Can It Be Irrational to Knowingly Choose the Best?

Jack Spencer

Abstract: Seeking a decision theory that can handle both the Newcomb problems that challenge evidential decision theory and the unstable problems that challenge causal decision theory, some philosophers recently have turned to 'graded ratifiability'. The graded ratifiability approach to decision theory is, however, despite its virtues, unsatisfactory; for it conflicts with the platitude that it is always rationally permissible for an agent to knowingly choose their best option.

Keywords: Decision theory; Newcomb's Problem; Rationality; Knowledge

1/ Introduction

The two leading approaches to decision theory are evidential decision theory (EDT) and causal decision theory (CDT). Both seem to admit of counterexamples.

Newcomb problems, like the following, seem to be counterexamples to EDT.

> *Newcomb*. There is a transparent box, an opaque box, and a very reliable predictor. The agent can take either only the opaque box or both boxes ('one-box' or 'two-box'). The transparent box contains $1. The opaque box contains either $0 or $10, depending on a prediction made yesterday by the predictor. If the predictor predicted that the agent would two-box, the opaque box contains $0. If the predictor predicted that the agent would one-box, the opaque box contains $10. The agent knows all of this.

EDT recommends one-boxing, but it seems clear to many of us that the only rationally permissible option is two-boxing.

CDT recommends two-boxing. But unstable problems, like the following, seem to be counterexamples to CDT.

> *Frustrating Button*. There is a button and a very reliable predictor. The agent can either press or refrain. The agent receives $10 if they refrain. What they receive if they press depends on a prediction made yesterday by the predictor. If the predictor predicted that the agent would press, the agent receives $0 if they press. If the predictor predicted that the agent would refrain, the agent receives $15 if they press. The agent knows all of this.[1]

A decision is *prediction-sensitive* if a change in how the agent divides their credence among their options can affect which options are rationally permissible. According to CDT, Frustrating Button is prediction-sensitive. CDT recommends refraining if the agent is very confident that they will press and recommends pressing if the agent is very confident that they will refrain. But it seems clear to many of us that Frustrating Button is prediction-*insensitive*: that the only rationally permissible

[1] This is a variant of Egan's [2007] Psychopath Button, which is a variant of Richter's [1984] Modified Death Case. Other unstable decisions include Gibbard and Harper's [1978] Death in Damascus and Ahmed's [2014] Dicing with Death.

option is refraining, *irrespective* of how the agent divides their credence among their options.

Say that someone is a *refraining two-boxe*r if they accept both of the following claims:

(1) In Newcomb, two-boxing is the only permissible option, irrespective of how the agent divides their credence among their options.

(2) In Frustrating Button, refraining is the only permissible option, irrespective of how the agent divides their credence among their options.

Refraining two-boxers are a sizeable contingent.[2] You may be one, yourself—perhaps you were one already, or perhaps you are one now that you consider the matter. My primary opponents in this essay are refraining two-boxers, and, for what it is worth, I am also one, myself. I think that we need a decision theory that, unlike either EDT or CDT, predicts both (1) and (2).

An initially tempting idea—one hit upon recently by several refraining two-boxers, including Gallow [2020] and Podgorski [forthcoming]—is to turn to 'graded

[2] See, for example, Barnett [MS], Egan [2007], Gallow [2020], Gustafsson [2011], Podgorski [forthcoming], Spencer [2021], Wedgwood [2013], and Weirich [1985, 1988, 2004].

ratifiability'.[3] As we will see when the notion is defined rigorously: The graded ratifiability of two-boxing exceeds the graded ratifiability of one-boxing; the graded ratifiability of refraining exceeds the graded ratifiability of pressing; and the graded ratifiability of an option is never sensitive to how the agent divides their credence among their options. Graded ratifiability approaches to decision theory thus naturally entail:

> **Insensitivity**. No decision is prediction-sensitive.[4]

Insensitivity is tempting in its own right. As Hare and Hedden [2016: 604] say, *en route* to defending it:

> [C]onsider how odd it would sound for me to say "I believe that I will do this, so I ought to do this," and consider how much yet odder it would sound for me to say "I believe that I will do this, so I ought not to do this" […].

[3] The name 'graded ratifiability' comes from Barnett [MS], an unpublished defence of the graded ratifiability approach.

[4] One could develop a graded ratifiability approach that falsified Insensitivity, but the view would be decidedly unnatural. The best-developed graded ratifiability approaches entail Insensitivity.

The graded ratifiability approach thus has much going for it. It deserves to be taken seriously.

That said, there is a compelling argument against the graded ratifiability approach. The main premise of the argument is a principle that—answering the question posed by the title above—says that it is always rationally permissible for an agent to knowingly choose their best option.

