Doing Your Best (While Making Do with Less): The Actual Value Conception of Instrumental Rationality

by

Ryan Doody

Submitted to the Department of Linguistics and Philosophy
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

August 2016 [September 2016]

Author . Signature redacted
Department of Linguistics and Philosophy
August 29, 2016

Certified by . . . Signature redacted
Agustín Rayo
Professor
Thesis Supervisor

Accepted by . . . . Signature redacted
Roger White
Chairman, Department Committee on Graduate Theses

Doing Your Best (While Making Do with Less): The Actual Value Conception of Instrumental Rationality

by
Ryan Doody

Submitted to the Department of Linguistics and Philosophy
on August 29, 2016, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

In this thesis, I sketch a decision-theoretic picture of instrumental rationality, which I call the **Actual Value Conception**: roughly, that you should align your preferences over your options to your best *estimates* of how the *actual values* of those options *compare*. Less roughly: for any options, ϕ and ψ, you are instrumentally rational if and only if you prefer ϕ to ψ when, and only when, your best estimate of the extent to which ϕ's actual value exceeds ψ's is greater than your best estimate of the extent to which ψ's actual value exceeds ϕ's, where an option's *actual value* equals the value you assign to the outcome that would *actually* result from performing it.

In the first chapter, I argue that this picture underlies causal decision theory by showing that, given some assumptions, the two are equivalent, and that the picture unifies and underlies the intuitive arguments offered for Two-Boxing over One-Boxing in the Newcomb Problem. I also show that the picture is incompatible with evidential decision theory. Evidential decision theory sometimes recommends preferring one option to another even though you are certain that the actual value of the latter exceeds the actual value of the former.

In the second chapter, I develop a decision theory for agents with incomplete preferences — called **Actual Value Decision Theory** — that, unlike its more popular competitors, is consistent with, and motivated by, the picture of instrumental rationality sketched in the first chapter. I argue that, in addition to being a generalization of causal decision theory, **Actual Value Decision Theory** is supported by many of the same considerations.

In the final chapter, I consider two powerful arguments against **Actual Value Decision Theory** — the Most Reason Argument and the Agglomeration Argument — and I argue that, while neither proves to be fatal, they each bring to light some interesting consequences of taking the **Actual Value Conception** seriously. In particular: that, first, we should reject the idea that instrumental rationality consists in doing what you *have* the most reason to do; and, second, that sometimes it is rationally permissible to have non-transitive (but not *cyclic*) instrumental preferences.

Thesis Supervisor: Agustín Rayo
Title: Professor

Acknowledgments

I owe thanks to more people than I can list, but I will try my best.

Foremost, I would like to thank my committee — Agustín Rayo, Caspar Hare, and Stephen Yablo — for their feedback and their advice, for their kindness and support, for their attention, and for their seemingly-endless, bordering-on-saintly patience. Writing a thesis — and philosophy, in general — is difficult. I don't know what I would've done without them. They each went *well beyond* what could be asked of any reasonably sane member of a thesis committee. And for that — and for everything — I cannot thank them enough.

The thesis grew out of some earlier work, which benefited greatly from discussions with very many people. As a philosopher, I, too, have benefited from discussions with, and support from, very many people. In particular, I'd like to thank: Arden Ali, Jennifer Carr, Dylan Bianchi, Colin Chamberlain, Nilanjan Das, Tom Dougherty, Lyndal Grant, David Gray Grant, Jordan Gray, Daniel Greco, Daniel Hagen, Sally Haslanger, Brian Hedden, Richard Holton, Sophie Horowitz, Brendan de Kenessey, Justin Khoo, Cole Leahy, Matt Mandelkern, Jack Marley-Payne, Vann McGee, Sofia Ortiz-Hinojosa, Milo Phillips-Brown, Kevin Richardson, Susanna Rinard, Damien Rochford, Bernhard Salow, Said Saillant, Miriam Schoenfield, Melissa Schumacher, Brad Skow, Jack Spencer, Bob Stalnaker, Ian Wells, Roger White, and — seriously — all the very many others that I've almost certainly inconsiderately neglected to mention.

I'd also like to thank the audiences of all the various MIT workshops and reading groups — the M.A.T.T.I.s and DOODYs and WIPs and ERGs — at which I've had the pleasure of presenting my work. Thank you, also, to the audience of the 2015 ASU Graduate Philosophy Conference at which I presented an early, proto-version of what would become the second chapter of my dissertation. In particular, thank you to Rebecca Chan for providing excellent comments.

Finally, I would like to thank my friends and family for their support, emotional and otherwise. To my parents, Dennis and Colleen. To my sister, Shannon. To my aunts and uncles, and my grandparents, and cousins. To my friends. My gratitude is unending. I don't have the words to adequately express my appreciation for all that you've done for me. All I can say is: thank you.

THIS PAGE INTENTIONALLY LEFT BLANK

Contents

Chapter 1

The Actual Value Conception of Instrumental Rationality

1.1 Introduction

Instrumental rationality is about taking the best means to your ends. Suppose you want to visit your grandmother as quickly as possible. And suppose there are only two things you can do: you can *walk* to Grandma's, or you can *drive*. If driving will get you to Grandma's faster than walking will, *driving* is the best means to your ends. Are you instrumentally irrational if you walk instead? Not necessarily. Perhaps you (incorrectly, but reasonably) *believe* that walking will be faster. Instrumental rationality doesn't require you to take the means to your ends that are *actually* best. If you know that ϕing is the best means to your ends, and you don't take it, then you've done something instrumentally irrational.

What does instrumental rationality require of you when you don't know which of those options available to you is the best means to your ends?

This chapter will look at some decision-theoretic answers to this question. I will outline a picture of rational decision-making, called the **Actual Value Conception**, which says, roughly, that the ranking of your means should, insofar as you're instrumentally rational, correspond to your *best estimates* of how your means *compare* in terms of their *actual values.* I will then argue that the **Actual Value Conception** underlies both causal and benchmark decision theory, and is incompatible with evidential decision theory.

1.2 Decision Theories & Instrumental Rationality

Normative Decision Theory provides an account of instrumental rationality. Given your perspective — that is, how you take the world to be — and your ends, Decision Theory tells you what you rationally ought to do. Instrumental rationality is about taking the best means to your ends, and decision theories provide mathematically rigorous accounts of what makes some means better than others. Decision theories have two parts: an *axiological* part and a *deontological* part. The former is about how to evaluate your available options — whether you rationally should prefer, or disprefer, an option ϕ to an option ψ, for example — given your ends and your perspective. The latter is about which of your available options you rationally ought to take, given the evaluations of those options. In other words, the theory's axiology specifies how the non-instrumental value of your ends, your perspective, and the instrumental value of your means should all relate; and the theory's deontology specifies what you rationally should choose to do, given the instrumental value of your means.

Here's an example. *Expected Utility Theory* is a decision theory that says you rationally ought to maximize expected utility.[1] It represents your *perspective* by attributing to you a probabilistically coherent credence-function Cr, and it represents your *ends* with a utility-function defined over the possible outcomes of your decisions.[2] Given the facts about how you non-instrumentally

[1] It's perhaps best to not think of Expected Utility Theory as a decision theory *itself,* but rather as a *genus* of decision theories containing, for example, causal decision theory, evidential decision theory, and others.

[2] When your preferences are incomplete, it's unclear what your "ends" are. In the standard case, when your preferences *are* complete, by representing you with a utility-function, we treat you *as if* you have a single unified end — there is a single measurable quantity of value that we represent you as seeking to maximize. Of course, utility is not some precious fluid that Decision Theorist presuppose we all want to amass; rather, utility is a theoretical posit whose extreme flexibility allows it to represent the intrinsic valuing of anything, whatever it happens to be: money, happiness, pleasure, other people's happiness, other people's pain, jumping jacks, etc., or even things that cannot be easily expressed by finitely long strings of English. That being said, by assuming that your

value your ends — which are encoded in your utility-function defined over outcomes — and your credence function, Expected Utility Theory says how you rationally ought to instrumentally value your means: namely, the instrumental value of an option is its *expected* utility. What you rationally ought to do, according to Expected Utility Theory, is take the available means that is most instrumentally valuable to you: that is, you should *maximize* expected utility.

Decision theories, like Expected Utility Theory, do not merely provide roundabout, mathematically precise ways of cashing out the idea that instrumental rationality requires taking the best means to your ends; rather, they provide substantive accounts of what you rationally ought to do when you don't know which of your options, if any, is *actually* the best means to your ends. Different decision theories will have different things to say about the requirements of instrumental rationality, even if they agree that instrumental rationality is about taking the best means your ends. As an account of instrumental rationality, what's substantive about a decision theory is what it has to say about decision problems under risk or uncertainty (that is, when which outcome an option will bring about depends on features of the world you are uncertain about). Decision problems can be represented with three different entities: there are your *options* (or "alternatives," or "acts"), which are the objects of your instrumental preferences; there are the *outcomes* that might result from performing your options, which are the objects of your *non*-instrumental preferences; and there are *states*, which are those features of the world not under your control that influence the outcomes that might result from performing one of your options.[3] Following Savage [1954], we can think of an option as a *function* from states to outcomes. Or, following Jeffrey [1983] and others, we can think of all three of these entities as *propo-*

various ends — in all their variety and complexity — can be represented with a single utility-function (or, more precisely, a set of utility-functions all of which are positive linear transformations of each other), we thereby assume that all the potential tensions between your ends are (or, would be were you to consider them) resolved. And, in so doing, we treat you as if you have a single overarching end.

[3] See [Briggs, 2014] or [Resnik, 1987] as examples of explanations of decision theory that set things out in this manner. Also, note that "not under your control" is intentionally ambiguous between a *causal* and an *evidential* reading so as to remain neutral between causal and evidential decision theory. Later, this intentional ambiguity will be disambiguated: we should understand the states (relevant to decision theory) to be *dependency hypotheses,* which are "maximally specific proposition[s] about how the things [you] care about do and do not depend causally on [your] present actions." [Lewis, 1981].

sitions (which I'll take to be sets of possible worlds), where an option ϕ is a proposition of the form ⌜I do such-and-such⌝, a state S is a proposition concerning how (for all you know) the world might be, and the outcomes are propositions of the form $(\phi \wedge S)$.

Your options (or "means") have instrumental value. According to Expected Utility Theory, for example, the instrumental value of an option is its expected utility. But your options also have, what I will call, *actual value.* The *actual value* of an option is equal to the value you assign to the outcome that would, as a matter of fact, result from performing it.[4]

> **Actual Value.** Let $K_{@}$ pick out the state of affairs that actually obtains.[5] It specifies how things are with respect to all of the features of the world (that you care about) which are outside your present influence.
>
> $$V_{@}(\phi) = V(\phi \wedge K_{@})$$
>
> The actual value of ϕ-ing is equal to the value you assign to the outcome picked out by $(\phi \wedge K_{@})$, which is the outcome that would actually result were you to ϕ.[6]

"Take the best means to your ends" is ambiguous. It could mean "take the option with the highest *actual* value," or it could mean "take the option with highest *instrumental* value" (where an option's instrumental value is given by a decision theory's axiology). If you happen to know, for each of your

[4] I define actual value in terms of dependency hypotheses, but we could just as well define it in terms of non-backtracking, causally-understood, subjunctive conditionals. Let $X \,\Box\!\!\rightarrow S$ be such a conditional. (It says: *if it were true that X, then it would be true that S.*) Then, we can say: if $\phi \,\Box\!\!\rightarrow o$, then $V_{@}(\phi) = V(o)$. In words: if, were you to ϕ, doing so would result in outcome o, then the actual value of ϕ-ing is equal to the value you assign to outcome o. In some recent work, Spencer and Wells [2016] cash out actual value these terms.

[5] Understand $K_{@}$ to be a *dependency hypothesis:* a maximally specific proposition about how the things you care about depend causally on your options [Lewis, 1981]. The dependency hypotheses form a partition, and each dependency hypothesis is causally independent of your options.

[6] It's important that $K_{@}$ be a dependency hypothesis as opposed to just any state-of-the-world that actually obtains. What you do can affect which state-of-the-world is actual — studying will make it more likely that you'll pass the test, for example — and, so, the outcome that would result were you to perform one of your options isn't guaranteed to be the outcome that option has in the state-of-affairs that is actual unless, like dependency hypotheses, the states are causally independent of your options.

available options, its actual value — that is, if you are facing a decision problem under *certainty* — then the two disambiguations aren't in conflict: each option's actual value should equal its instrumental value. If you don't know the actual values of your options, then, according to (any plausible) decision theory, you should take the option with the highest *instrumental* value. Decision Theory understands instrumental rationality to be both *subjective* and *internalistic:* it's subjective in that its requirements depend on your beliefs and your ends; it's internalistic in that its requirements should supervene on your perspective (so that you are, at least in principle, always in a position to follow its recommendations). Therefore, taking the best means to your ends, according to decision theory, is always a matter of taking the option with the most instrumental value.

There's no guarantee that the option with the most instrumental value is also the option with the most actual value. (In fact, sometimes, you can be certain it's not).[7] The sense in which the option with the most actual value is "the best means to your ends" is clear. It's less clear, however, why the option with the most instrumental value should be considered "the best means to your ends." How are the actual and instrumental values of your options related? And what justifies maximizing the latter when, ultimately, it's the former that you most care about?

[7] *The Drug Example* in [Jackson, 1991], as well as the famous Miners Puzzle (see, for example, [Kolodny and McFarlane, 2010]) are potential illustrations of this. Schematically, these examples have the following form. Suppose you have three options: a safe option, and two risky options. And suppose that you know that the actual value of the safe option exceeds the actual value of one of the risky options and is exceeded by the actual value of the other, but you don't know which is which. It might be best to opt for the safe option over the risky options even though you know the safe option doesn't have maximal actual value.

That being said, it's not straightforwardly obvious that these are cases in which you should be certain that the option with the most instrumental value (i.e., the *safe option*) doesn't maximize actual value. It depends on how we should think about your options. If you are deciding between *take the safe option*, on the one hand, and *reject the safe option*, on the other, then you aren't certain that the actual value of former fails to exceed the actual value of the latter. The actual value of *reject the safe option* depends on which of the two risky options you would choose were you to decide to take the safe option off the table. And so you should be certain that *reject the safe option* has more actual value than *take the safe option* only if you're certain that, were you to reject the safe option, you would choose the risky option with most actual value. But you don't know which of the two risky options has the most actual value, so you have no reason to think that you would select the one that maximizes actual value. And so you have no reason to think that *reject* has more actual value than *take.*

1.3 The Actual Value Conception of Instrumental Rationality

In the previous section, we distinguished between an option's actual value and its instrumental value. Being instrumentally rational involves wanting to take the means to your ends that is actually best. But, when you're uncertain about the actual values of options, you aren't in a position to reliably do so. So instead, decision theories recommend taking the option with the most instrumental value. I will argue that in order for a decision theory to provide an adequate account of the requirements of instrumental rationality, it must characterize instrumental value in such a way so that maximizing it can be justified in terms of your concern for maximizing actual value.

Here's the idea. The regulative ideal governing instrumental rationality is to align your preferences over your options with the facts concerning those options' actual values. Ideally, you would prefer one option to another when, and only when, it *actually* does a better job promoting your ends.

> **The Regulative Ideal of Instrumental Rationality:** "Aim to be such that you strictly prefer one option to another if and only if the actual value of the former exceeds the actual value of the latter; aim to be indifferent between two options if and only if their actual values are equal."

$$\phi \succ \psi \text{ when, and only when, } V_{@}(\phi) > V_{@}(\psi)$$
$$\phi \approx \psi \text{ when, and only when } V_{@}(\phi) = V_{@}(\psi)$$

Preference is, by nature, comparative. So, the facts concerning your options' actual values that should be relevant to satisfying the Regulative Ideal should, likewise, be comparative. Whether your preference for ϕ over ψ conforms to the Regulative Ideal depends on how the actual value of ϕ *compares* to the actual value of ψ — the actual values of ϕ and ψ don't themselves matter *per se*. What matters is whether ϕ has more, or less, or the same amount of actual value as ψ (as well as the *extent* to which ϕ has more, or less, or the same amount of actual value as ψ.) The absolute amount of actual value had by ϕ (and ψ) matter only derivatively: the absolute amounts of actual value *determine* the comparative facts, but it's the comparative facts — not the absolute ones — that matter. For example, suppose that I know the actual value of ϕ but I don't know the actual value of ψ. I'm not in a position to determine whether or not my preference for ϕ over ψ satisfies the Regulative Ideal. On the other hand, suppose that I don't know the actual value of ϕ and I don't know the actual value of ψ, but I do know that, whatever they happen to be, the actual value of ϕ exceeds the actual value of ψ. In such a case, I *am* in a position to know that my preference for ϕ over ψ satisfies the Regulative Ideal.

