to choose so as to maximize actual value. Everyone rejects the rule of actual value, and for good reason. Counterexamples abound. Rational permission and actual value maximization often come apart. But the rule of c-expected value also fails: rational permission and c-expected value maximization come apart. An agent is rationally required to choose so as to maximize c-expected value only when she has stable access to c-expected value. We think that the same holds for actual value. We accept the *restricted rule of actual value*: that agents who have stable access to actual value are rationally required to choose so as to maximize actual value.

The restricted rule of actual value entails the uncontroversial claim that rational permission and actual value maximization can come apart when agents lack access to actual value. In *Boxes like Miners*, for example, the agent is rationally required to choose the middle box, even though choosing the right box uniquely maximizes actual value.

The restricted rule of actual value also entails that rational permission and actual value maximization can come apart when an agent has access but lacks stable access to actual value. Not much attention has been paid to the question of whether rational permission and actual value maximization can come apart in such cases, in part because it requires some fancy footwork to devise an example. Here is one:

Unstable Boxes like Miners. There are four boxes, the outside-left

box, the middle-left box, the middle-right box, and the outside-right box. The outside boxes are opaque and the middle boxes are transparent. The middle-left box and the middle-right box each contain \$9. One of the outside boxes contains \$0 and the other contains \$10. Which outside box contains which sum depends on a prediction made yesterday by a reliable predictor. If the predictor predicted that the agent would choose either the middle-left box or the outside-left box, the outside-right box contains \$10. If the predictor predicted that the agent would choose either the middle-right box or the outside-right box, the outside-left box contains \$10. The agent knows all this. The agent also believes that she will choose the middle-left box.

Since the agent believes that she will choose the middle-left box and believes that the predictor is extremely reliable, she believes that the outside-right box contains \$10. Moreover, let us suppose that it is true that the outside-right box contains \$10. Presumably, then, if the predictor is reliable enough, the agent *knows* that choosing the outside-right box uniquely maximizes actual value. But her epistemic position is unstable. Conditional on choosing the outside-right box, she ceases to be in a position to know that choosing the outside-right box maximizes actual value. It seems to us clear that an agent facing *Unstable Boxes like Miners* is rationally required to choose either the middle-left box or the middle-right box, and rationally permitted to choose either. Even when an agent knows which option uniquely maximizes actual value, rational permission and actual value maximization can come apart, if the agent's knowledge is unstable. This is a somewhat surprising result.

But, as concerns *Classic Newcomb*, the real substance of the restricted rule of actual value is what is says about the coincidence between rational permission and actual value maximization: namely, that rational permission and actual value maximization cannot come apart if the agent has stable access to actual value.

The simplest cases in which an agent has stable access to actual value are

cases in which the agent knows the actual values of her options. (Imagine an agent choosing among transparent boxes, each containing a sum of money.) *Classic Newcomb* is interesting in part because it is a case in which the agent has stable access to actual value without being in a position to know the actual values of her options. The agent is not in a position to know whether the actual value of A_{2B} is 1,000 or 1,001,000, for example, because she does not know whether the opaque box contains \$0 or \$1,000,000. Nevertheless, she is in a position to know that taking A_{2B} (uniquely) maximizes actual value, and, conditional on A_{2B}, she still is in a position to know that A_{2B} (uniquely) maximizes actual value.

The restricted rule of actual value entails that if an agent is stably in a position to know of an option that it uniquely maximizes actual value, the agent is rationally required to choose the option. This claim is the crucial premise of the *Actual Value Argument* for two-boxing:

(P1) If an agent is stably in a position to know of an option that it uniquely maximizes actual value, then the agent is rationally required to choose the option.

(P2) An agent facing *Classic Newcomb* is stably in a position to know of taking both boxes that it uniquely maximizes actual value.

(C) Therefore, an agent facing *Classic Newcomb* is rationally required to take both boxes.

The Causal Dominance Argument and the Actual Value Argument are closely related. If A_i causally dominates A_j, then, unless the agent is otherwise epistemically disabled, the agent is stably in a position to know that the actual value of A_i exceeds the actual value of A_j. Pointing out that taking both boxes causally dominates taking only the opaque box therefore helps to justify the minor premise of the Actual Value Argument.

The crucial difference between the Causal Dominance Argument and the

Actual Value Argument lies in their respective major premises.[32] The major premise of the Actual Value Argument amounts to the claim that agents are rationally required to be guided by actual value when they are capable of being guided by actual value. Or to put the point in deontological terms (since agents are always objectively required to choose so as to maximize actual value): in the rare cases in which the objective 'ought' provides the agent with guidance, the guidance provided by the rational 'ought' cannot conflict with the guidance provided by the objective 'ought'.