> **Knowingly**. If an agent knows that they will choose option *a* and knows that option *a* is strictly better than every other option available to them, then it is rationally permissible for the agent to choose option *a*.

Not only do I accept Knowingly, I accept a stronger principle:

> **Strengthened Knowingly**. If an agent knows that they will choose option *a* and knows that *a* is strictly better than every other option available to them, then the agent is rationally required to choose option *a*.

But the added strength of Strengthened Knowingly is not needed. There is a compelling argument that the graded ratifiability approach should be rejected if Knowingly holds.

Indeed, there is a compelling argument that Insensitivity should be rejected if Knowingly holds. Refraining two-boxers should want to rid decision theory of a

certain kind of prediction-sensitivity, but they should not want to rid decision theory of prediction-sensitivity entirely.

2/ EDT and Knowingly

To get a better sense of Knowingly, it will be helpful to start with the conflict between it and EDT.

Some formalism: W is a (finite) set of possible worlds; C and u are the agent's *credence function* and *utility function*, respectively; A is a (finite) set of *options*, construed as propositions that the agent can make true by deciding; and K is a (finite) set of *dependency hypotheses*, which are maximal ways things the agent cares about might and might not depend on the agent's choice (*cf.* Lewis [1981: 11]). We assume that, at any possible world, exactly one option and exactly one dependency hypothesis hold, and we assume that each option is compossible with each dependency hypothesis.

The V-value of any proposition p (relative to C and u) is $\Sigma_W C(w|p)u(w)$; hence we have the V-value of option a (relative to C and u):

$$V(a) = \Sigma_W C(w|a)u(w) = \Sigma_K C(k|a)V(ak).$$

According to EDT, an option is rationally permissible just if it is V-maximising.

V-value is not sensitive to prediction. Let C^* be a credence function that the agent could have by redistributing their credence among their options; in other

words, let C^* be a credence function that can be obtained from C by Jeffrey conditionalising over A. The V-value of an option (relative to C^* and u) is equal the V-value of the option (relative to C and u), so a change in how the agent divides their credence among their options does not affect the V-values.

In Newcomb, one-boxing is the only V-maximising option. If we equate dollars and utils (as I will, hereafter), $V(a_{1\text{BOX}}) \approx 10$ and $V(a_{2\text{BOX}}) \approx 1$.[5] EDT thus recommend one-boxing—and this recommendation leads to the conflict between it and Knowingly.

Option a is strictly better than option b, in the sense relevant to Knowingly, if the objective value of a exceeds the objective value of b, where the objective value of an option is the value of the outcome that would result if the agent were to choose the option. Since we are equating dollars and utils, the objective value of one-boxing can be equated with the number of dollars contained in the opaque box, and the objective value of two-boxing can be equated with the number of dollars contained in the two boxes together. An agent facing Newcomb does not know the objective values of their options—the opaque box is opaque, after all. But the agent *does* know that the objective value of two-boxing exceeds the objective value of one-boxing; for

[5] Either the opaque box contains \$0 or \$10 (k_0 or k_{10}). $V(a_{1\text{BOX}}) = C(k_0|a_{1\text{BOX}})V(a_{1\text{BOX}}k_0) + C(k_{10}|a_{1\text{BOX}})V(a_{1\text{BOX}}k_{10}) \approx (0)(0) + (1)(10)$. $V(a_{2\text{BOX}}) = C(k_0|a_{2\text{BOX}})V(a_{2\text{BOX}}k_0) + C(k_{10}|a_{2\text{OX}})V(a_{2\text{BOX}}\text{k}_{10}) \approx (1)(1) + (0)(11)$.

the agent knows that the two boxes together contain more money than the opaque box does alone.

According to EDT, an agent is rationally required to one-box, even if the agent knows both that they will two-box and that two-boxing is strictly better than one-boxing. So EDT conflicts with Knowingly.[6]

3/ Maxrat

CDT recommends two-boxing. We can characterise the U-value of option a (relative to C and u):

$$U(a) = \Sigma_K C(k) V(ak).$$

[6] Strictly speaking, the conflict is between EDT and Knowingly+, the conjunction of Knowingly and the claim that options have objective values, since Knowingly is trivial if options do not have objective values. There is reason to think that EDT'ists should deny that options have objective values (see Ahmed and Spencer [2020]). But refraining two-boxers, being two-boxers, should accept that options have objective values (see Spencer and Wells [2019]). So, hereafter, I ignore the distinction between Knowingly and Knowingly+.

According to CDT, an option is rationally permissible just if it is U-maximising. And in Newcomb, the only U-maximising option is two-boxing.[7]

U-value is sensitive to prediction. Even if C^* can be obtained from C by Jeffrey conditionalising over A, the U-value of an option (relative to C and u) need not equal the U-value of the option (relative to C^* and u).