In order to make clearer the comparative nature of the Regulative Ideal, we can reformulate it in terms of actual value comparisons, using the following measure of comparative value:

$$\mathcal{CV}_{@}\,(\phi, \psi) = V_{@}(\phi) - V_{@}(\psi)$$

More generally, the function $\mathcal{CV}_K(\phi, \psi)$ measures the degree to which option ϕ does better than option ψ in state K — in other words, $\mathcal{CV}_K\,(\phi, \psi) = V(\phi \wedge K) - V(\psi \wedge K)$. And $\mathcal{CV}_{@}\,(\phi, \psi)$ measures the degree to which option ϕ is *actually* better than option ψ.[8] Given that $V_{@}(\phi) > V_{@}(\psi)$ if and only if $\mathcal{CV}_{@}\,(\phi, \psi) > \mathcal{CV}_{@}\,(\psi, \phi)$; and, given that $V_{@}(\phi) = V_{@}(\psi)$ if and only if $\mathcal{CV}_{@}\,(\phi, \psi) = \mathcal{CV}_{@}\,(\psi, \phi)$, we can reformulate the Regulative Ideal in the following, equivalent, way:[9]

[8] Alternatively, we could understand $\mathcal{CV}_{@}$ as *primitive* — $\mathcal{CV}_{@}(\phi, \psi)$ measures the extent to which the actual value of ϕ exceeds the actual value of ψ, it equals $V_{@}(\phi) - V_{@}(\psi)$ when $V_{@}(\phi)$ and $V_{@}(\psi)$ are both well-defined, but it isn't analyzed in terms of them. And, so, for example, $\mathcal{CV}_{@}(\phi, \psi)$ might be well-defined even if $V_{@}(\phi)$ and $V_{@}(\psi)$ aren't. This is particularly relevant to the phenomenon at issue in the next chapter: decisions involving outcomes whose "values" you regard as incommensurable. If you regard some of the potential outcomes of your decisions to be incommensurable, then your preferences over outcomes will fail to be complete. If your preferences are incomplete, we cannot represent your ends with a utility-function; and, so, $V_{@}$ isn't guaranteed to be well-defined. There are other phenomena, in addition to having incomplete preferences, that might make taking the *comparisons* of your options' actual values to be the primitive notion helpful. For example, if you have intransitive preferences, $V_{@}$ isn't well-defined, but $\mathcal{CV}_{@}(\phi, \psi)$ might be. Or, for example, if you regard some possible outcomes (e.g., spending an eternity in heaven) as infinitely valuable, $V_{@}(\phi) - V_{@}(\psi)$ might fail to be well-defined but not $\mathcal{CV}_{@}(\phi, \psi)$. For now, though, we will assume that $\mathcal{CV}_{@}(\phi, \psi) = V_{@}(\phi) - V_{@}(\psi)$.

[9] If we take $\mathcal{CV}_{@}$ as primitive, following the point made in the previous footnote, then this reformulation of the Regulative Ideal is not, strictly speaking, equivalent to the original. They are equivalent only when you can assign well-defined values to the outcomes of your options.

Assuming, as we are in this chapter, that $\mathcal{CV}_{@}(\phi, \psi) = V_{@}(\phi) - V_{@}(\psi)$, then $\mathcal{CV}_{@}(\phi, \psi) >$

> **The Regulative Ideal (ver 2):** "Aim to be such that you strictly prefer one option to another if and only if the actual value of the former exceeds the actual value of the latter; aim to be indifferent between two options if and only if their actual values are equal."
>
> $$\phi \succ \psi \text{ when, and only when } \mathcal{CV}_{@}(\phi, \psi) > \mathcal{CV}_{@}(\psi, \phi)$$
> $$\phi \approx \psi \text{ when, and only when } \mathcal{CV}_{@}(\phi, \psi) = \mathcal{CV}_{@}(\psi, \phi)$$

In order for a decision theory's account of instrumental value to adequately reflect the requirements of instrumental rationality, it must appropriately respect the Regulative Ideal by, somehow, connecting up the facts about your options' instrumental values to the facts about those options' actual values. Roughly, the instrumental value of an option should, in some sense, be a good guide to that option's actual value.

1.3.1 Respecting the Regulative Ideal

The Regulative Ideal says that you should aim, as best you can, to align your preferences over options with the facts concerning the comparisons of those options' actual values. It doesn't say, however, what rationality requires of you when you aren't in a position to know how the actual values of your options compare. Rather, the Regulative Ideal specifies a *criterion of correctness:* your preference for ϕ over ψ is *"correct"* when, and only when, the actual value of ϕ exceeds the actual value of ψ. As an analogy, consider William James' two "commandments as would-be knowers": *Believe truth! Shun error!* [James, 1896].

> **Jamesian Criterion for Belief:** "Aim to believe what's true; aim to disbelieve what's false."
>
> $$\text{Believe } p \quad \text{when, and only when} \quad p \text{ is true.}$$

The Jamesian Criterion is a regulative ideal. You can fail to satisfy the

$\mathcal{CV}_{@}(\psi, \phi)$ just in case $V_{@}(\phi) - V_{@}(\psi) > V_{@}(\psi) - V_{@}(\phi)$, which holds just in case $2 \cdot V_{@}(\phi) > 2 \cdot V_{@}(\psi)$, which holds just in case $V_{@}(\phi) > V_{@}(\psi)$.

ideal without thereby being irrational. Just as your belief that p can be rational even when p is false, your preference for ϕ over ψ can be rational but "incorrect." Or, to make use of one more example, consider the regulative ideal governing criminal trials: punish someone when, and only when, they are guilty of committing the crime.

> **Criterion for Just Punishment:** "Aim to punish someone for committing a crime when, and only when, they are, in fact, guilty of committing the crime."

$$\text{Punish } X \quad \text{when, and only when} \quad X \text{ is guilty.}$$

We might punish, or fail to punish, someone when we shouldn't. But failures of this sort aren't necessarily incidents of injustice. In each case, the criterion specifies conditions of correctness. And, if there were a way to devise a useable procedure that, if followed, guarantees that the criterion is satisfied, we should follow it. But, unless you're omniscient, there will be no such procedure. A criterion for success *itself* is no such procedure because, although it's true that, if followed, success is guaranteed, you won't always (or even usually) be in a position to follow it. If you don't know how things stand with respect to the actual values of your options (or the truth of some propositions, or the guilt of the defendant), then you, likewise, don't know whether your preference for ϕ over ψ (or your belief in p, or punishing the defendant) satisfies the criterion.

The Regulative Ideal of Instrumental Rationality (just like the one governing belief, and just punishment) isn't a procedure that practical rationality (or epistemic rationality, or justice) requires you to follow. Ought implies can. And, because you aren't ideal, the rule *"Prefer those options with greater actual value to those with less"* isn't one that you can reliably follow. Is there a procedure or rule that instrumental rationality requires you to follow? And, if so, what is it? The answer must satisfy two desiderata: First, it must respect the Regulative Ideal of Instrumental Rationality; and, second, it must be operationalizable: you must always be in a position to know what rationality requires of you (and, so, the correct rule should only make reference to material that's immediately accessible to you). In order to satisfy the second desideratum, the facts about the instrumental value of your options should supervene on your perspective. What does it take to satisfy the first desideratum? In order to get clear on the answer, it will be instructive to look at a tempting, but incorrect answer: *respecting the Regulative Ideal requires you to do whatever it is that, by your lights, is most likely to satisfy it.*

> **Rule 1 (preference):** "Prefer option ϕ to option ψ when, and only when, you're more confident that ϕ's actual value exceeds ψ's actual value than that ψ's exceeds ϕ's; be indifferent between ϕ and ψ when, and only when, you're more confident that they have the same actual values than that they don't."

$$\phi \succ \psi \quad \text{when, and only when,} \quad Cr\Big(V_{@}(\phi) > V_{@}(\psi)\Big) > Cr\Big(V_{@}(\phi) \not> V_{@}(\psi)\Big)$$
$$\phi \approx \psi \quad \text{when, and only when,} \quad Cr\Big(V_{@}(\phi) = V_{@}(\psi)\Big) > Cr\Big(V_{@}(\phi) \neq V_{@}(\psi)\Big)$$

> **Rule 1 (belief):** "Believe a proposition p when, and only when, you are more confident that p is true than you are that p is false"

$$\text{Believe } p \quad \text{when, and only when} \quad Cr(p) > Cr(\neg p)$$

> **Rule 1 (just punishment):** "Punish person X for committing a crime when, and only when, you are more confident that X is guilty of committing the crime than you are that X is not guilty of committing the crime."

$$\text{Punish } X \quad \text{when, and only when} \quad Cr\big(X \textit{ is guilty}\big) > Cr\big(X \textit{ is not guilty}\big)$$

These rules satisfy the second desideratum: because your credences are accessible to you, you are always in a position to know how to follow the rule. And, upon first glance, these rules appear to meet the first desideratum: what better way is there to respect a regulative ideal than to do whatever it is that you regard as most likely to satisfy it? But first glances can deceive. It's true that if all you care about is whether or not your preferences (or beliefs, or incidents of punishment) satisfy the criterion set forth by the

Regulative Ideal, then these rules are acceptable.

	$\mathbf{V}_{@}(\phi) > \mathbf{V}_{@}(\psi)$	$\mathbf{V}_{@}(\phi) \not> \mathbf{V}_{@}(\psi)$
$\phi \succ \psi$	Satisfy (1)	Violate (−1)
$\phi \not\succ \psi$	Violate (−1)	Satisfy (1)

	p is true	**p is false**
Believe p	Satisfy (1)	Violate (−1)
$\neg$*Believe p*	Violate (−1)	Satisfy (1)

	Guilty	**Not Guilty**
Punish	Satisfy (1)	Violate (−1)
$\neg$*Punish*	Violate (−1)	Satisfy (1)

Although **Rule 1** outlines the best strategy, by your lights, for satisfying the Regulative Ideal, it fails to appropriately *respect* the Regulative Ideal because it isn't sensitive to the different ways in which the Regulative Ideal might be violated. For example, the Jamesian Criterion can be violated in two ways: when you believe something false, and when you fail to believe something true. Similarly for Just Punishment: we might punish the innocent, or let the guilty go free. And not all violations count equally. We care more about incorrectly punishing the innocent than we do about letting the guilty go free, for example. The criterion, although true, doesn't capture everything that's relevant to the situation. (Of course, omniscient agents — who are always in a position to tell whether someone is guilty or not — have no need to worry about the differences between violating the criterion in different ways, because they are in no danger of violating it in the first place). If it is *much worse* to incorrectly punish the innocent than it is to fail to punish the guilty — which is, presumably, what we do think (e.g., "it is better that ten guilty persons escape than that one innocent suffers") — then, in order to properly respect the Regulative Ideal, one must be a great deal more confident in someone's guilt than **Rule 1 (just punishment)** prescribes.

A similar point holds for **Rule 1 (preference).** Although it's *not* true that it's worse to prefer ϕ to ψ when ϕ's actual value doesn't exceed ψ's than it is to *fail* to prefer ϕ to ψ when it does (or *vice versa*), it *is* true that there are different ways of satisfying and violating the Regulative Ideal that ought to matter. Let's say that if you satisfy the Regulative Ideal of Instrumental Rationality, you are a *winner.* If you're instrumentally rational, you should want to be a winner. (In other words, if you are instrumentally rational, you should want to prefer ϕ to ψ when, and only when, $V_{@}(\phi) > V_{@}(\psi)$).

You don't, however, *merely* care about being a winner. Some wins are more important than others. You care about winning because you care about actual value — you, ideally, want to take the option that has the most of it — and, if you win (by correctly matching your preferences over options to those options' actual values, and then acting on those preferences), then you'll have brought about the greatest amount of actual value that it was within your power to bring about. So, although you do care about winning, you care about it derivatively: you want to bring about the greatest amount of actual value that it's within your power to bring about, and, if you're a winner, then you'll have done just that. What ultimately matters to you, if you're instrumentally rational, is actual value — not "winning" *per se*. And so, in addition to winning, you also care about how much you win by. But **Rule 1 (preference)** isn't sensitive to the fact that some wins are more important to you than others. Here's an example. Suppose that it is very likely that ϕ has more actual value than ψ, but, if it does, only slightly so; and if it doesn't, then it loses badly.

	TICKET #1-#99	TICKET #100
ϕ	\$2	\$1
ψ	\$1	\$700

In this case, you should be very very confident — 99% confident — that ϕ has more actual value than ψ. If all you care about is satisfying the Regulative Ideal, then your best bet would be to prefer ϕ to ψ (just as **Rule 1 (preference)** recommends). But satisfying the ideal — being a "winner" — is not all that should matter to you. Some ways of winning are better than others. In this case, if you follow **Rule 1**'s advice, you have a very large chance at a small win but a small chance at a devastating loss. No matter how great the loss — replace "700" in the problem above with any finite number, no matter how large — **Rule 1 (preference)** will recommend preferring ϕ to ψ. If you don't prefer ϕ to ψ, and (as you should regard as likely) in so doing, you violate the Regulative Ideal, your preferences are "incorrect" but only slightly so; on the other hand, if you prefer ϕ to ψ, and (as you should regard as fairly unlikely) in so doing, you violate the Regulative Ideal, your preferences are *gravely* "incorrect."

So **Rule 1 (preference)** isn't a very good rule. Although it provides a way to approximate the Regulative Ideal, it doesn't properly *respect* it. In order for a rule to properly respect the Regulative Ideal, it should be sensitive to the various possible ways the ideal might be satisfied or violated. Correctly aligning your preferences over ϕ and ψ to the facts concerning their actual values should count for more when the difference between their actual values is great, and it should count for less when the difference is small; and *mutatis mutandis* for when your preferences are *in*correctly aligned.

In order to appropriately respect the Regulative Ideal, then, our rule must be sensitive to more than just how likely you take it to be that ϕ's actual value exceeds ψ's. Because satisfying (or violating) the ideal matters more to you the greater the *differences* in actual value between your options, our rule must also be sensitive to your beliefs about how big or how small the difference, for all you know, might be.

What might such a rule look like? Here's the suggestion. In evaluating the respective merits of options ϕ and ψ, first, use your credences to *estimate* the extent to which the actual value of ϕ exceeds the actual value of ψ, and compare that estimate with your estimate of the extent to which the actual value of ψ exceeds the actual value of ϕ. Ideally (if you were omniscient, for example), you would align your preferences over options with the facts concerning those options' actual values. Because we aren't often in a position to do that, we should, instead, align our preferences over options with our best estimates of the comparisons in actual value of our options. Call this the **Actual Value Conception.** It says that one is instrumentally rational insofar as one's instrumental preferences match, not the comparisons in actual value of one's options *themselves*, but one's best *estimates* of these comparisons. The view holds: (1) that, ultimately, it's the facts concerning your opinions about your options' *actual values* that should ground what instrumental rationality requires of you; and (2) that, in particular, you should aim to bring your subjective evaluations of your options in line with the facts concerning the *comparisons* of their actual values by *estimating.*

> **Estimate Comparisons of Actual Value Rule:** "Prefer option ϕ to option ψ when, and only when, your best estimate of the extent to which ϕ's actual value exceeds ψ's actual value is greater than your best estimate of the extent to which ψ's actual value exceeds ϕ's actual value."

$$\phi \succ \psi \quad \text{when, and only when,} \quad \textsc{Estimate}\Big[\mathcal{CV}_{@}(\phi,\psi)\Big] > \textsc{Estimate}\Big[\mathcal{CV}_{@}(\psi,\phi)\Big]$$

$$\phi \approx \psi \quad \text{when, and only when,} \quad \textsc{Estimate}\Big[\mathcal{CV}_{@}(\phi,\psi)\Big] = \textsc{Estimate}\Big[\mathcal{CV}_{@}(\psi,\phi)\Big]$$

Estimating the comparisons in actual value between your options, and then aligning your preferences with these estimates (in the manner described), is a way of attempting to satisfy the criterion set forth by the Regulative Ideal that is sensitive to what justifies the criterion in the first place: your concern for actual value. Because your estimates are accessible to you, this rule satisfies the second desideratum. And, because your estimates of the extent to which one option's actual value exceeds another's are, by their very nature, sensitive to the different ways the actual values of your options might compare, the rule properly respects the Regulative Ideal and, therefore, satisfies the first desideratum as well.

Comparing your best estimate of $\mathcal{CV}_{@}(\phi,\psi)$ to you best estimate of $\mathcal{CV}_{@}(\psi,\phi)$ is a fairly natural way to approximate the Regulative Ideal given that, as we've just seen, there are better and worse ways of satisfying (or violating) the criterion set forth by the Regulative Ideal.

