Can It Be Irrational to Knowingly Choose the Best?

Words: 4,050

Words, including footnotes: 4,802

There is a burgeoning research program born of partial satisfaction with causal decision theory (CDT). Those pursuing the program think that much is right about CDT, including its handling of Newcomb's Problem. But they think that CDT mishandles certain unstable decisions, such as Egan's Psychopath Button.[1] The aim is thus to find some successor to CDT: some decision theory that preserves what is right about CDT, while better handling unstable decisions.[2]

[1] Andy Egan, "Some Counterexamples to Causal Decision Theory," *Philosophical Review*, CXVI, 1 (2007); 93-114. Other unstable decisions include Death in Damascus, *cf.* Allan Gibbard and William L. Harper, "Counterfactuals and Two Kinds of Expected Utility," in C. A. Hooker, J. J. Leach, and E. F. McClennan, eds., *Foundations and Applications of Decision Theory* (Boston: D. Reidel, 1978): 125-62, and Dicing with Death, *cf.* Arif Ahmed, "Dicing with Death," *Analysis*, LXXIV, 4, (2014): 587-92.

[2] Some contributions to this research program include: Johan Gustafsson. "A Note in Defense of Ratificationism." *Erkenntnis*, LXXV, 1 (2011): 147-50; William MacAskill, "Smokers, Psychos, and Decision-Theoretic Uncertainty," this JOURNAL, CXIII, 9, (2016): 425-45; Ralph Wedgwood, "Gandalf's Solution to the Newcomb Problem." *Synthese*, CXC, 14 (2013): 2643-75; Paul Weirich, "Decision Instability," *Australasian Journal of Philosophy*, LXIII, 4 (1985):465-72; Paul Weirich, "Hierarchical Maximization of Two Kinds of Expected Utility," *Philosophy of Science*, LV, 4 (1988): 560-82; and [redacted].

The immediate concern of this paper is two recently proposed successors: a view put forward in this JOURNAL by J. Dmitri Gallow (GDT), and a distinct, but structurally similarly view put forward by David James Barnett (BDT).[3] I share the sentiments that motivate these proposals. Like Gallow and Barnett, I think that we should be looking for a decision theory that preserves what is right about CDT, while better handling unstable decisions. But I think that we should reject both GDT and BDT.

The argumentative bedrock of this paper 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*.

In fact, 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*.

[3] J. Dmitri Gallow, "The Causal Decision Theorist's Guide to Managing the News," this JOURNAL CXVII, 3, (2020): 117-49, and David James Barnett, "Graded Ratifiability," (unpublished manuscript).

But I have two main argumentative aims in this paper, and neither requires the added strength of Strengthened Knowingly.

The first aim is negative and mentioned already. I think that any decision theory that conflicts with Knowingly is false. It is a familiar fact that evidential decision theory (EDT) conflicts with Knowingly. I will argue that GDT and BDT also conflict with Knowingly and should, like EDT, be rejected therefore.

The second aim is positive. It is tempting to think that rationality is *prediction-insensitive*: that whether an option is rationally permissible never depends on how the agent divides their credence among their options. As Caspar Hare and Brian Hedden say, en route to arguing that rationality is prediction-insensitive:

> [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 odder it would sound for me to say "I believe that I will do this, so I ought not to do this" […].[4]

But there is a compelling argument that rationality is prediction-sensitive if Knowingly is true, and I accept Knowingly. The positive aim is thus to argue that rationality is prediction-sensitive. Many decision theories entail (as EDT, GDT, and BDT do) that rationality is prediction-insensitive. So this positive thesis, if true, has far-reaching consequences.

[4] Caspar Hare and Brian Hedden, "Self-Reinforcing and Self-Frustrating Decisions," *Noûs* L, 3, (2016): 604-28, p. 604.