Newcomb is an illustration. If the agent is very confident that they will one-box, then $U(a_{1\text{BOX}}) \approx 10$ and $U(a_{2\text{BOX}}) \approx 11$; if the agent is very confident that they will two-box, then $U(a_{1\text{BOX}}) \approx 0$ and $U(a_{2\text{BOX}}) \approx 1$.[8] No matter how the agent divides their credence among their options, $U(a_{1\text{BOX}}) = U(a_{2\text{BOX}}) - 1$. So CDT does not predict that Newcomb is prediction-sensitive. But the U-values of the options are prediction-sensitive, so it should come as no surprise that CDT predicts that some decisions are prediction-sensitive.

And Frustrating Button is an illustration. No matter how the agent divides their credence among their options, $U(a_{\text{REFRAIN}}) = 10$. But the U-value of pressing is sensitive to prediction. If the agent is very confident that they will refrain, then $U(a_{\text{PRESS}}) \approx 15 > 10 = U(a_{\text{REFRAIN}})$; if the agent is very confident that they will press,

---

[7] Either the opaque box contains \$0 or \$10 (k_0 or k_{10}). $U(a_{1\text{BOX}}) = C(k_0)V(a_{1\text{BOX}}k_0) + C(k_{10})V(a_{1\text{BOX}}k_{10}) = C(k_0)(V(a_{2\text{BOX}}k_0) - 1) + C(k_{10})(V(a_{2\text{BOX}}k_{10}) - 1) = U(a_{2\text{BOX}}) - 1$.

[8] $U(a_{1\text{BOX}}) = C(k_0)(0) + C(k_{10})(10)$. $U(a_{2\text{BOX}}) = C(k_0)(1) + C(k_{10})(11)$. If the agent is very confident that they will one-box, $C(k_{10}) \approx 1$; hence $U(a_{1\text{BOX}}) \approx 10$ and $U(a_{2\text{BOX}}) \approx 11$. If the agent is very confident that they will two-box, $C(k_0) \approx 1$; hence $U(a_{1\text{BOX}}) \approx 0$ and $U(a_{2\text{BOX}}) \approx 1$.

then $U(a_{\text{PRESS}}) \approx 0 < 10 = U(a_{\text{REFRAIN}})$.[9] So, according to CDT, Frustrating Button is prediction-sensitive.

If one wants to factor the prediction-sensitivity out of CDT, there is a natural way to do so. For any option b, we can characterise the b-conditional U-value of option a (relative to C and u):

$$U(a|b) = \Sigma_K C(k|b)V(ak).$$

The self-conditional U-value of an option is its V-value. But each option has a conditional U-value on each option, and the (unconditional) U-value of an option can be thought of as a credence-weighted average of its conditional U-values: $U(a) = \Sigma_A C(b)U(a|b)$.[10]

Conditional U-values are not sensitive to prediction. If C^* can be obtained from C by Jeffrey conditionalising over A, then, for any option a and for any option b, the b-conditional U-value of a (relative to C and u) equals the b-conditional U-value of a (relative to C^* and u). So a natural way to factor the prediction-sensitivity out of

[9] Either the agent gets \$0 or \$15 by pressing (k_0 or k_{15}). $U(a_{\text{PRESS}}) = C(k_0)(0) + C(k_{15})(15)$. If the agent is very confident that they will refrain, $C(k_{15}) \approx 1$; hence $U(a_{\text{PRESS}}) \approx 15$. If the agent is very confident that they will press, $C(k_0) \approx 1$; hence $U(a_{\text{PRESS}}) \approx 0$.

[10] $\Sigma_A C(b)U(a|b) = \Sigma_A C(b)(\Sigma_K C(k|b)V(ak)) = \Sigma_K C(k)V(ak)$.

CDT is to turn from (unconditional) *U*-value to conditional *U*-values. And as it turns out, that is exactly what proponents of the graded ratifiability approach do.

The graded ratifiability of option a, relative to b, is $U(a|a) - U(b|a)$. Graded ratifiability is a relational measure of regret/gladness. If negative, it is the degree to which the agent will regret having chosen a instead of b. If positive, it is the degree to which the agent will be glad to have chosen a instead of b.