	$V_\alpha(\phi) > V_\alpha(\psi)$				$V_\alpha(\phi) \not> V_\alpha(\psi)$		
	$\mathcal{CV}_\alpha(\phi, \psi) = v_1$	$\ldots$	$\mathcal{CV}_\alpha(\phi, \psi) = v_n$	$\ldots$	$\mathcal{CV}_\alpha(\psi, \phi) = v_1^*$	$\ldots$	$\mathcal{CV}_\alpha(\psi, \phi) = v_n^*$
$\phi \succ \psi$	v_1	$\ldots$	v_n		$-v_1^*$	$\ldots$	$-v_n^*$
$\phi \not\succ \psi$	$-v_1$	$\ldots$	$-v_n$		v_1^*	$\ldots$	v_n^*

$$\phi \succ \psi? \quad \sum_i Cr\Big(\mathcal{CV}_\alpha(\phi, \psi) = v_i\Big) \cdot v_i - \sum_j Cr\Big(\mathcal{CV}_\alpha(\psi, \phi) = v_j^*\Big) \cdot v_j^*$$

vs

$$\phi \not\succ \psi? \quad \sum_j Cr\Big(\mathcal{CV}_\alpha(\psi, \phi) = v_j^*\Big) \cdot v_j^* - \sum_i Cr\Big(\mathcal{CV}_\alpha(\phi, \psi) = v_i\Big) \cdot v_j$$

> The former is greater than the latter if and only if $\sum_v Cr(\mathcal{CV}_\alpha(\phi, \psi) = v) \cdot v > \sum_{v^*} Cr(\mathcal{CV}_\alpha(\psi, \phi) = v^*) \cdot v^*$. In other words, use your credences to estimate the extent to which the actual value of ϕ exceeds the actual value of ψ, and compare that estimate with your estimate of the extent to which the actual value of ψ exceeds the actual value of ϕ.

Earlier, we distinguished between an option's actual value and an option's instrumental value. A decision theory's axiology provides an account of instrumental value, and a decision theory's deontology advises you to take the option, out of those available to you, with the highest instrumental value. Instrumental rationality is about taking the "best" means to your ends. And the means to your ends that is actually best is the option with the most actual value. Insofar as a decision theory can be said to be providing an account of instrumental rationality, there must be some suitable connection between its account of instrumental value and actual value. There's no reason to think, though, that the option with the most instrumental value is, necessarily, the option with the most actual value. However, according to the **Actual Value Conception**, an option's instrumental value is (or, at least, is very closely related to) your best estimate of that option's actual value. And your best estimate of some quantity is, by definition, the amount that, in some sense, given your uncertainty, you should expect to be "closest" to the truth.

In the next section, I will flesh out the **Actual Value Conception** by motivating the form that one's best estimates ought to take. There are two, related but different, ways of estimating some quantity: in terms of, either, (i) your *unconditional* expectations, or (ii) your *conditional* expectations. I'll focus on the former rather than the latter. In the next section, I show that if we take your best estimates to be given by (i), then the **Actual Value Conception** entails causal decision theory. In the appendix, I show that if we take your best estimates to be given by (ii), then the **Actual Value**

Conception entails Wedgwood [2013]'s benchmark decision theory.[10] In either case, however, the **Actual Value Conception** is incompatible with evidential decision theory.

1.4 Instrumental Value & Actual Value Estimates

Here's the central argument of this chapter.

ACTUAL VALUE ESTIMATE ARGUMENT

P1 **[Actual Value Conception]** You are instrumentally rational if and only if you prefer an option ϕ to an option ψ when, and only when, $\textsc{Estimate}[\mathcal{CV}_{@}(\phi,\psi)] > \textsc{Estimate}[\mathcal{CV}_{@}(\psi,\phi)]$.

P2 Your best estimate of the extent to which ϕ's actual value exceeds ψ's actual value is $\textsc{Estimate}[\mathcal{CV}_{@}(\phi,\psi)] = \sum_v Cr_{\phi}(\mathcal{CV}_{@}(\phi,\psi) = v) \cdot v$.

C You are instrumentally rational if and only if $\phi \succ \psi$ when, and only when, $\sum_v Cr_{\phi}(\mathcal{CV}_{@}(\phi,\psi) = v) \cdot v > \sum_{v^*} Cr_{\psi}(\mathcal{CV}_{@}(\psi,\phi) = v^*) \cdot v^*$.

The first premise of this argument is a statement of the **Actual Value Conception**: being instrumentally rational involves aligning your preferences over options to your best estimates of how the actual values of those options compare. The second premise says that your best estimates of how the actual values of two options compare should be expectational: the best estimate of $\mathcal{CV}_{@}(\phi,\psi)$ should be the weighted average of how much the actual value of ϕ might exceed the actual value of ψ, where the weights correspond to your

[10] Furthermore, I think there are very compelling reasons to think that we should, in general, regard (i)-estimates as *better estimates* than (ii)-estimates. One reason is that (i)-estimates *minimizes expected error,* according to a family of plausible measures of error. Another reason is that the package of (ii)-estimates are *accuracy-dominated* by the package of (i)-estimates, according to that same family of plausible measures of error: no matter how the world turns out to be, the package of (ii)-estimates are less accurate than the package of (i)-estimates. See [Pettigrew, 2015] for a related argument, and a defense of the family of error measures appealed to in these arguments. A fuller discussion of these arguments is outside of the scope of this chapter.

credences in hypotheses about how the actual values of these options might compare. We can distinguish between two different versions of premise **P2**, depending on whether we take the relevant credences in these hypotheses to be your *unconditional* credences or your *conditional* credences.

	$Cr_{\phi}(X) = \,?$
Unconditional:	$Cr(X)$
Conditional:	$Cr(X \mid \phi)$

We will focus on the *unconditional estimate* version of premise **P2**, but I'll present both versions here, and relegate the discussion of the *conditional estimate* version to the appendix.

> [Comparisons of Actual Value Estimates (Unconditional)]
>
> The best estimate of the extent to which the actual value of ϕ exceeds the actual value of ψ:
>
> $$\textsc{Estimate}\Big[\mathcal{CV}_{@}\,(\phi,\psi)\Big] = \sum_{v} Cr\big(\mathcal{CV}_{@}\,(\phi,\psi) = v\big)\cdot v$$

In other words, this version of **P2** says that your best estimate of $\mathcal{CV}_{@}\,(\phi,\psi)$ is the weighted average of how much the actual value of ϕ might exceed the actual value of ψ, where the weights correspond to your current *unconditional* credences in the various hypotheses about how the actual values of ϕ and ψ might compare.

> [Comparisons of Actual Value Estimates (Conditional)]
>
> The best estimate of the extent to which the actual value of ϕ exceeds the actual value of ψ:
>
> $$\textsc{Estimate}\Big[\mathcal{CV}_{@}\,(\phi,\psi)\Big] = \sum_{v} Cr\big(\mathcal{CV}_{@}\,(\phi,\psi) = v \mid \phi\big)\cdot v$$

This version says that your best estimates of $\mathcal{CV}_{@}\,(\phi,\psi)$ is the weighted average of how much the actual value of ϕ might exceed the actual value of

ψ, where the weights correspond, roughly, to the credences you *would have* in the hypotheses about how the actual values of these options might compare *were you to learn that you ϕed.*

I'll show that, on either proposal, the **Actual Value Conception** is incompatible with evidential decision theory. Furthermore, on the first proposal, the **Actual Value Conception** is equivalent to causal decision theory. (In fact, the causal expect utility of an option *just is* your best unconditional estimate of that option's actual value). On the second proposal, the **Actual Value Conception** entails Wedgwood [2013]'s benchmark decision theory.

1.4.1 Unconditional Actual Value Estimates are Causal Expected Utilities

If the **Actual Value Conception** sounds familiar, it should: it's one particular way of spelling out the central idea motivating causal decision theory. One way of describing what causal decision theory says is as follows: when facing a decision, first partition the states-of-the-world into dependency hypotheses (which are maximally specific propositions about how the things you care about depend causally on your options); then, for each of these dependency hypotheses, find the values your option has if that dependency hypothesis is true; the value of your option is the weighted average of these values, where the weights correspond to your credence in each dependency hypothesis being the one that actually holds..[11]

[11] Two quick clarifications. First, I will, following Lewis [1981], use U to denote an option's *causal* expected value and V to denote the evidential expected value of a proposition: $V(X) = \sum_Z Cr(Z \mid X) \cdot V(X \wedge Z)$. The evidential expected value (or "news value") of a proposition measures how good you would expect the actual world to be were you to learn that it's true. Within a dependency hypothesis, a proposition's value is its evidential expected value. Second, there are several other versions of decision theory which don't make reference to dependency hypotheses. Some versions, like [Sobel, 1978] and [Joyce, 1999], define expected value using *imaging.* Other versions, like [Gibbard and Harper, 1978] and [Stalnaker, 1981], appeal to probabilities of subjunctive conditionals. However, as Lewis [1981] convincingly argues, given various plausible assumptions, these other versions of causal decision theory are notational variants of each other. What I say here could just as well, although perhaps less perspicuously, be formulated using one of these other versions.

> [Causal Expected Utility]
>
> The causal expected utility of an option, ϕ, is the weighted average of the values you assign, for each dependency hypothesis, to the outcome that would result from ϕing if that dependency hypothesis is true, and where the weights correspond to your unconditional credences in the dependency hypotheses.
>
> $$U(\phi) = \sum_{K} Cr\,(K) \cdot V(\phi \wedge K)$$

The causal utility of ϕ is greater than the causal utility of ψ if and only if your unconditional estimate of the extent to which the actual value of ϕ exceeds the actual value of ψ is greater than zero. (It's also true that an option's unconditional estimated actual value is equal to its causal expected utility). The **Actual Value Conception,** therefore, underlies causal decision theory.

I'll present only a sketch of the idea here. A fuller statement of the proof can be found in the appendix.

Recall that your unconditional estimate of $\mathcal{CV}_{@}(\phi, \psi)$ is the weighted average of all the ways the actual value of ϕ might exceed the actual value ψ, where the weights correspond to your unconditional credences in the various hypotheses about the ways the actual values of ϕ and ψ might compare:

$$\text{Estimate}\Big[\mathcal{CV}_{@}\,(\phi, \psi)\Big] = \sum_{v} Cr\big(\mathcal{CV}_{@}\,(\phi, \psi) = v\big) \cdot v$$

Given how we've characterized actual value in terms of dependency hypotheses, the proposition that $\mathcal{CV}_{@}\,(\phi, \psi) = v$ is equivalent to the following disjunction of conjunctions:

$$\bigvee_{K_i} \Big(\mathcal{CV}_{K_i}\,(\phi, \psi) = v \,\wedge\, K_i\Big)$$

Because the dependency hypotheses are mutually exclusive, your credence in the hypothesis that $\mathcal{CV}_{@}\,(\phi, \psi) = v$ can be expressed as the sum of your credences in each of the disjuncts.

$$Cr\big(\mathcal{CV}_{@}(\phi,\psi)=v\big)=\sum_K Cr\Big(\mathcal{CV}_K(\phi,\psi)=v\wedge K\Big)$$

And because each dependency hypothesis determines a way that the actual values of your options might compare, your credence in the hypothesis that $\mathcal{CV}_{@}(\phi,\psi)=v$ equals the sum of your credences in those dependency hypotheses according to which, if it is actual, then the actual value of ϕ exceeds the actual value of ψ by the amount v. In other words, for all dependency hypotheses K, if $\mathcal{CV}_K(\phi,\psi)=v$, then your credence that $\mathcal{CV}_{@}(\phi,\psi)=v$ equals the sum of your unconditional credences in each of these Ks. Because each dependency hypothesis corresponds to exactly one hypothesis concerning how the actual values of your options might compare, it follows that:

$$\textsc{Estimate}\Big[\mathcal{CV}_{@}(\phi,\psi)\Big]=\sum_K Cr(K)\cdot\mathcal{CV}_K(\phi,\psi)$$

Therefore, according to the **Actual Value Conception**, you should prefer ϕ to ψ when, and only when, $\sum_K Cr(K)\cdot\mathcal{CV}_K(\phi,\psi)>\sum_K Cr(K)\cdot\mathcal{CV}_K(\psi,\phi)$. And, because, for each K, $\mathcal{CV}_K(\phi,\psi)=V(\phi\wedge K)-V(\psi\wedge K)$, that inequality holds just in case:

$$\sum_K Cr(K)\cdot V(\phi\wedge K)>\sum_K Cr(K)\cdot V(\psi\wedge K)$$

Which is to say: *just in case the causal expected utility of ϕ is greater than the causal expected utility of ψ*. Therefore, the **Actual Value Conception** entails causal decision theory.

1.4.2 Evidential Decision Theory & The Actual Value Conception

Evidential decision theory says that you should prefer one option to another if and only if the expected *evidential* value of the former exceeds that of the latter, where the expected evidential value of an option is, roughly, your estimate of how good the actual world would be were you to learn that you performed that option.

[Evidential Decision Theory]

You should prefer ϕ to ψ if and only if the evidential expected value of ϕ exceeds the evidential expected value of ψ.

$$\phi \succ \psi \iff \sum_Z Cr(Z \mid \phi) \cdot V(\phi \wedge Z) > \sum_Z Cr(Z \mid \psi) \cdot V(\psi \wedge Z)$$

The evidential expected utility of ϕ, $V_E(\phi)$, is equivalent to your conditional estimate of ϕ's actual value — that is, the amount of actual value you would estimate ϕ to have were you to learn that you ϕed.

$$V_E(\phi) = \sum_v Cr\big(V_@(\phi) = v \mid \phi\big) \cdot v$$

But Evidential Decision Theory doesn't conform to the **Actual Value Conception** because it will sometimes recommend preferring one option to another even when you're certain that the former has less actual value than the latter. The Newcomb Problem serves as an example.

> **Newcomb Problem.** You have two boxes before you: an opaque box, which either contains a million dollars or nothing, and a transparent box, which contains a thousand dollars. You have the option to, either, take only the opaque box (One-Boxing) or to take both the opaque and the transparent box (Two-Boxing). Whether the opaque box contains a million dollars or no dollars has been determined by a super-reliable predictor. If the predictor predicted that you'd One-Box, she put a million dollars in the opaque box; if the she predicted that you'd Two-Box, she put nothing in the opaque box.

	Predicts: "One-Box"	Predicts: "Two-Box"
One-Box	$\$M$	$\$0$
Two-Box	$\$M+K$	$\$K$

Assume that you take the predictor to be so reliable that your credence that she predicted correctly is close to one. (And, for simplicity, assume that you value money linearly). Evidential Decision Theory says that you should prefer One-Boxing to Two-Boxing.

$$\begin{aligned} V_E(\text{One-Box}) &= Cr\,(\text{``One-Box''} \mid \text{One-Box}) \cdot V(\$M) + Cr\,(\text{``Two-Box''} \mid \text{One-Box}) \cdot V(\$0) \\ &= .99 \cdot V(\$M) + .01 \cdot V(\$0) \\ &\approx 990,000 \\ V_E(\text{Two-Box}) &= Cr\,(\text{``One-Box''} \mid \text{Two-Box}) \cdot V(\$M+K) + Cr\,(\text{``Two-Box''} \mid \text{Two-Box}) \cdot V(\$K) \\ &= .01 \cdot V(\$M+K) + .99 \cdot V(\$K) \\ &\approx 11,000 \end{aligned}$$

However, you are in a position to be certain that the extent to which the actual value of Two-Boxing exceeds the actual value of One-Boxing is greater than the extent to which the actual value of One-Boxing exceeds the actual value of Two-Boxing — in fact, you should be certain that $\mathcal{CV}_{@}$ (Two-Box, One-Box) $= V(\$K) > 0$, and certain that $\mathcal{CV}_{@}$ (One-Box, Two-Box) $= -V(\$K) < 0$. So, according to the **Actual Value Conception**, you should prefer Two-Boxing to One-Boxing. Whereas, according to evidential decision theory, you should prefer One-Boxing to Two-Boxing.

	Which quantity?	
Which estimate?	$\mathcal{CV}_{@}$	$V_{@}$
Unconditional	CDT	CDT
Conditional	BDT	EDT

1.5 The Arguments for Two-Boxing & the Actual Value Conception

The **Actual Value Conception** recommends Two-Boxing in the Newcomb Problem. There are several intuitive arguments that have been offered in support of Two-Boxing. In this section, we will looks at three of them — the Deference Argument, the Reflection Argument, and the Dominance Argument — and I will argue that the **Actual Value Conception** underlies them all. Each of these arguments can be viewed as a way of dramatizing the fact that, in cases like the Newcomb Problem, you are in a position to be certain about the *comparative* facts concerning your options' actual values.

Three Arguments for Taking Both Boxes

- *The Deference Argument:* Imagine that a friend, who wants the best for you, knows what's in the opaque box. She would want it that you had taken both boxes.
- *The Reflection Argument:* After discovering what's in the opaque box, you will want your(past)self to have taken both boxes.
- *The Dominance Argument:* No matter how the predictor predicted, the outcome of taking both boxes is better than the outcome of taking only the one box.

In order to bring out the connection between the **Actual Value Conception** and these three intuitive arguments, it will be helpful to show that the **Actual Value Conception** entails a general principle relating your beliefs about actual value to the preferences you ought to have if you're instrumentally rational.

1.5.1 The Credence Preference Coherence Principle

Consider the following claim: You rationally shouldn't prefer one option to another if you *are certain* that its actual value doesn't exceed the actual value of the other.

[Credence Preference Coherence]

If you are certain that the actual value of ϕ doesn't exceed the actual value of ψ, then you shouldn't strictly prefer ϕ to ψ.

$$\text{If } Cr\big(V_{@}(\phi) > V_{@}(\psi)\big) = 0, \text{ then } \phi \not\succ \psi$$

The **Actual Value Conception** entails CREDENCE PREFERENCE COHERENCE.[12] According to the **Actual Value Conception,** you should prefer ϕ to ψ if and only if $\sum_K Cr(K) \cdot \mathcal{CV}_K(\phi, \psi) > 0$. In order to show that the principle follows, we'll assume that $Cr\big(V_{@}(\phi) > V_{@}(\psi)\big) = 0$ and, then, show that $\sum_K Cr(K) \cdot \mathcal{CV}_K(\phi, \psi) \not> 0$.