The major premise of the Causal Dominance Argument—namely, Causal Dominance—is refuted by cases like *The Dominating Fire* and *The Semi-Frustrater*. But such cases pose no threat to the major premise of the Actual Value Argument, since they are not cases in which the agent has stable access to actual value.

8

The foregoing discussion provides us not only with a sound argument for two-boxing, but also with the resources needed to explain why Causal Dominance admits of counterexamples.

[32] Ahmed (2014a, ch. 7) claims that the best argument for two-boxing goes through a principle, akin to Causal Dominance, which he calls CDB: "If you know that a certain available option makes you worse off, given your situation, than you would have been on some identifiable alternative, then that first option is irrational" (p. 202). He then formulates a weaker principle, CDB-sequence: "If you know that a certain available sequence of choices makes you worse off, given your situation, than you would have been on some identifiable alternative, then that first sequence is irrational" (p. 211, italics original). He offers a counterexample to CDB-sequence and argues that "accepting CDB and not CDB-sequence looks completely unmotivated" (p. 211). As it turns out, both *Unstable Boxes like Miners* and *The Semi-Frustrater* are counterexamples to CDB. But we do not need anything nearly as strong as CDB to motivate two-boxing. Neither *Unstable Boxes like Miners* nor *The Semi-Frustrater* are counterexamples to the restricted rule of actual value. As for Ahmed's counterexample to CDB-sequence—namely, *Newcomb Insurance*—it matters whether there is a single choice or a sequence of choices, since the value quantities to which the agent has stable access depends on it. If there is a single choice, even a single choice among sequences, we agree with Ahmed's judgments. If there is a sequence of choices, each among non-sequential options, we agree with the recommendations of the rule of c-expected value.

Nothing about what an agent rationally ought to do follows from the relations of causal dominance among the agent's options. Of course, both actual value and c-expected value respect causal dominance, so if A_i causally dominates A_j, the actual value of A_i exceeds the actual value of A_j, and the c-expected value of A_i exceeds the c-expected value of A_j. But nothing about what an agent rationally ought to do follows from the actual values of the options, and nothing about what an agent rationally ought to do follows from the c-expected values of the options. There is no direct connection between dominance or value quantity maximization and rational choice. In order to derive conclusions about what an agent rationally ought to do, we need to know, in addition to the facts about which options maximize which value quantities, the facts about which value quantities the agent has stable access to.

Once we appreciate that stable access mediates the connection between value quantity maximization and rational choice, we can explain the patterns of counterexamples to Causal Dominance that we find. Since both actual value and c-expected value respect causal dominance, and since both the restricted rule of actual value and the restricted rule of c-expected value are true, we should expect counterexamples to Causal Dominance to arise when, but only when, agents lack stable access both to actual value and to c-expected value. This is exactly what we find. There are non-ideal counterexamples to Causal Dominance because a non-ideal agent may lack stable access both to actual value and to c-expected value, despite the fact that one of her options causally dominates another (e.g., *The Dominating Fire*). There are ideal counterexamples to Causal Dominance because an ideal agent might lack stable access both to actual value and to c-expected value, despite the fact that one of her options causally dominates another (e.g., *The Semi-Frustrater*). There are non-ideal counterexamples to Causal Dominance Selection because a non-ideal agent might lack stable access both to actual value and to c-expected value, despite the fact that one of her options causally dominates all others (e.g., *The Dominating Fire*). There are no ideal counterexamples to Causal Dominance Selection because an ideal agent is guaranteed to have stable access to actual value if one

of her options causally dominates all others.[33]

References

Ahmed, A. 2014a. *Evidence, Decision and Causality*. Cambridge University Press.

—. 2014b. "Dicing with Death." *Analysis* 74:587–94.

Arntzenius, F. 2008. "No Regrets, or: Edith Piaf Revamps Decision Theory." *Erkenntnis* 68:277–297.

Bostrom, N. 2001. "The Meta-Newcomb Problem." *Analysis* 61:309–10.

Briggs, R. 2010. "Decision-Theoretic Paradoxes as Voting Paradoxes." *The Philosophical Review* 119:1–30.

—. 2015. "Costs of Abandoning the Sure-Thing Principle." *Canadian Journal of Philosophy* 45:827–40.

Buchak, L. 2015. "Revisiting Risk and Rationality: A Reply to Pettigrew and Briggs." *Canadian Journal of Philosophy* 45:841–62.

Conee, E. and Feldman, R. 2004. *Evidentialism*. Oxford University Press.