There is no such thing as the graded ratifiability of an option if there are more than two options. Each option has a graded ratifiability relative to each other. To handle *multi-option* decisions, in which there are more than two options, proponents of the graded ratifiability approach turn to tournaments. They begin with pairwise comparisons: comparing, for each pair of options, the graded ratifiability of the one relative to the other to the graded ratifiability of the other relative to the one. They then offer some tournament format: some algorithm for deriving the rationally permissible options from the various pairwise comparisons. When it comes to the tournament stage of the theory, there is great latitude. Proponents of the graded ratifiability approach can adopt one of the many algorithms for deriving a top set from pairwise preferences that have been developed in the voting theory literature (see, for example, Laslier [1997]), or they

can plump for some algorithm of their own devising.[11] But by virtue of their preferred way of comparing options pairwise, all graded ratifiability approaches are committed to a partial decision rule. When there are only two options, we can ignore the relationality of graded ratifiability, letting $R(a) = U(a|a) - U(b|a)$ and $R(b) = U(b|b) - U(a|b)$ be the graded ratifiability of options a and b, respectively. The partial decision rule, then, is this:

> **Maxrat.** If an agent has just two options, the agent is rationally required to maximise graded ratifiability.

I will argue against the graded ratifiability approach by arguing against Maxrat.[12]

---

[11] A tournament approach is a graded ratifiability approach only if it compares options pairwise by appeal to graded ratifiability. I am not aware of a tournament approach that has been marketed to refraining two-boxers that is not also a graded ratifiability approach.

[12] Gallow defines $N(A,B)$ to be $(U(A|A) - U(B|A)) - (U(B|B) - U(A|B))$ and, accepting Maxrat, says that when there are just two options, 'you should prefer A to B iff $N(A,B) > 0$, and you should be indifferent between A and B iff $N(A,B) = 0$' [2020: 133].

Podgorski calls Maxrat, 'The Promising Thought' [forthcoming: section 3]. He writes $U(Y|X)$, '$V_X(Y)$', and says, 'For all two-option cases, an agent ought to perform X over Y iff $V_X(X) - V_X(Y) > V_Y(Y) - V_Y(X)$.' He also considers a weakening of Maxrat [forthcoming: n. 10], which I criticise in n. 18.

At first blush, Maxrat looks like something refraining two-boxers should welcome. Unlike either EDT or CDT, Maxrat predicts both (1) and (2); for irrespective of how the agent divides their credence among their options, $R(a_{2\text{BOX}}) = 1 > -1 = R(a_{1\text{BOX}})$,[13] and $R(a_{\text{REFRAIN}}) \approx -5 > -10 \approx R(a_{\text{PRESS}})$.[14] Nevertheless, I think we should reject Maxrat; for Maxrat conflicts with Knowingly.

4/ An Argument Against Maxrat

My argument against Maxrat begins with the following decision:

> *Asymmetry*. There are two opaque boxes, *a* and *b*, and a very reliable predictor. The agent can take either box. What the boxes contain depends on a prediction made yesterday by the predictor. If the predictor predicted that the agent would choose *a*, then *a* contains \$10 and *b* contains \$0. If the predictor predicted that the agent would choose *b*, then *a* contains \$0 and *b* contains \$15. The agent knows all of this.

---

[13] Either the opaque box contains \$0 or \$10 (k_0 or k_{10}). $R(a_{2\text{BOX}}) = U(a_{2\text{BOX}}|a_{2\text{BOX}}) - U(a_{1\text{BOX}}|a_{2\text{BOX}}) = C(k_0|a_{2\text{BOX}})(V(a_{2\text{BOX}}k_0) - V(a_{1\text{BOX}}k_0)) + C(k_{10}|a_{2\text{BOX}})(V(a_{2\text{BOX}}k_{10}) - V(a_{1\text{BOX}}k_{10})) = 1$. $R(a_{1\text{BOX}}) = U(a_{1\text{BOX}}|a_{1\text{BOX}}) - U(a_{2\text{BOX}}|a_{1\text{BOX}}) = C(k_0|a_{1\text{BOX}})(V(a_{1\text{BOX}}k_0) - V(a_{2\text{BOX}}k_0)) + C(k_{10}|a_{1\text{BOX}})(V(a_{1\text{BOX}}k_{10}) - V(a_{2\text{BOX}}k_{10})) = -1$.

[14] Either pressing yields \$0 or \$15 (k_0 or k_{15}). $R(a_{\text{REFRAIN}}) = U(a_{\text{REFRAIN}}|a_{\text{REFRAIN}}) - U(a_{\text{PRESS}}|a_{\text{REFRAIN}}) = C(k_0|a_{\text{REFRAIN}})(V(a_{\text{REFRAIN}}k_0) - V(a_{\text{PRESS}}k_0)) + C(k_{15}|a_{\text{REFRAIN}})(V(a_{\text{REFRAIN}}k_{15}) - V(a_{\text{PRESS}}k_{15})) \approx -5$. $R(a_{\text{PRESS}}) = U(a_{\text{PRESS}}|a_{\text{PRESS}}) - U(a_{\text{REFRAIN}}|a_{\text{PRESS}}) = C(k_0|a_{\text{PRESS}})(V(a_{\text{PRESS}}k_0) - V(a_{\text{REFRAIN}}k_0)) + C(k_{15}|a_{\text{PRESS}})(V(a_{\text{PRESS}}k_{15}) - V(a_{\text{REFRAIN}}k_{15})) \approx -10$.