Here's a sketch of the idea (see the appendix for a fuller presentation of the proof). If $Cr\big(V_{@}(\phi) > V_{@}(\psi)\big) = 0$, then, for every dependency hypothesis K, either: (1) the value of ϕ's outcome in K *doesn't* exceed the value of ψ's outcome in K, or (2) you are certain that K is not true (or both). Therefore, $Cr(K) \cdot \mathcal{CV}_K(\phi, \psi) \leq 0$, for every K. And thus, $\sum_K Cr(K) \cdot \mathcal{CV}_K(\phi, \psi) \not> 0$.

Therefore, if $Cr\big(V_{@}(\phi) > V_{@}(\psi)\big) = 0$, then $\sum_K Cr(K) \cdot \mathcal{CV}_K(\phi, \psi) \not> 0$. And, so, according to the **Actual Value Conception,** you should not prefer option ϕ to option ψ.

Evidential Decision Theory, on the other hand, does *not* entail CREDENCE PREFERENCE COHERENCE. The Newcomb Problem, discussed above, serves as a counterexample: your rational credence that the actual value of One-Boxing exceeds the actual value of Two-Boxing should be zero, and yet evidential decision theory recommends One-Boxing. That said, in the special case when the relevant states of the world and your actions are *probabilistically* independent — so that, for each option ϕ, and each state S, $Cr(S \mid \phi) = Cr(S)$ — your beliefs about your options' actual values and the expected evidential value of those options will align in the manner CREDENCE PREFERENCE COHERENCE describes.[13]

[12] Both Causal and Benchmark Decision Theory, as one might suspect, entail CREDENCE PREFERENCE COHERENCE. In the main text, I will show that the version of the **Actual Value Conception** that entails CDT also entails the principle. The proof that Benchmark Decision Theory entails the principle can be found in the Appendix.

[13] Nevertheless, the fact that this doesn't hold in general illustrates that CREDENCE PREFERENCE COHERENCE doesn't capture an idea central to the role formal accounts of practical rationality are meant to play *per se;* rather, it captures an idea central to the role *some* formal accounts of practical rational are meant to play — it captures a central idea behind causal decision theory and benchmark decision theory.

1.5.2 Deference, Reflection, Dominance & Actual Value

The **Actual Value Conception** helps explain the intuitive appeal behind the Deference, Reflection, and Dominance Arguments. Each argument succeeds, if it does, by showing that you are in a position to know that the actual value of taking only the one box doesn't exceed the actual value of taking them both.

In order to motivate this claim, it's helpful to, first, notice that the principles appealed to in these arguments (unless suitably qualified) offer bad advice in certain kinds of cases. Here's a familiar example.[14]

> **The Big Test.** You have an important test tomorrow. You'd very much like to pass the test rather than fail it. Tonight, you have two options: you can *Study* or you can *Goof.* All else equal, you prefer goofing around to studying. What should you do?

	PASS	FAIL
Study	20	0
Goof	25	5

Consider first: **Deference** and **Reflection**. After the results of the test have been made available, future-you (or anyone who is fully-informed, rational, and has your best interests at heart) will want it to be the case that (past)you goofed around rather than studied.[15] (To bring this out more clearly, imagine that future-you — or any fully-informed, rational person who has your best interests at heart — doesn't know whether you chose to study or to goof. If you passed the test, forgetful future-you will hope that you opted to goof; and if you failed the test, forgetful future-you will hope that you opted to goof.) You are now in a position to know that you will not prefer having studied to having goofed. According to **Deference** and **Reflection**, then, you should prefer *Goof* to *Study.* But that can't be right! It's not (always)

[14] For a similar example, see ([Joyce, 1999], pg. 115-8).

[15] As we'll see in a moment, it's not obvious what it means to be "fully-informed" in such situations. I'll argue that it's not enough to know whether or not you passed the test in order to count as fully-informed (at least, for the sense of "fully-informed" relevant to these principles. Instead, one must know which *dependency hypothesis* is actual in order to count as fully-informed. But, in this case, PASS and FAIL are not dependency hypotheses.

irrational to study!

This cases raises, essentially, the same problem for dominance reasoning. Relative to the partition {Pass, Fail}, goofing around dominates studying. Dominance reasoning, then, recommends goofing over studying. And, again, that's bad advice!

What's gone wrong here? These principles give the wrong results in cases, like this one, where the states of the world fail to be *independent* of your actions. And, intuitively, your performance on the test is partially determined by what you will choose to do tonight: if you choose to study, it's more likely you'll pass; if you choose to goof around, it's less likely you'll pass. Furthermore, you'd much rather pass the test than fail it. And so these principles don't provide an acceptable guide to what to do in such cases.

In order to apply these principles so that they do provide acceptable guidance, we need to represent the decision-problem you face in a particular way: we must ensure that the states of the world are independent of your options. The states are independent of your options when your relevant beliefs about them won't change depending on which option you choose to perform. (The states, Pass, Fail, are not independent of your options because you'll assign higher credence to passing the test if you decide to study than you will if you decide to goof around). Let's, then, represent the decision you face in **The Big Test** in the following way, where the states of the world are dependency hypotheses.[16]

	$\overbrace{S \mathrel{\Box\!\!\rightarrow} \textsc{Pass}}^{K_1}$	$\overbrace{S \mathrel{\Box\!\!\rightarrow} \textsc{Fail}}^{K_2}$	$\overbrace{S \mathrel{\Box\!\!\rightarrow} \textsc{Pass}}^{K_3}$	$\overbrace{S \mathrel{\Box\!\!\rightarrow} \textsc{Fail}}^{K_4}$
	$G \mathrel{\Box\!\!\rightarrow} \textsc{Pass}$	$G \mathrel{\Box\!\!\rightarrow} \textsc{Pass}$	$G \mathrel{\Box\!\!\rightarrow} \textsc{Fail}$	$G \mathrel{\Box\!\!\rightarrow} \textsc{Fail}$
Study	20	0	20	0
Goof	25	25	5	5

When the decision-problem is reformulated with dependency hypotheses,

[16] Recall that a *dependency hypothesis* is a maximally specific proposition about how the things you care about depend causally on your options [Lewis, 1981]. The dependency hypotheses form a partition, and each dependency hypothesis is causally independent of your options. Here, we're representing your dependency hypotheses in terms of conjunctions of (causally-understood, non-backtracking) subjunctive conditionals.

it's no longer true that goofing dominates studying. In particular, K_3 — the dependency hypothesis according to which studying results in passing and goofing results in failing — is a state in which studying does better than goofing. (K_3 is also, intuitively, the dependency hypothesis that you should give the most credence to). Furthermore, in order for a rational person with your best interests at heart to count as *truly* fully-informed in this case, she would need to know which of these four dependency hypotheses is the actual one. Knowing whether or not your passed the test isn't enough.[17] So each of the principles must be qualified: they are applicable only when the decision-problem is formulated in terms of states of the world, like dependency hypotheses, that are independent of your options.

But why? Why are these principles only applicable in cases where the states of the world are independent of your options? The **Actual Value Conception** can help explain. Here's the idea. When these principles are applied properly — that is, when we deploy dominance-reasoning only relative to a partition of dependency hypotheses, and we understand "fully-informed" in the Deference and Reflection principles to mean "knows which dependency hypothesis is actual" — you will be in a position to know something about how the actual values of your options compare. However, when these principles are misapplied, the connection to actual value is lost. For example, from the fact that *goofing* dominates *studying* relative to the partition {PASS, FAIL}, you cannot infer anything of interest about how the actual values of your options compare; it would be wrong to conclude, for example, that *goofing* has more actual value than *studying*; if K_3 describes the way the world actually is, then *studying* has more actual value than *goofing.*

So long as these principles are applied properly, then, whenever their antecedents hold, you will be in a position to be rationally certain about how the actual values of your options compare. When you're rationally certain

[17] Here's why. Suppose that you are in a position to know that any rational person with your best interests at heart who knows whether or not you passed the test would hope that you had *goof*ed rather than *studied.* Suppose that this better-informed rational person knows that you passed the test. Given our way of describing the problem, knowing that you passed is equivalent to knowing that, either, $(S \wedge K_1)$ or $(G \wedge K_1)$ or $(G \wedge K_2)$ or $(S \wedge K_3)$. Out of these possibilities, the G-worlds (i.e., the worlds in which you opted to goof around) are better than the S-worlds (i.e., the worlds in which you opted to study). But those comparisons shouldn't be relevant to what it's rational to *choose* because your actions only have the power to influence which outcome *within* a dependency hypothesis is actualized, not which dependency hypothesis is true.

about how the actual values of your options compare, the **Actual Value Conception** recommends aligning your preferences over those options with what you know about those comparisons. The Deference, Reflection, and Dominance arguments are all ways of dramatizing that you are in a position to know something relevant about your options' actual values.

Let's look at each of these arguments turn:

1. *Deference.* You are in a position to know that your friend is aware of your options' actual values. According to the Actual Value Conception, given that your friend knows the options' actual values, she should prefer one to the other if and only if the former has more actual value than the later. Because your friend wants you to have taken both boxes, you can infer that Two-Boxing has more actual value than One-Boxing.

2. *Reflection.* You are in a position to know that you will prefer having taken both boxes. This is, in part, because you know that, after making your decision, the actual values of your options will be revealed to you. So, you now know that future-you will be better-informed — in fact, fully-informed about all of the things that are relevant to this decision problem. Furthermore, assuming that future-you will value things in exactly the same way that you do now, you are in a position to infer from the fact that future-you will prefer having taken both boxes that the actual value of doing so exceeds the actual value of only taking the one box. So, you are now in a position to know that the actual value of Two-Boxing is greater than the actual value of One-Boxing.

3. *Dominance.* Partition the states of the world into dependency hypotheses. If ϕ is dominated by ψ, then you are in a position to know that the actual value of ϕ doesn't exceeded the actual value of ψ. Here's why. You know that $V_@(\phi) = V(\phi \wedge K_@)$ and $V_@(\psi) = V(\psi \wedge K_@)$. If ϕ is dominated by ψ, then you are in a position to know, for all dependency hypotheses K, that $V(\phi \wedge K) \leq V(\psi \wedge K)$. And so, even though you might not know which dependency hypothesis is actual (i.e., for each K, you don't know if $K = K_@$), you are in a position to know that, whichever it is, the value of ϕ's outcome in that state doesn't exceed

> the value of ψ's outcome. But the values of these outcomes correspond to their respective option's actual values. Therefore, you are in a position to know that the actual value of ϕ doesn't exceed the actual value of ϕ.

Each argument dramatizes the fact that you are in a position to know that the actual value of Two-Boxing exceeds the actual value of One-Boxing. If you know that the actual value of some option exceeds the actual value of another, then, by CREDENCE PREFERENCE COHERENCE, rationality requires you to prefer it.

1.6 Conclusion

This chapter outlined a view about decision-theoretic instrumental rationality — the **Actual Value Conception** — and demonstrated how this view relates to causal and evidential decision theory. I argued that the view underlies causal decision theory and that it unifies some of the intuitive arguments offered for Two-Boxing in the Newcomb Problem. I also argued that the view is incompatible with evidential decision theory.

Although the **Actual Value Conception** underlies causal decision theory, the two are not equivalent — the former is broader than the latter. In particular, in order for causal decision theory to issue recommendations, you must be able to assign precise values to all potential outcomes; that is, your non-instrumental preferences must be such that you can be represented with a *utility-function*.[18] But instrumental rationality shouldn't require you to value your ends in this manner. It's not irrational to value various things in various ways without settling, once and for all, how the various things we care about weigh-off against each other. And so, it's not irrational to have non-instrumental preferences that cannot be represented with a utility-function.

Because the **Actual Value Conception** instructs you to align your preferences over your options to your best estimates of how the actual val-

[18] Or, more carefully, a set of utility-functions that are unique up to positive linear transformations.

ues of those options *compare*, it requires less information about your non-instrumental preferences in order to issue recommendations. In particular, it doesn't require you to assign precise values to all potential outcomes; rather, it merely requires that you're able to *compare* the values of outcomes "residing" in the same state.

The next chapter addresses this issue explicitly. What does instrumental rationality require of you when your non-instrumental preferences cannot be represented with a utility-function? I will argue that the most popular way of generalizing Expected Utility Theory to cover such cases is incompatible with the **Actual Value Conception**, and suggest a novel alternative.

Chapter 2

A Guide to Gambling for the Indecisive

2.1 Introduction

Classic decision-theoretic models of practical rationality require that one's preference be *complete* (or, *trichotomous*): for any things, X and Y, that you regard as comparable, either you prefer X to Y, or you prefer Y to X, or you are indifferent between the two.[1] There are, however, a growing number of philosophers and economist who argue that practical rationality requires no such thing.[2] The power of the classic decision-theoretic models resides in what they say about what it's rational to do when facing a decision under risk or uncertainty — that is, when which outcome your action will bring about depends on features of the world you are uncertain about. Without the Completeness Axiom, however, it's no longer clear what rationality requires under risk or uncertainty.

[1] See, for example, [Savage, 1954], [von Neumann and Morgenstern, 1944], and [Anscombe and Aumann, 1963].

[2] See, for example, [Chang, 2002, 2005] [Dubra et al., 2004], [Evren and Ok, 2011], [Galaabaatar and Karni, 2013], [Hare, 2010], [Herzberger, 1973], [Joyce, 1999], [Levi, 1986, 1999, 2006], [Nau, 2006], [Ok et al., 2012], [Raz, 1985], [Seidenfeld et al., 1990, 1995], [Sen, 2004]. Even the developers of the classic models, for example [Aumann, 1962] and [Savage, 1954], express doubts that the Completeness Axiom is an honest-to-goodness constraint imposed by rationality.

In particular, Expected Utility Theory is a mathematically rigorous account of how your preferences over outcomes should relate to your preferences over options. We can think of Expected Utility Theory as a view about how the values you place on the possible consequences of a decision *transfers* to the option that might, if performed, result in those consequences. When your preferences over outcomes are incomplete, however, it's not clear that the possible consequences of a decision can be said to have an unequivocal value — and, so, it's also unclear how your (incomplete) preferences over outcomes should constrain your preferences over your options.

One popular way of generalizing Expected Utility Theory to handle incomplete preferences goes like this.[3] There are two steps. First, we represent your incomplete preference ordering (over outcomes) with *the set* of all complete preference orderings which are coherent extensions of your partial ordering. An ordering $\succeq^+$ is a *coherent extension* of a partial ordering $\succeq$ just in case: (i) $\succeq^+$ is complete, and (ii), for any outcomes X, Y, $X \succeq^+ Y$ if and only if $X \succeq Y$. (So, for example, if your incomplete preference ordering ranks outcome X ahead of outcome Y, then *every* complete preference ordering included in the set will likewise rank X ahead of Y; and so on and so forth; if, however, outcome X and outcome Y don't stand in any of the three traditional preference-relations to each other, then, for each of the ways the two can be ranked, there will be a complete preference ordering included in the set that does rank them that way).[4] Then, we can apply the traditional

[3] Among economists, frameworks of this general nature are nearly the only game in town. See, for example, [Dubra et al., 2004], [Evren and Ok, 2011], [Galaabaatar and Karni, 2013], [Ok et al., 2012].

[4] Every partial ordering can be represented, in the manner described, by a set of complete orderings. The converse, however, doesn't hold: there are sets of complete orderings that cannot be faithfully represented by a partial ordering. Here's an example. Suppose you are deciding between three dessert options: an apple pie (A), a bowl of blueberries (B), and a cantaloupe cake (C). And, at least as far as desserts are concerned, you only care about two things: how *healthy* the dessert is, and how *delicious* it is. Suppose that A is the most delicious, B is the least delicious, and C is just slightly more delicious than B; and suppose that B is the healthiest option, A is the least healthy option, and C is just slightly healthier than A. Consequently, in terms of your all-things-considered preferences, none of the three options stands in any of the traditional preference-relations to any of the others. But, in such a case, we might want to represent your motivational-state with a set of complete orderings which includes orderings that rank C ahead of A and C ahead of B, but *doesn't* include any orderings that ranks C ahead of *both* A and B. In other words, there are no admissible way of evaluating your options, resolving your concern for health and your concern for deliciousness, according to which C is the dessert that is most desirable to you. (See [Levi, 1985, 2008] for a discussion of cases with this structure). This distinction won't matter for our purposes, however.

machinery of Expected Utility Theory to each of the complete orderings in your set. You should prefer one option to another just in case every complete ordering in your set ranks things that way; you should be indifferent between two options just in case every complete ordering in your set ranks them that way; etc. Let's call the family of views with this basic structure *Supervaluational Expected Utility Theory.*[5]

2.2 Puzzle: the Vacation Boxes

Consider the following decision problem (borrowed from [Hare, 2010]). Suppose you have incomplete preference with respect to two different vacations: an alpine ski vacation (A) and a beach vacation (B). You don't strictly prefer one to the other, nor are your indifferent between the two. Following Chang [2002], let's say that the two are *on a par.*[6] The difference between parity, on the one hand, and indifference, on the other, is that the former is insensitive to mild sweetening while the latter is not. Let us suppose, for example, that you don't prefer the alpine ski vacation plus a dollar (A^+) to the beach vacation, nor do you prefer the beach vacation plus a dollar (B^+) to the alpine ski vacation. (If you were *indifferent* between the alpine ski vacation and the beach vacation, however, then you would prefer the alpine ski vacation plus a dollar to the beach vacation and you would prefer the beach vacation plus a dollar to the alpine ski vacation).[7]

There are two opaque boxes: a Larger box and a Regular box. A fair coin

[5] Here are some examples of views that fall into this class: I.J. Good's QUANTIZATIONISM [Good, 1952]; Caspar Hare's PROSPECTISM [Hare, 2010]; Isaac Levi's V-ADMISSIBILITY [Levi, 1986, 2008]; Amartya Sen's INTERSECTION MAXIMIZATION [Sen, 2004]. There are also a number of decision theories designed to handle similar cases that arise not because of incomplete preferences but because of imprecise (or unsharp) credences: for example, Susanna Rinard's MODERATE [Rinard, 2015]; Weatherson's CAPRICE [Weatherson, 2008]; and [Joyce, 2010].