Eells, E. 1982. *Rational Decision and Causality*. Cambridge University Press.

Egan, A. 2007. "Some Counterexamples to Causal Decision Theory." *Philosophical Review* 116:94–114.

Feldman, F. 2006. "Actual Utility, the Objection from Impracticality, and the Move to Expected Utility." *Philosophical Studies* 129:49–79.

Fudenberg, D. and Tirole, J. 1991. *Game Theory*. MIT Press.

[33] An ideal agent knows that the states $\mathcal{S} = \{S_1, S_2, ..., S_n\}$ are causally act-independent. Hence, for each $A \in \mathcal{A}$ and each $S \in \mathcal{S}$, she knows that $(S \leftrightarrow [A \mathrel{\square\!\!\rightarrow} S])$. The states are causally act-independent, so there is some $S_j \in \mathcal{S}$ such that, for any $A \in \mathcal{A}$, $av(A) = V(AS_j)$. Hence, if option A_i dominates all other options, the ideal agent knows that A_i uniquely maximizes actual value. Moreover, conditional on A_i, the $V(AS)$'s remain unchanged, and the ideal agent still knows that some $S \in \mathcal{S}$ obtains; hence the ideal agent still is in a position to know that A_i uniquely maximizes actual value.

Gibbard, A. and Harper, W. 1978. "Counterfactuals and Two Kinds of Expected Utility." In Leach J. Hooker, A. and E. McClennen (eds.), *Foundations and Applications of Decision Theory*, 125–162. Reidel.

Hare, C. and Hedden, B. 2015. "Self-Reinforcing and Self-Frustrating Decisions." *Nous* 50:1–26.

Harper, W. 1986. "Mixed Strategies and Ratifiability in Causal Decision Theory." *Erkenntnis* 24:25–36.

Hedden, B. 2015. "Options and Diachronic Tragedy." *Philosophy and Phenomenological Research* 90:423–45.

Jeffrey, R. 1965. *The Logic of Decision*. University of Chicago Press.

Joyce, J. 1998. "A Nonpragmatic Vindication of Probablism." *Philosophy of Science* 65:575–603.

—. 1999. *The Foundations of Causal Decision Theory*. Cambridge University Press.

—. 2012. "Regret and Instability in Causal Decision Theory." *Synthese* 187:123–45.

Kagan, S. MS. "The Paradox of Methods." Unpublished. Accessed Spring 2012.

Levi, I. 1975. "Newcomb's Many Problems." *Theory and Decision* 6:161–75.

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

Myerson, R. 1991. *Game Theory: Analysis of Conflict*. Harvard University Press.

Nozick, R. 1969. "Newcomb's Problem and Two Principles of Choice." In N. Rescher (ed.), *Essays in Honor of Carl G. Hempel*, 114–146. Reidel.

Pettigrew, R. 2015. "Risk, Rationality, and Expected Utility Theory." *Canadian Journal of Philosophy* 45:798–826.

Pollock, J. 2002. "Rational Choice and Action Omnipotence." *Philosophical Review* 111:1–23.

—. 2010. "A Resource-Bounded Agent Addresses the Newcomb Problem." *Synthese* 176:57–82.

Rabinowicz, W. 1988. "Ratifiability and Stability." In P. Gärdenfors and N. Sahlin (eds.), *Decision, Probability, and Utility*, 406–25. Cambridge University Press.

Ramsey, F. 1990 [1926]. "Truth and Probability." In D. H. Mellor (ed.), *Philosophical Papers*. Cambridge University Press.

Savage, L. 1954. *The Foundations of Statistics*. Wiley Publications in Statistics.

Skyrms, B. 1984. *Pragmatics and Empiricism*. Yale University Press.

Sobel, J. H. 1984. "Circumstance and Dominance in a Causal Decision Theory." *Synthese* 63:167–202.

—. 1994. *Taking Chances: Essays on Rational Choice*. Cambridge University Press.

Spencer, J. and Wells, I. MS. "Rational Decision as a Metaethical Optimization Problem." Unpublished.

Stalnaker, R. 1981. "Letter to David Lewis." In Stalnaker R. Harper, W. L. and G. Pearce (eds.), *Ifs: Conditionals, Belief, Decision, Chance and Time*, 151–53. Reidel.

Weirich, P. 1988. "Hierarchical Maximization of Two Kinds of Expected Utility." *Philosophy of Science* 55:560–82.

—. 2004. *Realistic Decision Theory: Rules for Nonideal Agents in Nonideal Circumstances*. Oxford University Press.

Williamson, T. 2000. *Knowledge and its Limits*. Oxford University Press.