Let an *a-confident* Asymmetry be a version of Asymmetry in which the agent is very confident that they will choose a, and let an *a-veridical* Asymmetry be an a-confident Asymmetry in which, in fact, a contains \$10 and the agent will choose a.

According to Maxrat, an agent facing an a-veridical Asymmetry is rationally required to choose b, since $R(a) \approx 10 < 15 \approx R(b)$.[15] But the following principle seems true:

> **Known**. An agent facing an a-veridical Asymmetry knows both that they will choose a and that a is strictly better than b.

And Knowingly and Known together entail that it is rationally permissible for an agent facing an a-veridical Asymmetry to choose a.

The argument against Maxrat is thus straightforward. Knowingly, Known, and Maxrat cannot all be true. Knowingly and Known are true. So Maxrat isn't.

5/ Four Possible Reponses

Let me consider four ways a proponent of Maxrat might respond.

---

[15] Either b has \$0 or \$15 (k_0 or k_{15}). $R(a) = U(a|a) - U(b|a) = C(k_0|a)(V(ak_0) - V(bk_0)) + C(k_{15}|a)(V(ak_{15}) - V(bk_{15})) \approx 10$. $R(b) = U(b|b) - U(a|b) = C(k_0|b)(V(bk_0) - V(ak_0)) + C(k_{15}|b)(V(bk_{15}) - V(ak_{15})) \approx 15$.

5.1. A New Rational Requirement

Knowingly and Known enjoy considerable intuitive support, so a proponent of Maxrat might start by seeking a reconciliation. The three claims—Knowingly, Known, and Maxrat—cannot be reconciled if an *a*-veridical Asymmetry is possible, and there is no latent contradiction in the specification of the case. It is not impossible for an agent to face an *a*-veridical Asymmetry. But perhaps a proponent of Maxrat could deny that it is possible for an *ideal* agent to face an *a*-veridical Asymmetry. If it is impossible for an ideal agent to face an *a*-veridical Asymmetry and Maxrat is restricted to ideal agents, then the conflict disappears; for Maxrat then makes no prediction about which options are rationally permissible for an agent facing an *a*-veridical Asymmetry.

The initial burden of this response is finding a principle of ideal rationality that must be violated by an agent facing an *a*-veridical Asymmetry. The usual suspects won't do. The utilities of the agent are coherent and well-behaved. The credences satisfy the probability axioms. The violated principle almost certainly will be one that is not yet acknowledged as a principle of ideal rationality.

We can envisage principles that would do the work. Some CDT'ists claim that agents are rationally required to divide their credence among their options in a way that reflects their *U*-values (see, for example, Joyce [2018] and Skyrms [1990]). In a similar spirit, a proponent of Maxrat could say that agents choosing between two options are rationally required to be confident that they will choose an option that

maximises graded ratifiability. This principle (or another to a similar effect) would ensure that it is impossible for an ideal agent to face an *a*-veridical Asymmetry.

But motivating this principle is not easy; for the considerations that tell against Maxrat also tell against it. If this principle is true, then it is irrational for an agent facing an *a*-veridical Asymmetry to be confident that they will choose *a*. But it is very far from obvious that it is irrational for an agent facing an *a*-veridical Asymmetry to be confident that they will choose *a*. After all, they know that they will choose *a*, care only about money, and know that *a* contains more money than *b* does.

So the first challenge to a proponent of Maxrat who pursues this response is finding some principle that is necessarily violated by an agent facing an *a*-veridical Asymmetry and defending the claim that the principle is indeed a principle of ideal rationality.

The second challenge is motivating Maxrat once Maxrat is restricted to ideal agents and the new principle of rationality is imposed. Much of the appeal of Maxrat is owed to the fact that it entails both (1) and (2). But if we restrict Maxrat to ideal agents and insist that an ideal agent choosing between two options is always confident that they will choose an option that maximises graded ratifiability, then Maxrat no longer entails (1) or (2); for Maxrat then makes no predictions about which options are rationally permissible for an agent who is confident that they will one-box or press.

The motivation for Maxrat could be regained if we coupled Maxrat with some principles of rational decision-making that apply to agents who are not quite ideal,

like agents who are confident that they will one-box or press. But this combo package is vulnerable to the conjunction of Knowingly and Known in more or less the same way that Maxrat, without the newly proposed principle of ideal rationality, is.

5.2. Restricting Maxrat

One can reconcile Maxrat, Knowingly, and Known, without proposing a new principle of ideal rationality, just by restricting Maxrat appropriately. For example, the conflict goes away if one restricts Maxrat to cases in which an agent does not foreknow which option they will choose.