[6] For Chang, *parity* is a fourth *sui generis* value relation that hold between two comparable goods. The other philosophers who argue that the Completeness Axiom should be relaxed, not because there is a fourth value relation, but rather because, e.g., preferences can be vague or indeterminate ([Broome, 1997], [Gert, 2004]). I don't intend to take sides on this issue. When I say that two things are "on a par," one should feel free to substitute in whichever analysis of the phenomenon one likes.

[7] The "sweetener" needn't be a dollar. The same issue would arise if we sweetened one of the options with 50¢, or an ice-cream cone, or 1¢, or a lottery ticket with a one-in-a-millionth chance at netting 1¢, etc. So long as you as a good contributes *some* positive value to outcome A and B, it's a potential sweetener.

has been tossed. If the coin landed heads, then A was placed in the Larger box and B was placed in the Regular box; if the coin landed tails, then B was placed in the Larger box and A was placed in the Regular box. In either case, you don't know which prize is in which box. You are asked to choose one of the two boxes, taking home whichever prize is in the box you choose.

$$\textbf{Larger box} = \begin{cases} A & \text{if Heads} \\ B & \text{if Tails.} \end{cases} \qquad \textbf{Regular box} = \begin{cases} B & \text{if Heads} \\ A & \text{if Tails.} \end{cases}$$

Given your attitudes about A and B, what attitude ought you have between the option of taking the Larger box and the option of taking the Regular box? Offhand, it might seem like the answer is that you ought to be *indifferent* between these two options. Both options afford you a 50% chance of getting A and a 50% chance of getting B. Everything that can be said in favor of choosing the one box can just as easily be said in favor of choosing the other.

Now imagine that \$1 is added to the Larger box. If you choose the Larger box, you will win whichever prize it contains plus a \$1. Nothing is added to the Regular box. Would it now be irrational to choose the Regular box? If you ought to have been indifferent between the option of taking the Larger box and the option of taking the Regular box (prior to \$1 being added to the mix), then you now ought to strictly prefer taking the Larger box to taking the Regular box.

	HEADS	TAILS
Take Larger box	A^+	B^+
Take Regular box	B	A

Standard Expected Utility Theory says nothing about cases like these because in order for *expected* utility to be well-defined, *utility* must be well-defined. But if you have incomplete preferences (as you do here), you cannot be represented with a *single* utility-function.[8] *Supervaluational Expected*

[8] Here is why you are unable to place a single, absolute value on any of these outcomes. Suppose, to the contrary, that you could. You assign the number $r \in \mathbb{R}$ to A, $u(A) = r$. Because you don't prefer A to B, the number you assign to B, $u(B)$, cannot be less than r. Because you don't prefer B to A, $u(B)$ also cannot be greater than r. Therefore, $u(B) = r$. And, because you prefer A^+ to A, it must be that $u\left(A^+\right) > r$. But, because you don't

Utility Theory, on the other hand, does have something to say about this case: namely, that it would be irrational to take the Regular box rather than the sweetened Larger box. Here's why. As mentioned above, *Supervaluational Expected Utility Theory* (which represents your motivational-state with *a set* of utility-functions) endorses the following two principles:

(1) (**Preference Supervaluationism**) For any two options ϕ, ψ, you prefer ϕ to ψ just in case option ϕ is ranked ahead of option ψ according to every utility-function in your representor.

$$\phi \succ \psi \quad \text{if and only if} \quad u(\phi) > u(\psi), \ \forall u \in \mathcal{U}$$

(2) (**Expected Valueism**) An option ϕ is ranked ahead of another option ψ according to a utility-function u in your representor just in case the *expected value* of ϕ relative to u is higher than the *expected value* of ψ relative to u. That is, each function $u \in U$ is an expected value function.

$$\forall u \in \mathcal{U}, \ \ u(\phi) = \sum_{S} Cr(S \text{ if } \phi) \cdot u(S \wedge \phi)$$

Together, these two entail that you ought to prefer taking the Larger box to taking the Regular box. You prefer A^+ to A and B^+ to B, so, by (1), every utility-function in your set ranks A^+ ahead of A and ranks B^+ ahead of B. Let u^* be an arbitrary utility-function in your set. By (2),

$$u^*(\textit{take Larger}) = \frac{1}{2} \cdot u^*\left(A^+\right) + \frac{1}{2} \cdot u^*\left(B^+\right)$$
$$u^*(\textit{take Regular}) = \frac{1}{2} \cdot u^*\left(B\right) + \frac{1}{2} : u^*\left(A\right)$$

No matter how u^* ranks A vs B, $u^*(\textit{take Larger}) > u^*(\textit{take Regular})$.[9] And because u^* was chosen arbitrarily, it follows that every utility-function in your representor ranks taking the Larger box ahead of taking the Regular

prefer A^+ to B, it cannot be the case that $u\left(A^+\right) > r$. And that's a contradiction.

[9] Here's why. $u^*(\textit{take Larger}) > u^*(\textit{take Regular})$ just in case $u^*(\textit{take Larger}) - u^*(\textit{take Regular}) > 0$. Because every utility-function in your set ranks A^+ ahead of A and ranks B^+ ahead of B, $u^*\left(A^+\right) > u^*(A)$ and $u^*\left(B^+\right) > u^*(B)$. So, $u^*\left(A^+\right) - u^*(A) + u^*\left(B^+\right) - u^*(B) > 0$. So, $\frac{1}{2}(u^*\left(A^+\right) - u^*(A)) + \frac{1}{2}(u^*\left(B^+\right) - u^*(B)) > 0$. Thus, $\frac{1}{2}(u^*\left(A^+\right) + u^*\left(B^+\right)) - \frac{1}{2}(u^*(B) + u^*(A)) > 0$.

box. So, by (1), you ought to prefer taking the Larger box to taking the Regular box.

So, *Supervaluational Expected Utility Theory* entails that you're rationally required to prefer taking the Larger box to taking the Regular box. But does rationality require such a thing? I think that, ultimately, the answer is *no:* the two options are on a par, so it's not irrational to pick the Regular box over the Larger one.

But there is a tension here. On the one hand, it seems like the answer should be *yes.* After all, there is a reason (of which you are aware) which tells in favor of choosing the Larger box — namely, that you are guaranteed to receive a dollar — which cannot be said in favor of taking the Regular box. Whatever reasons there are for choosing the Regular box are also reasons for choosing the Larger box, and you have an *additional* reason to choose the Larger one.[10] On the other hand, you know that no matter how the coin lands, you don't prefer the prize you will get by choosing the Larger box to the prize you will get by choosing the Regular box. And you might think: if you know that no matter how the world turns out to be, you don't prefer one option to the other, then rationality doesn't require you to prefer one to the other.

2.3 An Argument For Taking the Larger Box: *Prospectism*

Let's look at what I take to be the strongest line-of-thought supporting the claim that rationality requires you to take the Larger box.[11]

The idea goes like this. The prospects associated with taking the Larger box are better than the prospects associated with taking the Regular box.

[10] See [Hare, 2010] for further elaboration on this argument, and see [Bales et al., 2014] for some criticism of it. I'll address this argument in the next chapter.

[11] This argument is presented in [Hare, 2010]. Hare offers some other compelling arguments against the permissibility of taking the Regular box — the *Reasons Argument* [Hare, 2010] and the *Agglomeration Argument* [Hare, 2015] — which will be discussed in the next chapter. Because, as I'll argue in the next section, the **Actual Value Conception** entails that it's rationally permissible to take the Regular box, these arguments pose a challenge to the **Actual Value Conception.**

The *prospects* associated with an option are the various outcomes you think might result from taking that option, weighted by how likely you think it is that they will result if you take it. In this case, the prospects associated with each of your options are as follows:

$$\begin{aligned} \textsc{Prospects}(\textit{take Larger box}) &= \left\{\langle\frac{1}{2}, A^+\rangle, \langle\frac{1}{2}, B^+\rangle\right\} \\ \textsc{Prospects}(\textit{take Regular box}) &= \left\{\langle\frac{1}{2}, A\rangle, \langle\frac{1}{2}, B\rangle\right\} \end{aligned}$$

In other words, the prospects associated with taking the Larger box are a 50% shot of getting A^+ and a 50% shot at getting B^+ and the prospects associated with taking the Regular box are a 50% shot of getting A and a 50% shot of getting B. The former are better than the latter.[12] And, in general, if the prospects associated with one option are better than the prospects associated with another, you should prefer it. What rationality requires of you depends *only* on what you think might happen if you take the options and how likely you think it is for those things to happen. So rationality requires you to prefer taking the Larger box over taking the Regular box.

The argument has two premises. The first is that the prospects associated with taking the Larger box are better than the prospects associated with taking the Regular box. The second premise is that if the prospects associated with ϕing are better than the prospects associated with ψing, then you ought to prefer ϕing to ψing. Both premises are supported by the idea that the value you place on your options should only be sensitive to facts about their corresponding prospects.

> [Prospector Principle]
>
> You should prefer ϕing to ψing if and only if you regard the prospects associated with ϕing to be better than the prospects associated with ψing.

[12] To bring this out, Hare [2010] considers a different decision problem: there is only one box; it either contains prize A or prize B; you are offered a choice between taking the box *as is* or taking the box *plus a dollar*. Clearly, you should prefer the latter. But the prospects associated with each of these options — a 50% shot at getting A and a 50% shot at getting B vs a 50% shot at getting A^+ and a 50% shot at getting B^+ — are exactly the same as the prospects associated with options in the original case. So, if rationality requires that the evaluations of your options only be sensitive to those options' corresponding prospects, then you should also prefer taking the Larger box to taking the Regular box.

Prospectism — roughly, identifying the value of an option with the value of its corresponding prospects — entails that you're rationally required to prefer taking the Larger box to taking the Regular box. The view agrees with what *Supervaluational Expected Utility Theory* recommends in this case. Whether it supports what *Supervaluational Expected Utility Theory* recommends is *all* cases depends on what it is, in general, for the prospects associated with one option to be *better than* the prospects associated with another.

Hare [2010] doesn't offer a full account of what it is for the prospects associated with one option to be better than those associated with another, but here are some sufficient conditions: if, for example, one option dominates another (relative to an appropriate partition), then the prospects associated with the former are better than the prospects associated with the latter; or, for example, if the expected utility of one option exceeds that of another (supposing both are well-defined), then the prospects associated with the former are better than those associated with the latter.[13] If the prospects associated with one option are better than the prospects associated with another (in one of the ways just described), then *Supervaluational Expected Utility Theory* will recommend preferring the former option to the latter. Is the converse true? The answer, of course, depends on what it is for some prospects to be better than others. But for the sake of argument, let's grant

[13] Both of these sufficient conditions follow from the PROSPECTOR PRINCIPLE: if ϕing dominates ψing (relative to an appropriate partition), you should prefer ϕing to ψing, and so you should regard the prospects associated with the former to be better than the prospects associated with the latter; similarly, if ϕing has greater expected utility than ψing, again, you should prefer ϕing to ψing, and so you should regard the prospects associated with the former to be better than those associated with the latter. Moreover, if the prospects associated with ϕ^* are the same as those associated with ϕ and if the prospects associated with ψ^* are the same as those associated with ψ, and if you should prefer ϕing to ψing, you should, according to the PROSPECTOR PRINCIPLE, also prefer ϕ^*ing to ψ^*ing.

that it is.[14]

> **The Relationship between Supervaluational Expected Utility & Prospectism:** *Supervaluational Expected Utility Theory* recommends preferring ϕ to ψ if and only if the prospects associated with ϕing are better than the prospects associated with ψing.

The prospects associated with an option abstract away from information about which *states of the world* the outcomes that might result from taking that option reside in. As we've seen, decision problems can be represented with three different entities: there are your *options* (or "alternatives," or "acts"), which are the objects of your instrumental preferences; there are the *outcomes* that might result from performing your options, which are the objects of your *non*-instrumental preferences; and there are *states*, which are those features of the world not under your control that influence the outcomes that might result from performing your various options. And we can think

[14] In order for the argument to succeed, the proponents of *Prospectism* need to provide an account of what it is for some prospects to be better than others which satisfies the following:

> [Better Prospects]
>
> The prospects associated with option ϕ are better than the prospects associated with option ψ if and only if, for every utility-function u in your representor, u ranks ϕ ahead of ψ.

To my knowledge, no such account has been offered. However, some remarks in Rabinowicz [2016] — as well as the representation theorems for agents with incomplete preferences in [Seidenfeld et al., 1995], [Ok et al., 2012], [Nau, 2006], [Evren and Ok, 2011], and elsewhere — suggest that such an account could be provided. The prospects associated with an option are, more or less, *lotteries*: probability distributions over possible prizes. Classic representation theorems (like, for example, those in [von Neumann and Morgenstern, 1944] and [Anscombe and Aumann]), as well as the representation theorems for agents with incomplete preferences just mentioned, put forth various axioms governing rational preference over lotteries. We could, instead, interpret the axioms as jointly providing an account of what it is for some prospects to be better than others. The aforementioned representation theorems for agents with incomplete preferences say, roughly, that if your preferences over lotteries obey the axioms, you can be represented with a *set* of utility-functions according to which you prefer one option to another when, and only when, every utility-function in the set ranks the former ahead of the latter. Adapting the result to our reinterpretation of the axioms, we can say: there is a set of utility-functions according to which you regard the prospects associated with one option to be better than the prospects associated with another just in case every utility-function in the set ranks the former ahead of the latter. A full account would need to justify each axiom, on this new interpretation, as well as show that, not only is there *some* set of utility-functions that all rank options with better prospects ahead of options with worse prospects, but that this set corresponds to the one we would get from taking all the coherent extensions of your incomplete preferences. For our purposes, though, let's grant, for the sake of argument, that something like this holds: i.e., that Prospects(ϕ) are better than Prospects(ψ) if and only if *Supervaluational Expected Utility Theory* recommends preferring ϕ to ψ.

of your options as *functions* from states to outcomes. If you know what your options are, you thereby know which outcomes reside in which states; if you know only the prospects associated with each option, however, you might not know this. What role do these states play in evaluating the choiceworthiness of your options? According to the PROSPECTOR PRINCIPLE, the answer is: *an eliminable one* — facts about which outcomes reside in which states are relevant only insofar as they affect either (i) your assessments of how likely a particular outcome is to result from performing one of your options or (ii) exactly which outcomes might result from performing one of the options.

Offhand, this all might seem exactly right. After all, instrumental rationality is a matter of doing what is best given your perspective — how you *believe* the world is and how you *desire* it to be. And one might think that facts about your perspective wholly supervene on the facts about (i) how likely you take it to be that the possible outcomes of your decision will result from performing the options available to you and (ii) the value you place on the various possible outcomes that might result. And, so the thought goes, the prospects associated with your options — by their very nature — take into account all of the information that is relevant to instrumental rationality.

As we'll see, however, this thought is incompatible with the **Actual Value Conception**: states play an *ineliminable* role in evaluating the choiceworthiness of your options (at least when your preferences over the outcomes are incomplete). Roughly, the problem is that the argument sketched above — which turns on the idea that instrumental rationality is matter of doing what's best given your perspective — doesn't support the PROSPECTOR PRINCIPLE. One's "perspective" encompasses more than merely the credences one assigns to the possible outcomes of one's decisions: one's perspective also encompasses beliefs regarding the *actual values* of one's options. And, insofar as one is sympathetic to the picture of practical rationality underpinning the **Actual Value Conception**, these beliefs are relevant to what it's rational to do.

2.4 It's Okay to Take the Regular Box

As we've just seen, the prospects associated with taking the Larger box are better than the prospects associated with taking the Regular box. If the PROSPECTOR PRINCIPLE is correct, then rationality requires you to prefer opting for the Larger box over the Regular box. However, the **Actual Value Conception**, which I consider to be an intuitively compelling picture of instrumental rationality, entails that rationality requires no such thing.