1/ EDT

Let me start with Knowingly and EDT. Let A be the set of *options*;[5] let C be the agent's *credence function*; let u be the agent's *utility function*; let W be the set of *possible worlds*;[6] and let K be the set of *dependency hypotheses*.[7] The V-value of some option a, $V(a)$, is then $\Sigma_W C(w|a)u(w) = \Sigma_K C(w|k)V(ak)$.

According to EDT, agents are always rationally required to maximize V. EDT thus recommends one-boxing in Newcomb's Problem.[8]

In Newcomb's Problem, there is an opaque box and a transparent box. The agent has two options; they can take only the opaque box (one-box) or both boxes (two-box). The transparent box contains $1,000. What the opaque box contains depends on a prediction made yesterday by a very reliable predictor. If the predictor predicted that the agent would one-box, the opaque box contains $1,000,000. If the predictor predicted that the agent would two-box, the opaque box contains $0. The agent knows all of this.

---

[5] Construed as propositions the agent can make true by choosing.

[6] Assumed, for simplicity, to be finite.

[7] A dependency hypothesis is a maximal description of how things the agent cares about do and do not depend causally on the agent's choice; *cf.* David Lewis, "Causal Decision Theory," *Australasian Journal of Philosophy* LIX, 1, (1981): 5-30, p. 11.

[8] There is some disagreement about whether EDT recommends one-boxing; see *e.g.* Ellery Eells, *Rational Decision and Causality* (New York Cambridge University Press, 1982). But many contemporary proponents of EDT regard the recommendation of one-boxing as an important advantage of their view, *cf.* Arif Ahmed, *Evidence, Decision and Causality* (New York: Cambridge University Press, 2014). So I will continue to assume that EDT recommends one-boxing.

If we equate dollars and units of value, then the *V*-value of one-boxing is close to 1,000,000, and the *V*-value of two-boxing is close to 1,000.[9] So EDT recommends one-boxing.

But this recommendation conflicts with 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. If we assume that the agent cares only about money and values dollars linearly, then 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 taken together. An agent facing Newcomb's Problem 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 the agent knows that the two boxes together contains $1,000 more than does the opaque box alone.[10]

[9] Either the opaque box contains $0 ($k_z$) or $1,000,000 ($k_m$). $V(\mathrm{a}_{1\mathrm{BOX}}) = C(k_z|a_{1\mathrm{BOX}})(0) + C(k_m|a_{1\mathrm{BOX}})(1{,}000{,}000) \approx 1{,}000{,}000$. $V(a_{2\mathrm{BOX}}) = C(k_z|a_{2\mathrm{BOX}})(1{,}000) + \mathrm{C}(k_m|a_{2\mathrm{BOX}})(1{,}001{,}000) \approx 1{,}000$.

[10] *Cf.* [redacted].

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.[11]

2/ Maxrat

I will not present either GDT or BDT in full. Both are subtle and sophisticated, and there are interesting differences between them. But they agree about how to handle two-option decisions, and my criticism does not require looking at decisions with more than two options.

Both emphasize graded ratifiability, a relational property of options. The U-value of option a, $U(a)$, is $\Sigma_K C(k)V(ak)$; and for any option b, the b-conditional U-value of a, $U(a|b)$, is $\Sigma_K C(k|b)V(ak)$. Graded ratifiability is function of conditional U-values. The *graded ratifiability* of a, relative to b, is $U(a|a) - U(b|a)$.

There is no such thing as the graded ratifiability of an option if there are more than two options. Graded ratifiability is a relational measure of regret/gladness. If negative, the graded ratifiability of a, relative to b, is the degree to which the agent will regret having chosen a instead of b; if positive, the graded ratifiability of a, relative to b, is the degree to

[11] Strictly speaking, the conflict is between EDT and Knowingly+, the conjunction of Knowingly and the claim that options have objective values. There is reason to think that EDT'ists should deny that options have objective vales; see [redacted]. But anyone looking for a successor to CDT should accept that options have objective values, so I ignore the distinction between Knowingly and Knowingly+, henceforward.

which the agent will be glad to have chosen a instead of b. But if there are only two options, we can ignore the relationality, letting $R(a) = U(a|a) - U(b|a)$ and $R(b) = U(b|b) - U(a|b)$ be the graded ratifiability of a and b, respectively.