But a proponent of Maxrat who wants to respond to the argument by restricting Maxrat faces three challenges.

The first is a motivational problem. If Maxrat is restricted to cases in which the agent does not foreknow what they will choose, then Maxrat no longer entails (1) and (2). It thus becomes unclear whether a motivation for Maxrat can be evinced.

The second challenge is finding a suitably general restriction. Knowingly and Known are concerned with knowledge, but rationality is a function of credences and utilities. If it is rationally permissible for an agent facing an a-veridical Asymmetry to choose a, then it is rationally permissible for an agent facing any a-confident Asymmetry to choose a. Not every a-confident Asymmetry is one in which the agent foreknows what they will choose. An agent facing an a-confident Asymmetry might

deviate from their own expectations and choose *b*, for example. An *a*-confident Asymmetry in which an agent does not foreknow what they will choose is thus a counterexample to Maxrat, even if we restrict Maxrat to cases in which an agent does not foreknow what they will choose. The foreknowledge restriction is thus not restrictive enough. To insulate Maxrat from the threat posed by Knowingly and Known, we need to ensure that Maxrat is silent about every *a*-confident Asymmetry.

The third challenge is exhibiting the philosophical interest of the restricted principle. One can arrive at an exceptionless principle by restricting Maxrat to cases in which it is unmistaken, but the resultant principle sheds no light on rational decision-making, and without some rather convincing argument that we should expect Maxrat to need some restriction, one worries that the principle we arrive at by restricting Maxrat will be, even if counterexample-free, of little philosophical interest.

5.3. Deny Known

Since the prospects of reconciling Maxrat, Knowingly, and Known seem dim, perhaps a more straightforward response is preferable. Could a proponent of Maxrat deny Known?

The argument against Maxrat does not require that Known be true of every *a*-veridical Asymmetry. It requires only that Known be true of some *a*-veridical Asymmetry, and the prima facie case for this existential claim is strong.

We often know what we will choose before choosing, and we sometimes know that we will choose what we have been predicted to choose. Prior to visiting nytimes.com, I knew that I would visit nytimes.com, and I knew that Google's algorithms predicted that I would visit nytimes.com. Prior to ordering the salad, I knew that I would order the salad, and I knew that my loved ones predicted that I would order the salad. When it comes to predicting my choices, Google's algorithms and my loved ones are reliable. But the predictor in Asymmetry is more reliable still. So if I can know both that I will order the salad and that my loved ones predicted that I would order the salad, then an agent facing an *a*-veridical Asymmetry can know both that they will choose *a* and that the predictor predicted that they would choose *a*.

Of course, there are theses that contradict Known. One could deny that an agent ever foreknows what they will choose, for example, or one could deny that anyone ever knows anything about the future. But it is not unreasonable to demand that a decision theory cohere with our ordinary epistemic standards. A decision theory should not commit us to skeptical theses that are otherwise unwanted. And judging by the epistemic standards that underlie our ordinary attributions of knowledge, it seems clear that Known is true: that an agent facing an *a*-veridical Asymmetry does (or anyway could) know both that they will choose *a* and that *a* is strictly better than *b*.

5.4. Deny Knowingly

If Knowingly, Known, and Maxrat cannot be reconciled, and Known cannot reasonably be denied, the last strategy available to a proponent of Maxrat is denying Knowingly.

But Knowingly is, I think, undeniable. Imagine trying to convince someone of their irrationality in any putative counterexample. Whatever you say, whatever mathematical sophistication you bring to bear, whatever rhetoric about regret or degrees of ratifiability you offer, they can say, in reply, 'I knew that I would choose this option before I chose it, and I knew that this was my best option before I chose it.' And that, as a reply to alleged irrationality, seems, to my mind, dispositive.[16] Irrational decision-making is defective. But there is nothing defective about an agent choosing an option if the agent can choose the option because it is their best option; and if an agent can know both that they will choose an option and that the option is

[16] *Objection*: Knowingly fails when extreme outcomes are possible. If I know that my house will not catch fire this year, I know that refusing the insurance is strictly better than buying. But if the price of insurance is sufficiently low, I might be rationally required to buy. *Reply:* One does not know in such a case. One cannot know that a is strictly better than b if, relative to one's own credences and utilities, $U(b)$ exceeds $U(a)$. But note that the argument against Maxrat does not require the full strength of Knowingly; it could be run with a principle that restricts Knowingly to cases in which the option known by the agent to be best is also (known by the agent to be) expectedly best, as measured by U-value.

strictly better than every other option available to them, then the agent can choose the option because it is their best option.[17]

As I said, I accept Strengthened Knowingly. I think that two-boxing in Newcomb is not just rationally permitted, but rationally required, and I think that Strengthened Knowingly explains why two-boxing is rationally required. Similarly, I think that choosing *a* in an *a*-confident Asymmetry is not just rationally permitted, but rationally required, and I think that Strengthened Knowingly explains why choosing *a* is rationally required.