In this section, I will argue that the **Actual Value Conception** entails that it's rationally permissible to take the Regular box. I'll, then, present some of the arguments that have been given against it being a requirement of rationality to prefer the Larger box, and show that the **Actual Value Conception** underlies them all.[15] The upshot of the ensuing discussion in this. There is a tight connection between two, seemingly unrelated, issues in decision theory: the debate between Two-Boxing vs One-Boxing, on the one hand, and the debate about whether rationality requires you to prefer the Larger box over the Regular box, on the other. The same reasons that support Two-Boxing over One-Boxing also support regarding the Regular box to be on a par with the Larger box.

2.4.1 The Actual Value Conception & Taking the Regular Box

Recall the view sketched in the previous chapter: the **Actual Value Conception**. It is an account of instrumental rationality according to which you should align your preferences over your options to your best estimates of how the actual values of those options compare.

[15] These arguments are discussed in [Hare, 2010], [Schoenfield, 2014], [Bales et al., 2014], and [Rabinowicz, 2016].

[The Actual Value Conception]

You are instrumentally rational if and only if, for all options ϕ and ψ,

$$\phi \succ \psi \quad \text{when, and only when,} \quad \text{Estimate}\Big[\mathcal{CV}_{@}(\phi,\psi)\Big] > \text{Estimate}\Big[\mathcal{CV}_{@}(\psi,\phi)\Big]$$

$$\phi \approx \psi \quad \text{when, and only when,} \quad \text{Estimate}\Big[\mathcal{CV}_{@}(\phi,\psi)\Big] = \text{Estimate}\Big[\mathcal{CV}_{@}(\psi,\phi)\Big]$$

In the previous chapter, we defined $\mathcal{CV}_{@}$ in terms of $\mathcal{V}_{@}$ — in particular, we took $\mathcal{CV}_{@}(\phi, \psi)$ to equal the difference between $\mathcal{V}_{@}(\phi)$ and $\mathcal{V}_{@}(\psi)$. If your preferences are incomplete, your ends cannot be represented with a utility-function. And if your ends cannot be represented with a utility-function, $\mathcal{V}_{@}$ needn't be well-defined. However, even if $\mathcal{V}_{@}(\phi) - \mathcal{V}_{@}(\psi)$ isn't well-defined, there might nevertheless be a fact of the matter about the extent to which the actual value of ϕ exceeds the actual value of ψ.

Here's an example. Imagine, like before, that a coin has been tossed. If the coin landed heads, prize A is in the box; if the coin landed tails, prize B is the box. You have two options: you can take the box *with* a dollar, or your can take the box *without* a dollar.

	Heads	Tails
With	A^+	B^+
Without	A	B

Because you regard prize A to be on a par with prize B, $\mathcal{V}_{@}$ isn't well-defined: it makes no sense to say how valuable *full stop* any of the prizes are, and, because an option's actual value is equal to the value of the prize you would actually receive were you to perform it, your options fail to have well-defined actual values. Nevertheless, in this case, we can still *compare* the actual values of your options. If the coin landed heads, then, because you prefer A^+ to A, the actual value of *With* exceeds the actual value of *Without*; and if the coin landed tails, then, because you prefer B^+ to B, the actual value of *With* exceeds the actual value of *Without*. In fact, so long as you regard the value of getting \$1 as independent of getting prize A or B, then $\mathcal{CV}_{@}(\textit{With}, \textit{Without}) = \mathcal{V}(\$1) > 0$. So, we should relax the assumption, made in the previous chapter, that $\mathcal{CV}_@(\phi, \psi) = V_@(\phi) - V_@(\psi)$; rather, this equivalence only holds in the special case when your ends can be represented with a utility-function.

If performing option ϕ would result in an outcome that you regard as on a par with the outcome that would actually result from performing ψ, then $\mathcal{CV}_@(\phi, \psi) \ngtr 0$ and $\mathcal{CV}_@(\psi, \phi) \ngtr 0$. The actual value of ϕ doesn't exceed the actual value of ψ, and *vice versa*; the two options, just like the outcomes performing them would bring about in the actual world, are on a par. Ideally, if you were omniscient (and so knew which outcomes would actually result from performing your options), you would regard ϕ as on a par with ψ. Even if you aren't omniscient, though, if you happen to know that the outcomes that would actually result from performing your options are on a par, then, insofar as you're instrumentally rational, you should regard those options as on a par as well.[16]

> **The Regulative Ideal (ver 3):** "Aim to be such that you strictly prefer one option to another if and only if the actual value of the former exceeds the actual value of the latter; aim to be indifferent between two options if and only if their actual values are equal; aim to regard two options as on par if and only if you regard their actual outcomes as on a par."

$$\begin{aligned}
\phi \succ \psi &\text{ when, and only when } \mathcal{CV}_@(\phi, \psi) > \mathcal{CV}_@(\psi, \phi) \\
\phi \approx \psi &\text{ when, and only when } \mathcal{CV}_@(\phi, \psi) = \mathcal{CV}_@(\psi, \phi) \\
\phi \bowtie \psi &\text{ when, and only when } \mathcal{CV}_@(\phi, \psi) \ngeq \mathcal{CV}_@(\psi, \phi) \;\&\; \mathcal{CV}_@(\phi, \psi) \nleq \mathcal{CV}_@(\psi, \phi)
\end{aligned}$$

According to the **Actual Value Conception**, if you're certain that the actual value of an option doesn't exceed the actual value of another, then you shouldn't prefer it. In the previous chapter, we called this principle CREDENCE PREFERENCE COHERENCE and we showed that it's entailed by the **Actual Value Conception.**[17]

[16] Let's write $\ulcorner X \bowtie Y \urcorner$ to mean that X is on a par with Y.

[17] This principle is more-or-less equivalent to the second clause of a principle Schoenfield [2014] calls "LINK." It says (where p is your rational credence function, and $\ulcorner V(X) > V(Y) \urcorner$ says that *the outcome that would actually result from choosing X is better than the outcome that would actually result from choosing Y.*):

If $p(V(\phi) > V(\psi)) = 0$ & $p(V(\psi) > V(\phi)) = 0$, then $EV^p(\phi) \ngtr EV^p(\psi)$ &

> [CREDENCE PREFERENCE COHERENCE]
>
> If you are certain that the actual value of ϕ doesn't exceed the actual value of ψ, then you shouldn't strictly prefer ϕ to ψ.
>
> $$\text{If } Cr\Big(\mathcal{CV}_{@}(\phi, \psi) > 0\Big) = 0, \text{ then } \phi \not\succ \psi$$

If you're certain that the actual value of ϕ doesn't exceed the actual value of ψ, you shouldn't prefer ϕ to ψ. Ideally, you'd prefer one option to another when, and only when, the actual value of the former exceeds the actual value of the latter. If you're certain that the actual value of ϕ doesn't exceed the actual value of ψ, then you know that by preferring ϕ to ψ you'll thereby violate the ideal.

CREDENCE PREFERENCE COHERENCE entails that you shouldn't prefer the Larger box to the Regular box. Suppose that as a matter of fact (but, of course, unbeknownst to you) the Larger box contains A^{+} and the Regular box contains B. The actual outcome that would result, then, from choosing the Larger box is the one in which you get A^{+} and the actual outcome that would result from choosing the Regular box is the one in which you get B. If the world is as just described, it is *as if* you are choosing between prize A^{+} and prize B. Because you regard both prizes to be on a par, the actual value of taking the Larger box does not exceed the actual value of taking the Regular box. Suppose instead that the coin had landed the other way. Analogous reasoning gets us to the same conclusion: the actual value of taking the Larger box doesn't exceed the actual value of taking the Regular

$$EV^{p}(\psi) \not> EV^{p}(\phi).$$

In words: if you are rationally certain that the value of ϕing doesn't exceed the value of ψing (and *vice versa*), then neither should have higher expected value than the other.

Schoenfield [2014] defends LINK by arguing that "if LINK is rejected, expected value theory cannot play the role that it was intended to play: namely, providing agents with limited information guidance concerning how to make choices in circumstances in which value-based considerations are all that matter." (pg. 268). Schoenfield [2014] claims that it's central to the role we want expected value theory to play that it's recommendations not conflict with what you know about how the actual values of your options compare. As the discussion in the previous chapter suggests, however, it's not true that every version of expected value theory satisfies LINK. The Newcomb Example brings out that *evidential* decision theory violates the constraint. And, at least offhand, evidential decision theory is a satisfactory account of expected value. Schoenfield [2014]'s argument is persuasive only if we limit our attention to accounts of expected value that are supported by the **Actual Value Conception**.

box. You are in a position to know all of this, and are able to reason as follows:

Reasoning By Cases

P1 The coin has landed either Heads or Tails.

P2 If the coin has landed Heads, the actual value of taking the Larger box does not exceed the actual value of taking the Regular box.

P3 If the coin has landed Tails, the actual value of taking the Larger box does not exceed the actual value of taking the Regular box.

C The actual value of taking the Larger box does not exceed the actual value of taking the Regular box.

This is a valid argument.[18] And you are in a position to know each of the premises. You are, therefore, in a position to know the conclusion: *that the actual value of the Larger box doesn't exceed the actual value of taking the Regular box.* That is, you're in a position to know $\mathcal{CV}_{@}(L,R) \not> 0$ and it would be epistemically rational of you to have the following credence:

$$Cr\Big(\mathcal{CV}_{@}(L,R) > 0\Big) = 0$$

If you endorse Credence Preference Coherence — as you should if you are at all sympathetic to the **Actual Value Conception** — then you shouldn't think that rationality requires you to prefer taking the Larger box over taking the Regular box. Consequently, *Supervaluational Expected Utility Theory* is incorrect.

[18] Not all arguments of this form — i.e., reasoning by cases with indicative conditionals — are valid (as the much-discussed Miners Puzzle makes clear [Kolodny and McFarlane, 2010]). The reasoning leads us astray when the consequents of the indicative conditionals are "information-sensitive." But the As-a-Matter-of-Fact value of some option doesn't depend on what information you have; it depends only on which prizes are, as a matter of fact, in which box. The reasoning here is analogous to following non-puzzling Miners argument: "Either the miners are in shaft A or they are in shaft B; if they are in shaft A, then blocking neither shaft saves fewer lives than something else I could do; if they are in shaft B, then blocking neither shaft saves fewer lives than something else I could do; therefore, blocking neither shaft saves fewer lives than something else I could do." That's a fine argument. It would be a mistake, however, to take the conclusion to be a decisive reason to not block either shaft.

In the next sections, we will look at some of the arguments that have been given against it being a requirement of rationality to prefer the Larger box to the Regular box. I will argue that these arguments inherit their persuasiveness from the **Actual Value Conception.**

2.4.2 Compelling Reason 1: Deference & Reflection

In the service of presenting the first argument, let's consider a variation on our original case. Everything is the same as before — there are two boxes, one contains an alpine ski vacation, the other contains a beach vacation, a dollar has been added to the Larger box, etc. — except in this case *you* don't win whichever prize is in the box you choose, your best friend does. You are playing **The Vacation Box Game** for your friend, who is sitting in the studio audience and can see (and thus knows) which box contains which prize. You are a good friend and want what's best according to your friend's preferences. (You desperately want to avoid seeing your friend's face melt into an expression of disappointment upon the announcement of your choice). You know, however, that (much like you) your friend lacks complete preferences over the prizes — she, of course, prefers A^+ to A and B^+ to B but is otherwise torn. From your friend's (better-informed) perspective, you are facing one of two possible decision problems: one where you are choosing between A^+ or B, and another where you are choosing between B^+ or A. You don't know which of these two decision problems you are facing but you *do* know that, either way, your friend doesn't prefer that you take the Larger box over the Regular box; she isn't crossing her fingers and muttering under her breath: "take the Larger box, take the Larger box, please take the Larger box..."; she won't be particularly angry or disappointed or vexed with you for choosing the Regular box. Furthermore, suppose that from her position in the audience, your friend can secretly signal to you what she'd like you to do (by nodding her head toward one of the boxes, for example). Whether she could secretly send such a signal, you know that she wouldn't; she doesn't particularly care what you do.

Does rationality, nevertheless, require you to choose the Larger box? Given that you *knew* that your friend would be totally fine with you choosing the Regular box, and that all you cared about in this situation was doing right by your friend, what could possibly render choosing the Regular box rationally off limits? You know that, no matter which box contains which prize, your friend would be okay with you just randomly picking either.

Call this **The Argument from Deference.** You are in a position to know that your friend doesn't prefer that you take the Larger box over the Regular box on her behalf. And you want to do what your friend prefers that you do. Moreover, in this situation, wanting to do what your friend prefers you to do is all that you care about. You are in a position to know, then, that your goal — that is, doing whatever it is that your friend prefers you to do — isn't better served by taking the Larger box than it is by taking the Regular box. So, taking the Larger box isn't a better means to your ends than taking the Regular box. And so, rationality doesn't require you to prefer taking the Larger box.

Even if you're convinced that it's okay to not take the Larger box when you are playing the game for your best friend, what about in the original friendless version? What holds about what you ought to do when playing the game on your friend's behalf also holds when you are playing the game for yourself. Your best friend was only a rhetorical device meant to dramatize something that was true about you all along: you are in a position to know that you *will* not prefer having taken the Larger box to having taken the Regular box (and *vice versa*). Later, after having made your decision and opening the box, you will not prefer the prize you've won to the prize you could've won had you chosen differently. If, despite all this, you now do prefer taking the Larger box to the Regular box, you violate the following principle:

> [Preference Reflection]
>
> If you are now in a position to know that you will not prefer having ϕed to having ψed, then you shouldn't now prefer ϕing to ψing.

If this principle is correct, then rationality doesn't require you to prefer taking the Larger box. Call this **The Argument from Reflection.**[19] If

[19] These arguments are discussed in more detail in [Hare, 2010]. Two principles are appealed to. The first Hare [2010] calls *Deference*; it says: "If I know that any fully informed, rational person, with all and only my preferences between maximal states of affairs, would have a certain array of preferences between sub-maximal states of affairs

you now, in the situation described, only care about doing what future-you will prefer present-you to have done, then, because you are in a position to know that future-you will not prefer having taken the Larger box, you also know that taking the Larger box isn't a better means to your ends than taking the Regular box. And, so, rationality doesn't require you to prefer taking the Larger box.

2.4.3 Compelling Reason 2: Dominance

In this section, I will present one more species of argument against being rationally required to prefer the Larger box: **Dominance Arguments.**

The first Dominance Argument goes like this.[20] The coin can land either heads or tails. If the coin lands heads, then the outcome that would result from taking the Larger box isn't preferred to the outcome that would result from taking the Regular box. If the coin lands tails, then, similarly, the outcome that would result from taking the Larger box isn't preferred to the outcome that would result from taking the Regular box. Preferring the Larger box to the Regular box, then, violates the following dominance principle:

> [NO-PREFERENCE DOMINANCE]
>
> Let $\mathbf{Z} = \{Z_1, Z_2, \ldots, Z_n\}$ be a partition of the ways the world might be. For any two options, ϕ and ψ, if, for all $Z_i \in \mathbf{Z}$, you don't prefer $(\phi \wedge Z_i)$ to $(\psi \wedge Z_i)$, then rationality doesn't require you to prefer ϕing to ψing.

on my behalf, then it is rationally permissible for me to have that array of preferences between sub-maximal states of affairs" (pg. 242). The second says, roughly, that if it's rationally permissible to have no preference for one option over another, and *vice versa*, then it's rationally permissible to take either option.

[20] Both Bales et al. [2014] and Rabinowicz [2016] discuss arguments of this form (approvingly, in the former case, but disapprovingly in the latter). Rabinowicz [2016] discusses a principle he calls *Complementary Dominance (V)*, which says: "One action is not better than another if it under every state yields an outcome that is not better than the outcome of the other action." Bales et al. [2014] argue, on intuitive grounds, for a principle they call *Competitiveness*. It says that it's rationally permissible to perform a competitive action, where an action is *competitive* if "for every way the world could be, its consequences are no worse than the consequences of all alternative actions." (pg. 460). These two principles are formulated differently, but the differences stop there.

Partition the ways the world might be into two: the worlds in which the coin landed heads (H), and the worlds in which the coin landed tails (T). Because you don't prefer A^+ to B, you don't prefer $(L \wedge H)$ to $(R \wedge H)$; because you don't prefer B^+ to A, you don't prefer $(L \wedge T)$ to $(R \wedge T)$. So, according to NO-PREFERENCE DOMINANCE, rationality doesn't require you to prefer taking the Larger box to taking the Regular box.

In fact, *Supervaluational Expected Utility Theory* sometimes offers advice which violates an even weaker dominance principle. Consider one more variant of **Vacation Boxes**. Everything is the same as before, except this time your options are slightly different: you can take the Regular box for free, or you can pay a small fee to take the Larger box, or you can let the game show host flip *another* fair coin to decide which of the two boxes you get (if this coin lands heads, let's say, you get whichever prize happens to be in the Larger box; if it lands tails, you get whichever prize is in the Regular box).