According to both GDT and BDT,

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

I will argue against GDT and BDT by arguing against Maxrat.

At first blush, Maxrat appears to deliver what those seeking a successor to CDT seek. It recommends two-boxing in Newcomb's Problem,[12] but avoids some of the unintuitive recommendations that CDT makes regarding certain unstable decisions.

Consider Egan's Psychopath Button, for example. The agent has two options; they can press the button, thereby killing all psychopaths, or refrain. The agent has a medium-strong desire to rid the world of psychopaths and a strong desire not to die. The agent is also very confident that only a psychopath would press.

According to CDT, agents are rationally required to maximize U-value. So, according to CDT, an agent who is sufficiently confident that they are not a psychopath is rationally required to press.

But that recommendation seems wrong. There is a powerful intuition that an agent facing Egan's Psychopath Button is rationally required to refrain, irrespective of their

[12] See Gallow *ibid.*, pp. 131-2.

credence that they are, themself, a psychopath, and it is a credit to Maxrat that it accommodates this powerful intuition.[13]

Moreover, as Gallow and Barnett show, although Maxrat applies only to two-option decisions, one can use graded ratifiability to devise ways of choosing between more than two options, and the theory that results from combining Maxrat to these ways of choosing between more than two options appears to outperform CDT with regard to multi-option unstable decisions, too.[14]

Now, of course, none of this amounts to a derivation of Maxrat. It is just a dialectical motivation, and the dialectical motivation will not appeal to everyone. Not everyone is a two-boxer, and some two-boxers tolerate the recommendations CDT makes in unstable decisions.[15] But I am, myself, in the target demographic: a two-boxer who thinks CDT mishandles unstable decisions. So I think that Maxrat is very much worth considering in earnest.

Nevertheless, I reject Maxrat; for Maxrat conflicts with Knowingly.

3/ An Argument Against Maxrat

[13] See Barnett *ibid.*, §3.

[14] See Gallow *ibid.*, pp. 139-46.

[15] See *e.g.* James. M. Joyce, "Deliberation and Stability in Newcomb Problems and Pseudo-Newcomb Problems," in Arif Ahmed, ed., *Newcomb's Problem* (New York: Cambridge University Press, 2018), pp. 138-59.

Egan's Psychopath Button is an asymmetric self-frustrating decision. My argument against Maxrat begins from the following asymmetric self-reinforcing decision:

> *Asymmetry*. There are two opaque boxes, *a* and *b*. The agent has two options; they can take either box. What the boxes contains depends on a prediction made yesterday by a very reliable 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.

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*: $R(b) \approx 15 > R(a) \approx 10$.[16]

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*.

[16] $R(b) = U(b|b) - U(a|b) = \Sigma_K C(k|b)V(bk) - \Sigma_K C(k|b)V(ak) \approx (0)(0) + (1)(15) - ((0)(10) + (1)(0)) \approx 15$, and $R(a) = U(a|a) - U(b|a) = \Sigma_K C(k|a)V(ak) - \Sigma_K C(k|a)V(bk) \approx (1)(10) + (0)(0) - ((1)(0) + (0)(15)) \approx 10$.

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.

4/ Four Possible Reponses

To test the strength of this argument, let me consider four ways a proponent of Maxrat might respond.

4.1. 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 *ideally rational* 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 will not do. The utilities of the agent are coherent and well-behaved. The credences of the agent satisfy

the probability axioms. The violated principle will almost certainly 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.[17] 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 maximizes 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 the challenge of 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 is imposed. Two-boxers want a decision theory that entails that it is

[17]See Joyce *ibid*, who draws on Brian Skyrms, *The Dynamics of Rational Deliberation* (Cambridge, MA: Harvard University Press, 1990).

irrational to one-box in Newcomb's Problem, even if the agent is confident that they will one-box. Refrainers want a decision theory that entails that it is irrational to press, even if the agent is confident that they will press. Maxrat is marketed to two-boxing refrainers. 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 maximizes graded ratifiability, then Maxrat no longer delivers what two-boxers and refrainers want; 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 dialectical 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 ideally rational, 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.