But the added strength of Strengthened Knowingly is not needed in the argument against Maxrat. All that is needed is the apparent platitude that it is always rationally permissible for an agent to knowingly choose their best option.

6/ Two Takeaways

[17] This explains why, although I accept Knowingly, I reject:

> **Knowing**. If an agent knows that option *a* is strictly better than every other option available to them, then it is rationally permissible for the agent to choose *a*.

If an agent's knowledge that an option is their best would be destroyed by coming to believe that they will choose the option, then the agent cannot choose the option because it is their best, even though they know that the option is their best. Some cases of this sort are counterexamples to Knowing (see, for example, Richter [1984: 396–98] and Spencer and Wells [2019: 40]).

Graded ratifiability approaches to decision theory hold some initial appeal for refraining two-boxers. Refraining two-boxers seek a decision theory that, unlike either EDT or CDT, entails both (1) and (2), and graded ratifiability approaches entail Maxrat and thus entail both (1) and (2). But the initial appeal does not persist; for as we have seen, there is a compelling argument Maxrat. Maxrat conflicts with the platitude that it is always rationally permissible for an agent to knowingly choose their best option.

An argument against Maxrat is interesting in its own right, but there are two more general takeaways that can be wrung from the discussion heretofore.

6.1 Insensitivity

The first concerns Insensitivity. Formally, a decision, d, is a quadruple, $\langle A,K,C,u\rangle$, where A is the set of options, K is the set of dependency hypotheses, C is the credence function, and u is the utility function. If C^* is a credence function that can be obtained from C by Jeffrey conditionalising over A, we will say that $d^* = \langle A,K,C^*,u\rangle$ is *predictively accessible* from d. According to Insensitivity, the rationally permissible options relative to any decision d are also the rationally permissible options relative to any decision predictively accessible from d.

It is sometimes said that a decision theory is structurally defective if it conflicts with Insensitivity (see, for example, Hare and Hedden [2016] and Richter [1984]). But there is a straightforward argument against Insensitivity.

Consider a *b-confident* Asymmetry, in which the agent is very confident that they will choose *b*. All of the decision theories considered herein—EDT, CDT, and any theory that entails Maxrat—agree that an agent facing a *b*-confident Asymmetry is rationally required to choose *b*, and for good reason.[18] An agent facing a *b*-confident Asymmetry is confident that choosing *a* will lead to the worst outcome ($0) and that choosing *b* will lead to the best ($15). But if an agent facing an *a*-confident Asymmetry is rationally permitted to choose *a* and an agent facing a *b*-confident Asymmetry is rationally required to choose *b*, then Insensitivity is false.

The point thus cuts the other way. It is the decision theories that entail Insensitivity that are structurally defective.

6.2 Decision Instability

The second concerns decision instability. Refraining two-boxers think that some unstable decisions are counterexamples to CDT. But what marks a decision as unstable? Two proposals suggest themselves.

Say that decision $d = \langle A,K,C,u \rangle$ is *U-insensitive* if the options that maximise *U*-value relative to it are also the options that maximise *U*-value relative to every

[18] According to the weakening of Maxrat that Podgorski [forthcoming: n. 10] considers, an agent with just two options is rationally required to maximise graded ratifiability or choose an option with positive graded ratifiability. I take the weakened principle to be refuted by its prediction that *a* is rationally permissible relative to a *b*-confident Asymmetry.

decision predictively accessible from it. A decision is *U-sensitive* if it is not *U*-insensitive, and the first proposal identifies instability with *U*-sensitivity.

Say that decision $d = \langle A,K,C,u \rangle$ is *U-stable* if some option maximises both *U*-value and self-conditional *U*-value relative to it. A decision is *U-unstable* if it is not *U*-stable, and the second proposal identifies instability with *U*-instability.

The two proposals agree about Frustrating Button and other paradigm unstable decisions; for the paradigm unstable decisions are *U*-unstable, and every *U*-unstable decision is *U*-sensitive. But they disagree about Asymmetry and other *U*-sensitive-yet-stable decisions, and they lend themselves to different diagnoses of where CDT goes wrong.

The *U*-sensitive decisions are, according to CDT, the prediction-sensitive decisions. So if instability is *U*-sensitivity, then it is natural to think that the defect in CDT that unstable decisions exploit is the commitment to prediction-sensitivity.