L$^-$ You pay \$$\epsilon$ for the Larger box.

R You get the Regular box for free.

M Another coin is tossed. If it lands heads, you get the Larger box; if it lands tails, you get the Regular box.

Taking the Larger box has better prospects than option M. If *Supervaluational Expected Utility Theory* is correct, you rationally ought to prefer taking the Larger box to option M.[21] In general, if you prefer X to Y, then there should be some amount of positive value (however small) such that you'd be willing to pay that amount in order to get X rather than Y.[22] Pick a suitably small enough value for \$$\epsilon$ so that (according to *Supervaluational*

[21] One way to see this is by directly doing the calculations: option L nets you a 50% shot at getting A^+ and a 50% shot at getting B^+, whereas option M nets you a 25% shot at getting A^+, a 25% shot at getting B^+, a 25% shot at getting A, and a 25% shot at getting B. All of the utility functions in your representor, then, rank the former over the latter. (In fact, so long as the value of the \$1 sweetening is independent of the vacation prizes, there is precise amount by which option L exceeds option M in value: $\frac{1}{2} \cdot V(\$1)$. You should, according to *Supervaluational Expected Utility Theory* be willing to pay anything up to that amount in order to secure L over M.)

Another, more indirect, way to see it is to note that, at least in general, if you prefer X to Y, then you should also prefer X to all of the *probabilistic mixtures* of X and Y.

[22] Standardly, there is an amount such that it is *the most* you'd be willing to pay. If your preference are incomplete, however, we shouldn't assume that there is a unique value such that it is the most you'd be willing to pay.

Expected Utility Theory, at least) you're rationally required to prefer paying that amount for L over option M. You should, then, prefer L^- to M. But if the Larger box contains A^+ and the Regular box contains B, you would be fine with opting for option M; if the Larger box contains B^+ and the Regular box contains A, again, you would be fine with opting for option M. Either way, you *don't* prefer selecting the Larger box rather than option M.

Furthermore, if you choose L^- over M, there's a 50% chance that you'll be making yourself worse off than you could've been had you chosen M instead; and there's no chance that you'll be making yourself better off.

	Larger: A^+ & **Regular:** B		**Larger:** B^+ & **Regular:** A	
	HEADS	TAILS	HEADS	TAILS
Option L^-	$A^+ - \$\epsilon$	$A^+ - \$\epsilon$	$B^+ - \$\epsilon$	$B^+ - \$\epsilon$
Option R	B	B	A	A
Option M	A^+	B	B^+	A

You are in a position to know that M *might* do better than L^- and, also, in a position to know that L^- definitely won't do better than M. If, for all you know, there are some ways that some option can do better than another and there are no ways that it can do worse, it should, at the very least, be *permissible* to take it (when these are the only two options at play). But if *Supervaluational Expected Utility Theory* is correct, then there are some cases — this one, for example — in which you are rationally *required* to choose an option that is *weakly dominated* in the manner just described.[23] In other words, *Supervaluational Expected Utility Theory* violates the following dominance principle:[24]

[23] Standardly, it is said that one option X *weakly dominates* another Y, just in case, in every state, X leads to *at least as good* an outcome as Y, and there is at least one state in which X leads to a better outcome than Y. Option M, then, doesn't weakly dominate L^- in this sense because, e.g., B isn't *at least as good* as $A^+ - \$\epsilon$ (it isn't better; it isn't worse; they're on a par). Option M does weakly dominate L^- in a *weaker* sense, though: it never does worse and it sometimes does better.

[24] This principle, like all dominance principles, should be understood with the caveat that the partition of states is such that each of the states are independent of which option you take. If you like causal decision theory, the principle holds only with respect to a partition of states that are *causally* independent of your options. If you like evidential decision theory, on the other hand, more is required for the principle to hold: the partition of states and your options must be *probabilistically* independent. In the case under consideration,

[The Weak Quasi Dominance Principle]

> For any option ϕ, if there is some other available option ψ, such that for every state Z, you don't prefer outcome $(\phi \wedge Z)$ to outcome $(\psi \wedge Z)$, and there is some state Z' such that you do prefer outcome $(\psi \wedge Z')$ to $(\phi \wedge Z')$, then rationality *doesn't require* you to ϕ.

One option *weakly quasi dominates* another if it never does worse and might do better. This principle says that if an option is *weakly quasi dominated,* then rationality doesn't require you to choose it.[25]

2.4.4 The Deference, Reflection, and Dominance Arguments as Guides to Actual Value

The Actual Value Conception underlies the previously discussed arguments against being rationally required to prefer taking the Larger box. Each argument succeeds, if it does, by showing that you are in a position to know that the actual value of taking the Larger box doesn't exceed the actual value of taking the Regular box.

Let's look at each argument, in turn.

1. *Deference.* You are in a position to know that your friend is aware of your options' actual values. According to the Actual Value Conception, given that your friend knows the options' actual values, she should prefer one to the other if and only if the former has more actual value than the later. But, because you know that your friend doesn't prefer

the relevant partition of states — worlds in which the coin lands heads and worlds in which the coin lands tails — is *both* causally and probabilistically independent of your options.

[25] How does this dominance principle compare with Bales et al. [2014]'s *Competitiveness* (which, recall, says "that an action is rationally permissible if, for every way the world could be, its consequences are no worse than the consequences of all alternative actions")? When there are only two available options, my principle is weaker than theirs (if there are only two options, and one weakly quasi dominates the other, then the dominating option is competitive; and so, by their principle, it is permissible; and so you are not rationally required to take the other option). If there are more than two available options, the principles are logically independent. (Imagine a case in which there are three options such that each is weakly quasi dominated by one of the others, and so none of the three are competitive).

the Larger box to the Regular box, you are in a position to infer that the actual value of taking the Larger box isn't greater than the actual value of taking the Regular box.

2. *Reflection.* You are in a position to know that you will not prefer having chosen the Larger box. This is, in part, because you know that, after making your decision, the actual values of your options will be revealed to you. So, you now know that future-you will be better-informed — in fact, fully-informed about all of the things that matter you relevant to this decision problem. Furthermore, assuming that future-you will value things in exactly the same way that you do now, you are in a position to infer from the fact that future-you will not prefer having taken the Larger box that the actual value of doing so doesn't exceed the actual value of taking the Regular box. So, you are now in a position to know that the actual value of taking the Larger box isn't greater than the actual value of taking the Regular box.

3. *Dominance.* Partition the states of the world into dependency hypotheses. If ϕ is dominated by ψ — in either of the ways characterized by the dominance principles above — then you are in a position to know that the actual value of ϕing doesn't exceed the actual value of ψing. Here's why. You know that $\mathcal{CV}_{@}(\phi, \psi) = \mathcal{CV}_{K_{@}}(\phi, \psi)$. If ϕ is No-Preference- or Weakly Quasi- dominated by ψ, then you are in a position to know, for all dependency hypotheses K, that $\mathcal{CV}_{K}(\phi, \psi) \not> 0$. And so, even though you might not know which dependency hypothesis is actual (i.e., for each K, you don't know if $K = K_{@}$), you are in a position to know that, whichever it is, the value of ϕ's outcome in that state doesn't exceed the value of ψ's outcome. But the comparison between the values of these outcomes correspond to the comparison between their respective option's actual values. Therefore, you are in a position to know that the actual value of ϕing doesn't exceed the actual value of ϕing.

If you know that the actual value of some option doesn't exceed the actual value of another, then, by CREDENCE PREFERENCE COHERENCE, rationality shouldn't require you to prefer it.

2.5 A Decision Theory that Respects Actual Value

As we've seen, *Supervaluational Expected Utility Theory* is in conflict with the **Actual Value Conception.** The former entails that you are rationally required to prefer taking the Larger box to taking the Regular box, while the latter entails that, because you are justified in being certain that the actual value of the Larger box fails to exceed the actual value of the Regular box, rationality requires no such thing. In this section, I will present a competitor to *Supervaluational Expected Utility Theory* — a decision theory for agents with incomplete preferences that's supported by the **Actual Value Conception.**

Here's what we'll do. First, I'll present three desiderata that, in my view, any adequate decision theory for agents with incomplete preferences must satisfy in order to count as respecting the **Actual Value Conception**. Then, I'll present an idealized version of the decision theory (one that makes an unrealistic assumption about your value-structure: namely, that, whenever two goods, X and Y, are on a par, there is always a precise amount of value that can be added to X such that it is the least amount of value that needs to be added in order for X plus it to be preferred to Y). Next, I'll show that the view satisfies the three desideratum. Finally, I'll sketch how the view can be weakened to handle cases in which this unrealistic assumption fails to hold.

2.5.1 What Should Such a Decision Theory Look Like?

What must a decision theory for agents with incomplete preferences look like in order to be supported by the **Actual Value Conception**?

Think of a decision theory like a helpful advisor: you give your advisor information about how you take the the world to be, and information about how you value outcomes, and your advisor issues recommendations about what you rationally ought to do. Standard decision theories require a great deal of information about how you value outcomes in order to issue recommendations: they require you to have complete preferences over all possible outcomes, and that, for any two outcomes, there be a determinate fact about the precise degree to which you prefer the one to the other. Without this information, standard decision theories remain *silent* — they are unable to offer any recommendations; they have nothing to say about what rationality requires or permits you to do. As we've seen, the **Actual Value Conception** requires slightly less of you: your advisor only needs information about how the values of the outcomes in the same dependency hypothesis compare. In particular, it requires, for any options, ϕ and ψ, and for each dependency hypothesis K, that there be some real number r such that $\mathcal{CV}_K(\phi, \psi) = r$.[26] Even this, I think, requires more from you than is needed.

Any adequate decision theory for agents with incomplete preferences should be *robust:* it should require less input than the standard views in order to issue recommendations. In order to respect the **Actual Value Conception**, the proposal should be a generalization of a standard decision theory that's supported by the conception. Furthermore, in cases like **Vacation Boxes**, where you're certain that the actual outcomes of your decision are on a par, the proposal should avoid recommending that you prefer either option to the other. But, in order to be robust, the proposal shouldn't be rendered completely silent in all but the most trivial cases of parity. In other words, any adequate decision theory for agents with incomplete preferences that respects the **Actual Value Conception** should meet the following three desiderata:

First, the view should be a generalization of a version of *Expected Utility Theory* that is supported by the **Actual Value Conception.** The view developed here is a generalization of causal decision theory: if you had complete preferences, the view should give the same recommendations as causal decision theory.[27]

Second, in **Vacation Boxes**, the view shouldn't recommend preferring the Larger box to the Regular box. More generally, the view should entail CREDENCE PREFERENCE COHERENCE: if, when considering two options, you are position to be rationally certain that the actual value of the former

[26] More carefully, it requires, for any two options, ϕ and ψ, and for each K, that, *given a conventionally chosen zero-point and scale,* there be a real number r such that $\mathcal{CV}_K(\phi, \psi) = r$.

[27] It will not be difficult to see how the proposal can be amended in order to generalize benchmark decision theory instead. For the sake of presentational perspicuity, however, I will only focus on the version that generalizes causal decision theory.

doesn't exceed the actual value of the latter, you shouldn't prefer the former to the latter.

Lastly, if you are sufficiently confident that the actual value of ϕ exceeds the actual value of ψ to a significant extent, then, even though it *might* be the case that the outcome that would result from performing ϕ is on a par with the outcome that would result from performing ψ, the view should recommend preferring ϕ to ψ. In other words, the view should be capable of offering non-trivial recommendations in the face of parity. Here's an example.

> **Probabilistic Sweetening.** There are two boxes in front of you: the Larger box and the Regular box. A *biased* coin has been tossed. If it landed heads, then $1,000,000 has been place in the Larger box and $0 has been placed in the Regular box. If the biased coin landed tails, then a *fair* coin was tossed. If the fair coin landed heads, then A has been placed in the Larger box and B has been placed in the Regular box. If the coin landed tails, then B is in the Larger box and A is the Regular box.

	K$_1$	K$_2$	K$_3$
L	A	B	$1,000,000
R	B	A	$0

Even though you regard A and B as on a par, if your credence that the biased coin landed heads is sufficiently great, then you ought to prefer L to R. However, if your credence in the biased coin landing heads is sufficiently low, then you shouldn't be rationally required to prefer L to R. We want a decision theory that, in cases like these, offers recommendations that are sensitive to your credence in receiving the money if you take L. We also want the decision theory's recommendations, in cases of this sort, to be sensitive to how much money you might win if you take L, and how valuable you take receiving that sum of money to be.

2.5.2 Actual Value Decision Theory & the "Elasticity" of Parity

Recall that, according to the **Actual Value Conception**, you should align your preferences over your options with your best estimates of how the actual values of those options compare.[28] If we partition the ways the world could be into dependency hypotheses, then, in estimating the *comparison* of actual values of your options, you compare outcomes the "reside in" the same state. Each dependency hypothesis is, in effect, a hypothesis about how the actual values of your options might compare.

Can the view be generalized to cases in which the outcomes of your options might be on a par? If you give some credence to dependency hypothesis K being actual, and you regard outcome $(\phi \wedge K)$ as on a par with outcome $(\psi \wedge K)$, how should this be reflected in your estimation of the extent to which the actual value of ϕ might exceed the actual value of ψ?

Here's a sketch of the proposal. You want to estimate the extent to which the actual value of an option ϕ exceeds the actual value of an option ψ. Partition the ways the world might be into dependency hypotheses. Each dependency hypothesis determines a possible way the actual values of your options, for all you know, might compare. Speaking somewhat-metaphorically: the dependency hypotheses according to which ϕ does better than ψ (if there are any) contribute a *positive* amount to your estimate of the extent to which ϕ has more actual value than ψ; the ones according to which ψing does better (if there are any) contribute a *negative* amount to your estimate; the dependency hypotheses according to which ϕ and ψ are equally good (if there are any) contributes *nothing*, positive or negative, to your estimate. What about the dependency hypotheses (if there are any) according to which ϕ and ψ are on a par? If outcomes $(\phi \wedge K)$ and $(\psi \wedge K)$ are on a par, then, if K is the way the world actually is, the actual value of ϕ doesn't exceed the actual value of ψ, and *vice versa;* so, hypothesis K neither contributes a positive

[28] And, also recall, that the idea behind the **Actual Value Conception** is that, ideally, you should match your preference-like attitudes over your options to the facts concerning how the actual values of those options compare. So, for example, if you prefer outcome $(\phi \wedge K)$ to outcome $(\psi \wedge K)$, then, if K is the way the world actually is, the actual value of ϕ exceeds the actual value of ψ. Similarly, I think that if you regard $(\phi \wedge K)$ as on a par with $(\psi \wedge K)$, then, if K is the way the world actually is, the actual values of ϕ and ψ are on a par.

nor a negative amount to your estimate. Does it contribute *nothing*? No. Parity is *elastic:* if two outcomes are on a par, small improvements (in either direction) won't break the parity. On my proposal, hypotheses according to which your options are on a par contribute some "elasticity" to your estimate of the extent to which an option's actual value exceeds another.

What do I mean by *"elasticity"*? If A and B are on a par, there will be a range of improvements to A, and a range of diminishments to A, that will also be on a par with B; and likewise for B: there will be a range of improvements, and diminishments, that will be on a par with A. The parity between A and B is *maximally elastic* if no matter how much we improve (or diminish) one of them, you still regard them as on a par. Although there might be cases of parity which are maximally elastic, it's implausible that all cases are. (Surely you'd prefer the alpine ski vacation plus *a trillion dollars* to the beach vacation, for example!) In many cases, there are limits to the elasticity — we can place upper and lower bounds on the extent to which the outcomes are on a par. In particular, suppose that you strictly prefer prize A plus $\$y$ to prize B, and suppose that you strictly prefer prize B plus $\$x$ to prize A. Then, the extent to which A and B are on a par is bounded by the values you assign to those sums of money. We can interpret these bounds as placing limits on the extent to which the value of A fails to exceed the value of B (and *vice versa*).

$$-V(\$y) < \mathcal{CV}(A,B) < V(\$x)$$
$$-V(\$x) < \mathcal{CV}(B,A) < V(\$y)$$

If you would prefer B plus $\$x$ to A, then the extent to which A is on a par with B can't exceed $V(\$x)$; and if you would prefer A plus $\$y$ to B, then the extent to which B is on a par with A can't exceed $V(\$y)$. When A and B are on a par, your assessment of the extent to which A's value exceeds B's value reflects the extent to which the two are on a par by being *unsharp:* there is no precise amount such that the value of A exceeds, or falls short of, the value of B; rather, there is a range, or interval, of values capturing the extent to which the two are on a par.