4.2. Restricting Maxrat. One can reconcile Maxrat, Knowingly, and Known, without proposing a new principle of ideal rationality, just be 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 marketing problem. Two-boxers want a decision theory that entails that one-boxing is irrational, even if the agent foreknows that they will one-box. Refrainers want a decision theory that entails that pressing is irrational, even if the agent foreknows that they will press. If Maxrat is restricted to cases in which an agent does not foreknow

what they will choose, then Maxrat does not deliver what two-boxers and refrainers want, and we are left wondering whether any 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.

> **Supervenience**. If an agent facing an *a*-veridical Asymmetry is rationally permitted/required to choose *a*, then an agent facing an *a*-confident Asymmetry is rationally permitted/required to choose *a*.

The foreknowledge restriction is thus not restrictive enough. To insulate Maxrat from the threat posed by Knowingly and Known, one needs to ensures that Maxrat makes no prediction concerning any *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. (It is not for no reason that two-boxers and refrainers want a decision theory that applies equally to agents who foreknow what they will choose.)

4.3. 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 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 can) know both that they will choose *a* and that *a* is strictly better than *b*.

4.4. 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 the best option available to me before I chose it." And that, as a reply to alleged irrationality, seems, to my mind, dispositive. Irrational decision-making is defective and criticizable. But what could be defective or criticizable about an agent knowingly choosing their best option?[18]

As I said above, I accept Strengthened Knowingly. I think that two-boxing in Newcomb's Problem is not just rationally permitted, but rationally required, and I think that Strengthened Knowingly explains why two-boxing is rationally required. Similarly, I

[18] *Objection*: Knowingly fails when extremes 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, nevertheless. *Reply:* I suspect that the salience of possible extremes destroys one's knowledge in these cases. Indeed, I suspect that one cannot know that *a* is strictly better than *b* if $U(a) < U(b)$. But, that aside, it would suffice for the current purposes if Knowingly held for cases that did not involve extreme possibilities.

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.

5/ Two Takeaways

Both GDT and BDT entail Maxrat, so an argument against Maxrat is interesting in its own right. But two more general takeaways can be wrung from the discussion heretofore.

5.1. Prediction-sensitivity. The first concerns prediction-sensitivity. Rationality is prediction-sensitive if whether an option is rationally permissible can depend on how the agent divides their credence among their options. More formally: Say that a decision $d^* = \langle A^*,K^*,C^*,u^* \rangle$ is *predictively accessible* from a decision $d = \langle A,K,C,u \rangle$ just if $A^* = A$, $K^* = K$, $u^* = u$, and C^* can be obtained from C by Jeffrey conditionalizing over A. Then rationality is prediction-sensitive just if the permissible options relative to some decision differ from the permissible options relative to some predictively accessible decision.

One decision that is predictively accessible from an *a*-confident Asymmetry is a *b-confident Asymmetry*, in which the agent is highly confident that they will choose *b*. All of the decision theories considered herein—CDT, GDT, BDT, and EDT—agree that an agent facing a *b*-confident Asymmetry is rationally required to choose *b*, and for good reason. An agent facing a *b*-confident Asymmetry is very confident that choosing *a* will lead to the

worst outcome ($0) and that choosing *b* will lead to the best outcome ($15). But if an agent facing a *b*-confident Asymmetry is rationally required to choose *b*, and an agent facing an *a*-confident Asymmetry is rationally permitted to choose *a*—as Knowingly, Known, and Supervenience entail—then rationality is prediction-sensitive.