The identification of instability and *U*-instability suggests a different diagnosis. An agent can choose an option because it is *U*-maximising just if they face a *U*-stable decision. An agent can choose an option because it is *U*-maximising just if the agent can know both that they will choose the option and that the option is *U*-maximising, and an agent can know both that they will choose an option and that the option is *U*-maximising just if the option maximises both *U*-value and self-conditional *U*-value. So if instability is *U*-instability, then it is natural to think that the defect in CDT that unstable decisions exploit is not the commitment to prediction-sensitivity but rather the commitment to *reason elusiveness*—the

prediction that a U-maximising option is rationally permissible, even when an agent cannot choose the option because it is U-maximising. Reason elusiveness requires prediction-sensitivity; it can be thought of as a certain kind of prediction-sensitivity. But if the defect in CDT is its commitment to reason elusiveness, then no U-stable decision poses the distinctive threat to CDT that unstable decisions pose. Since I think that a commitment to prediction-sensitivity is a not a defect but rather a condition of adequacy for a decision theory, I favour the second proposal, which classifies U-stable-yet-sensitive decisions as stable.

A clearer understanding of decision instability is relevant to future work. Some refraining two-boxers—myself, included—seek a stability-preserving successor to CDT: a decision theory that agrees with CDT about every stable decision, but better handles unstable decisions. If instability is U-sensitivity, then Maxrat could be a part of a stability-preserving successor to CDT, since Maxrat and CDT agree about every U-insensitive two-option decision.[19,20] But if instability is U-instability, then no decision theory that entails Maxrat can be a stability-preserving successor to CDT; for Maxrat and CDT do not agree about every U-stable two-option decision, as witnessed by an a-confident Asymmetry.[21]

---

[19] If a two-option decision is U-insensitive, then $R(a) = U(a) - U(b)$ and $R(b) = U(b) - U(a)$.

[20] Many graded ratifiability approaches agree with CDT about every U-insensitive decision.

[21] My sincerest thanks, for comments, questions, and encouragement, to the editor, to two anonymous referees, and to David James Barnett, J. Dmitri Gallow, and Caspar Hare.

References

Ahmed, Arif 2014. Dicing with Death, *Analysis* 74/4: 587–92.

Ahmed, Arif and Jack Spencer 2020. Objective Value Is Always Newcombizable, *Mind* 129/516: 1157–1192.

Barnett, David James MS. Graded Ratifiability.

Egan, Andy 2007. Some Counterexamples to Causal Decision Theory, *Philosophical Review* 116/1: 93–114.

Gallow, J. Dmitri 2020. The Causal Decision Theorist's Guide to Managing the News, *Journal of Philosophy* 117/ 3: 117–49.

Gibbard, Allan and William L. Harper 1978. Counterfactuals and Two Kinds of Expected Utility, in *Foundations and Applications of Decision Theory*, ed. C. A. Hooker, J. J. Leach, and E. F. McClennan, Boston: D. Reidel: 125–62.

Gustafsson, Johan 2011. A Note in Defense of Ratificationism, *Erkenntnis* 75/1: 147–50.

Hare, Caspar and Brian Hedden 2016. Self-Reinforcing and Self-Frustrating Decisions, *Noûs* 50/3: 604–28.

Joyce, James M. 2018. Deliberation and Stability in Newcomb Problems and Pseudo-Newcomb Problems, in *Newcomb's Problem*, ed. Arif Ahmed, Cambridge: Cambridge University Press: 138–59.

Laslier, Jean-François 1997. *Tournament Solutions and Majority Voting*. Berlin: Springer-Verlag Berlin Heidelberg.

Lewis, David 1981. Causal Decision Theory, *Australasian Journal of Philosophy* 59/1: 5–30.

Podgorski, Abelard forthcoming. Tournament Decision Theory, *Noûs*.

Richter, Reed 1984. Rationality Revisited, *Australasian Journal of Philosophy* 62/4: 392–403.

Skyrms, Brian 1990. *The Dynamics of Rational Deliberation*. Cambridge, MA: Harvard University Press.

Spencer, Jack 2021. Rational Monism and Rational Pluralism, *Philosophical Studies* 178/6: 1769–1800.

Spencer, Jack and Ian Wells 2019. Why Take Both Boxes?, *Philosophy and Phenomenological Research* 99/1: 27–48.

Wedgwood, Ralph 2013. Gandalf's Solution to the Newcomb Problem, *Synthese* 190/14: 2643–75.

Weirich, Paul 1985. Decision Instability, *Australasian Journal of Philosophy* 63/4: 465–72.

Weirich, Paul 1988. Hierarchical Maximization of Two Kinds of Expected Utility, *Philosophy of Science* 55/4: 560–82.

Weirich, Paul 2004. *Realistic Decision Theory: Rules for Nonideal Agents in Nonideal Circumstances*. New York, NY: Oxford University Press.