It has been claimed that prediction-sensitivity is a structural defect: that any decision theory that entails that rationality is prediction-sensitive should be rejected on those grounds.[19] But, in fact, the reverse is true. Any decision theory that entails that rationality is prediction-insensitive is, for purely structural reasons, inadequate.

5.2. Decision Instability. The second takeaway concerns decision instability. A successor to CDT is meant to solve the distinctive problem that unstable decisions pose, so it is natural to demand that any proposed successor agree with CDT regarding every stable decision. But this demand prompts a question. What marks a decision as stable? Two proposals suggest themselves.

Say that a decision $d = \langle A,K,C,u \rangle$ is *U-insensitive* if the options that maximize *U*-value relative to it also maximize *U*-value relative to every decision predictively accessible from it. The first proposal takes stability to be *U*-insensitivity.

Say that a decision $d = \langle A,K,C,u \rangle$ is *U-stable* if some option *a* maximizes both *U*-value and *a*-conditional *U*-value relative to it. The second proposal takes stability to be *U*-stability.

[19] See *e.g.* Hare and Hedden *ibid.* and Reed Richter, "Rationality Revisited" *Australasian Journal of Philosophy* LXII, 4 (1984): 392-403.

The two proposals agree about Egan's Psychopath Button and the other paradigm unstable decisions; for the paradigm unstable decisions are *U*-unstable, and every *U*-unstable decision is *U*-sensitive. But the proposals disagree about Asymmetry and other decisions that are *U*-sensitive but *U*-stable.

Maxrat is a fitting solution to the problem of decision instability if stability is *U*-insensitivity—Maxrat agrees with CDT regarding every *U*-insensitive decision. But Maxrat is an ill-fitting solution if stability is *U*-stability—Maxrat does not agree with CDT regarding every *U*-stable decision, as witnessed by an *a*-confident Asymmetry. And, for two reasons, I think stability should be identified with *U*-stability.

The first is extensional. I think that CDT handles Asymmetry and other *U*-sensitive but *U*-stable decisions correctly. It is not clear that any *U*-stable decision poses the distinctive threat to CDT that unstable decisions do.

The second reason is metaethical. The burgeoning research program—the search for a successor to CDT—is only as philosophically significant as the division between stable and unstable decisions. It is not clear that the division between *U*-insensitive and *U*-sensitive decisions marks anything of philosophical significance. But the division between *U*-stable and *U*-unstable decisions is metaethically deep and important. When an agent faces a *U*-stable decision, the fact that an option maximizes *U* can be their reason for choosing the option. The agent can know both that they will choose the option and that choosing the option maximizes *U*. When an agent faces a *U*-unstable decision, the fact that an option maximizes *U* cannot be their reason for choosing the option. The agent cannot

know both that they will choose the option and that the option maximizes *U*.[20] Identifying stability with *U*-stability thus makes sense of our intuitions. It makes sense that CDT would handle stable decisions correctly, since those are the decisions in which *U*-maximization can be the agent's reason for choosing an option, and it makes sense that unstable decisions would pose a distinctive threat to CDT, since those are the decisions in which *U*-maximization cannot be the agent's reason for choosing an option.[21]

If we accept the identity between stability and *U*-stability, we can state more clearly what we want from a successor to CDT. A successor should agree with CDT regarding every *U*-stable decision, while better handling *U*-unstable decisions. And we can also state more clearly why Maxrat fails. No decision theory that entails Maxrat can be the successor we seek; for no decision theory that entails Maxrat agrees with CDT regarding every *U*-stable decision.[22]

---

[20] This point about reasons explains why, though 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 knows both that they will choose *a* and that *a* is their best option, then the fact that *a* is their best option can be their reason for choosing *a*. But if the agent's knowledge that *a* is their best option would be destroyed by coming to believe that they will choose *a*, then the fact that *a* is their best option cannot be their reason for choosing *a*, even though they know that *a* is their best options. And I think that some such cases are counterexamples to Knowing; see [redacted].

[21] See [redacted].

[22] [Acknowledgements].