Let $\lceil \mathcal{CV}_K(\phi,\psi)\rceil$ be the *least upper bound* on the extent to which the value of outcome $(\phi \wedge K)$ exceeds the value of outcome $(\psi \wedge K)$, and let $\lfloor \mathcal{CV}_K(\phi,\psi)\rfloor$

be the *greatest lower bound.*[29] If outcome $(\phi \wedge K)$ and outcome $(\psi \wedge K)$ are not on a par and there is a precise amount by which the former exceeds, or falls short of, the other, then $\lceil \mathcal{CV}_K(\phi, \psi) \rceil = \lfloor \mathcal{CV}_K(\phi, \psi) \rfloor$. If the two outcomes are on a par, then $\lfloor \mathcal{CV}_K(\phi, \psi) \rfloor < 0 < \lceil \mathcal{CV}_K(\phi, \psi) \rceil$. And, if you prefer $(\phi \wedge K)$ to $(\psi \wedge K)$, but there's no precise extent to which you prefer it, then $0 < \lfloor \mathcal{CV}_K(\phi, \psi) \rfloor < \lceil \mathcal{CV}_K(\phi, \psi) \rceil$. By allowing the comparisons of value to be imprecise in this way, we can develop a decision theory for agents with incomplete preferences that's inline with the **Actual Value Conception.**

To assume that $\lceil \mathcal{CV}_K(\phi, \psi) \rceil$ and $\lfloor \mathcal{CV}_K(\phi, \psi) \rfloor$ are well-defined is an unrealistic idealization. If you regard two goods to be on a par, we shouldn't expect there to be a precise amount of value such that it is the least amount that needs to be added to the one in order for it to be preferred to the other. I'll provisionally assume that these quantities *are* well-defined in order more easily state the proposal, and then show how this unrealistic idealization can be relaxed.

[29] If these outcomes are on a par (and the parity isn't maximally elastic), is it guaranteed that there will be a *least* upper bound and a *greatest* lower bound? I don't think so. There needn't be a precise amount of value such that were A improved by exactly that amount, nothing more and nothing less, you will strictly prefer it to B. This is a limitation of the way we're modeling parity, and it's a limitation that will be inherited by the decision theory proposed in the next section. However, this is a limitation it shares with *Supervaluational Expected Utility Theory.* In order to bring this out, consider a version of **Vacation Boxes** in which you believe that the coin which determined which prize is in which box is ever-so-slightly biased toward heads. Let your credence in Heads be .55, for example. According to *Supervaluational Expected Utility Theory,* you aren't required to prefer taking the Larger box to the Regular box so long as there is some utility-function in your representor which ranks taking the Regular box ahead of taking the Larger box. Let $u \in U$ be the utility-function that ranks the relevant outcomes as follows: $u\left(B^{+}\right) > u\left(B\right) > u\left(A^{+}\right) > u\left(A\right)$. Given your credences, u will rank taking the Regular box over taking the Larger box just in case $\frac{u(B)-u\left(A^{+}\right)}{u\left(B^{+}\right)-u(A)} > \frac{45}{55}$. Assuming that the value of receiving the \$1 sweetening is independent of the two prizes, u will rank the Regular box ahead of the Larger box just in case $u\left(B\right) - u\left(A\right) > 10$ (in general, if you believe the coin to be biased in favor of heads, $u\left(B\right) - u\left(A\right) > \frac{1}{Cr(H)-Cr(T)}$). So, according to *Supervaluational Expected Utility Theory,* you ought to prefer taking the Larger box just in case *every* $u^{*} \in U$ is such that $u^{*}\left(B\right) - u^{*}\left(A\right) \leq 10$. That holds just in case you strictly prefer A plus 10 value-points to B (and, more generally, you ought to prefer taking the Larger box just in case you prefer A plus $\frac{1}{Cr(H)-Cr(T)}$ to B). Now suppose that you become slightly more confident that the coin landed heads. Should you prefer taking the Larger box? Because there are determinate facts about which utility-functions are in your set, there must, likewise, be a determinate point at which *Supervaluational Expected Utility Theory* switches its recommendations — and, consequently, there must be a precise amount of value such that A improved by exactly that amount is preferred to B, and A improved by any less amount isn't preferred to B.

2.5.3 Actual Value Decision Theory

You should align your preferences over outcomes with your estimate of how the actual values of those options compare. Your best estimate of the extent to which the actual value of ϕ exceeds the actual value of ψ is the weighted average of all the various ways their actual values might compare, where the weights correspond to your credences in the hypotheses about those comparisons. However, if the actual values of your options might be on a par — so that there is no precise fact of the matter about the extent to which the actual value of one of the options exceeds, or falls short of, the actual value of the other — then your *estimate* might also fail to be precise.

Because we've placed upper and lower bounds on $\mathcal{CV}_K(\phi, \psi)$, we can, likewise, place bounds on your estimates in the following way:

$$\textsc{Estimate}\Big[\lceil \mathcal{CV}_{@}(\phi, \psi)\rceil\Big] = \sum_K Cr(K) \cdot \lceil \mathcal{CV}_K(\phi, \psi)\rceil$$

$$\textsc{Estimate}\Big[\lfloor \mathcal{CV}_{@}(\phi, \psi)\rfloor\Big] = \sum_K Cr(K) \cdot \lfloor \mathcal{CV}_K(\phi, \psi)\rfloor$$

You should prefer ϕ to ψ when, and only when, your estimate of the extent to which the actual value of ϕ exceeds the actual value of ψ is greater than zero. When your actual value estimate is interval-valued, you should prefer ϕ to ψ when, and only when, the *lower bound* of your estimate is greater than zero.[30] You should regard your options to be on par when, and only when, the lower bound of your estimate is less than zero and the upper bound is greater than zero. If both bounds are the same, and equal to zero, you ought to be indifferent between the two.

[30] Notice that $\lfloor \mathcal{CV}_K(\phi, \psi)\rfloor = -\lceil \mathcal{CV}_K(\psi, \phi)\rceil$, and $\lceil \mathcal{CV}_K(\phi, \psi)\rceil = -\lfloor \mathcal{CV}_K(\psi, \phi)\rfloor$. So, this is equivalent to saying that you should prefer ϕ to ψ when, and only when, $\textsc{Estimate}\big[\lceil \mathcal{CV}_{@}(\psi, \phi)\rceil\big] < 0$.

Actual Value Decision Theory: "Prefer option ϕ to option ψ when, and only when, the lower bound on your estimate of the extent to which the actual value of ϕ exceeds the actual value of ψ is greater than zero; the options are on a par when, and only when, the lower bound on your estimate is less than zero and the upper bound is greater than zero."

$$
\begin{aligned}
&\phi \succ \psi \text{ when, and only when } \textsc{Estimate}\left[\lfloor \mathcal{CV}_{@}(\phi,\psi)\rfloor\right] > 0\\
&\phi \approx \psi \text{ when, and only when } \textsc{Estimate}\left[\lfloor \mathcal{CV}_{@}(\phi,\psi)\rfloor\right] = 0 = \textsc{Estimate}\left[\lceil \mathcal{CV}_{@}(\phi,\psi)\rceil\right]\\
&\phi \bowtie \psi \text{ when, and only when } \textsc{Estimate}\left[\lfloor \mathcal{CV}_{@}(\phi,\psi)\rfloor\right] < 0 < \textsc{Estimate}\left[\lceil \mathcal{CV}_{@}(\phi,\psi)\rceil\right]
\end{aligned}
$$

As we've seen, even if it is more likely than not that the actual value of ψ exceeds the actual value of ϕ, if there's a sufficiently large enough chance that the actual value of ϕ might exceed the actual value of ψ to a great extent, then you should prefer ϕ to ψ. In estimating the actual value comparison between ϕ and ψ, the "losses" in some states can be outweighed by the "gains" in others. If outcome $(\phi \wedge K)$ and outcome $(\psi \wedge K)$ are on a par, then it's not the case that receiving either outcome constitutes a "loss" or a "gain." But, if we can place bounds on the extent to which these outcomes are on a par — that is, if we can say for each outcome, how much value would need to be added (or subtracted) in order for you to prefer (or disprefer) it to the other — then we can, in effect, also place bounds on how large, or small, the "gains" (or "losses") in the other states must be in order to outweigh the parity between outcomes in others. Just as the parity between two prizes can be overcome if one of the prizes is sufficiently improved, the "elasticity" of your actual value estimate, inherited from the parity of those options' outcomes, can be overcome if there's a sufficient chance that the actual value of one of your options might exceed the actual value of the other by a significant enough extent.[31]

[31] Here's one way to motivate part of what the decision theory says. Suppose you are deciding between option ϕ and ψ, which might, for all you know, be on a par. Now, let's introduce the following *"virtual"* option: ϕ^*, which is just like ϕ except that, in each state, it's outcome is augment by $\lceil \mathcal{CV}_K(\psi,\phi)\rceil$ (that is to say, if $\lceil \mathcal{CV}_K(\psi,\phi)\rceil > 0$, the outcome is improved by exactly that amount, and if $\lceil \mathcal{CV}_K(\psi,\phi)\rceil < 0$, it is diminished by exactly that amount). You ought to (weakly) prefer ϕ^* to ψ. Here's why. In those states in which the outcomes are *not* on a par, $\mathcal{CV}_K(\phi^*,\psi) = 0$ because the value of

2.5.4 The Three Desiderata

Actual Value Decision Theory satisfies the desiderata mentioned above: (1) it is a generalization of causal decision theory; (2) in **Vacation Boxes**, the proposal says that you should regard taking the Larger box and taking the Regular box as on a par; and (3) there are non-trivial cases in which the proposal *does* recommend preferring one option to another even though you regard some (but, crucially, not all) of the outcomes in the same states of the world to be on par.

A Generalization of Causal Decision Theory

This view, like *Supervaluational Expected Utility Theory*, is a generalization of Expected Utility Theory (in particular, this view generalizes causal decision theory). In other words, when your ends can be represented, in the canonical way, with a utility-function, this view will make the same recommendations as causal decision theory.

If your ends can be represented with a utility-function, then, for every dependency hypothesis K, $\lceil \mathcal{CV}_K(\phi, \psi) \rceil = \lfloor \mathcal{CV}_K(\phi, \psi) \rfloor = \mathcal{CV}_K(\phi, \psi)$.

Therefore, $\textsc{Estimate}\,[\lfloor \mathcal{CV}_@(\phi, \psi) \rfloor] > 0$ just in case $\textsc{Estimate}\,[\mathcal{CV}_@(\phi, \psi)] > 0$. And, as we saw in the first chapter, $\textsc{Estimate}\,[\mathcal{CV}_@(\phi, \psi)] > 0$ holds just in case the causal expected utility of ϕ exceeds the causal expected utility of ψ. Thus, **Actual Value Decision Theory** is a generalization of causal decision theory.

ϕ^*'s outcomes have been adjusted so as to equal the value of ψ's outcomes. In those states in which the outcomes *are* on a par, $\mathcal{CV}_K(\phi^*, \psi) > 0$ because (i) $\lceil \mathcal{CV}_K(\psi, \phi) \rceil$ is the smallest amount that $(\phi \wedge K)$ needs to be improved in order to be preferred to ψ, and (ii) $(\phi^* \wedge K)$ is just like $(\phi \wedge K)$ except that it's been improved by exactly that amount. So, you ought to prefer ϕ^* to ψ. But, also, if $\sum_K Cr(K) \cdot \lceil \mathcal{CV}_K(\psi, \phi) \rceil < 0$, you ought to prefer ϕ to ϕ^*. Our recipe for defining ϕ^* ensures that its outcomes are comparable to ϕ's in the same states. And, in particular, for each K, $\mathcal{CV}_K(\phi^*, \phi) = \lceil \mathcal{CV}_K(\psi, \phi) \rceil$. So, $\textsc{Estimate}[\mathcal{CV}_@(\phi^*, \phi)] = \sum_K Cr(K) \cdot \lceil \mathcal{CV}_K(\psi, \phi) \rceil < 0$. So, you ought to prefer ϕ to ϕ^*. Because your preferences ought to be transitive, you ought to prefer ϕ to ψ. This shows that, in general, if $\textsc{Estimate}\,[\lfloor \mathcal{CV}_@(\phi, \psi) \rfloor] > 0$, then you ought to prefer ϕ to ψ.

Parity in Vacation Boxes

According to **Actual Value Decision Theory**, you should regard the Larger box and the Regular box as on a par. Because A^+ is on a par with B and B^+ is on a par with A, the lower bound on your estimate of the comparison in actual values between the Larger box and the Regular box is less the zero, and the upper bound on your estimate is greater than zero.

$$\text{Estimate}\Big[\lfloor \mathcal{CV}_{@}(L,R)\rfloor\Big] < 0 < \text{Estimate}\Big[\lceil \mathcal{CV}_{@}(L,R)\rceil\Big]$$

$$\sum_K Cr(K)\cdot\lfloor \mathcal{CV}_K(L,R)\rfloor < 0 < \sum_K Cr(K)\cdot\lceil \mathcal{CV}_K(L,R)\rceil$$

Suppose that $\$y$ is the smallest amount that A^+ must be improved in order for it to be preferred to B; and suppose that $\$x$ is the smallest amount that B^+ needs to be improved to be preferred to A. It follows from this that $\$(y+2)$ is the smallest amount needed by A to be preferred to B^+, and that $\$(x+2)$ is the smallest amount needed by B to be preferred to A^+. These facts place bounds on the extent to which these outcomes are on a par:

Because $(A^+ + \$y) \succ B$, $\lceil \mathcal{CV}_H(R,L)\rceil = V(\$y)$

Because $(B^+ + \$x) \succ A$, $\lceil \mathcal{CV}_T(R,L)\rceil = V(\$x)$

Because $(A + \$(y+2)) \succ B^+$, $\lceil \mathcal{CV}_T(L,R)\rceil = V(\$(y+2))$

Because $(B + \$(x+2)) \succ A^+$, $\lceil \mathcal{CV}_H(L,R)\rceil = V(\$(x+2))$

We can use these quantities to arrive at the lower and upper bounds of your estimates.

$$\begin{aligned}\sum_K Cr(K)\cdot\lfloor \mathcal{CV}_K(L,R)\rfloor &= \frac{1}{2}\cdot -V(\$y) + \frac{1}{2}\cdot -V(\$x)\\ &= \frac{-(V(\$x)+V(\$y))}{2} < 0\end{aligned}$$

$$\begin{aligned}\sum_K Cr(K)\cdot\lceil \mathcal{CV}_K(L,R)\rceil &= \frac{1}{2}\cdot V(\$(x+2)) + \frac{1}{2}\cdot V(\$(y+2))\\ &= \frac{V(\$(x+2))+V(\$(y+2))}{2} > 0\end{aligned}$$

Because $V(\$y)$ and $V(\$x)$ are both greater than zero, the lower bound on your estimate is less than zero and the upper bound on your estimate is greater than zero. According to **Actual Value Decision Theory**, then, you shouldn't prefer taking the Larger box over the Regular box; the two options are on a par.

Going Beyond Parity

Consider the following decision-problem. There are two boxes: the Larger box and the Regular box. There is some chance, p, that L contains $\$z$ while R contains \$0; otherwise, a coin was flipped to determine whether prize A was placed in L and prize B in R or *vice versa.*

	K$_1$	K$_2$	K$_3$
L	A	B	$\$z$
R	B	A	\$0

If the chance, p, and the prize money, $\$z$, are large enough, then, according to **Actual Value Decision Theory**, rationality requires you to take the Larger box.

Suppose that $\$y$ is the smallest improvement to A to render it preferred to B, and that $\$x$ is the smallest improvement to B to render it preferred to A. These facts place bounds on the extent to which your outcomes are on a par.

$$\begin{array}{lll} \text{Because} & (A+\$y)\succ B, & \lceil \mathcal{CV}_{K_1}(R,L)\rceil = \lceil \mathcal{CV}_{K_2}(L,R)\rceil = V(\$y) \\ \text{Because} & (B+\$x)\succ A, & \lceil \mathcal{CV}_{K_2}(R,L)\rceil = \lceil \mathcal{CV}_{K_1}(L,R)\rceil = V(\$x) \\ \text{Because} & \$z\succ \$0, & \lceil \mathcal{CV}_{K_3}(L,R)\rceil = V(\$z) \end{array}$$

Your credence in K_3 is p, and your credences in K_1 and K_2 are both $\frac{1-p}{2}$. We can use these quantities to arrive at the lower bound on your estimate of the extent to which L has more actual value than R.

$$\sum_K Cr(K) \cdot \lfloor \mathcal{CV}_K(L,R) \rfloor = \frac{1-p}{2} \cdot -V(\$y) + \frac{1-p}{2} \cdot -V(\$x) + p \cdot V(\$z)$$

$$= p \cdot V(\$z) - \frac{1-p}{2} \cdot \Big(V(\$x) + V(\$y) \Big)$$

According to **Actual Value Decision Theory**, if $p \cdot V(\$z) - \frac{1-p}{2} \cdot (V(\$x) + V(\$y)) > 0$, then you are rationally required to prefer L to R.

$$p \cdot V(\$z) - \frac{1-p}{2} \cdot \Big(V(\$x) + V(\$y) \Big) > 0$$

$$p \cdot V(\$z) > \frac{1-p}{2} \cdot \Big(V(\$x) + V(\$y) \Big)$$

$$\frac{p}{1-p} > \frac{V(\$x) + V(\$y)}{2 \cdot V(\$z)}$$

$$\frac{p}{1-p} > \frac{V(\$x) + V(\$y)}{V(\$z) + V(\$z)}$$

If p and $V(\$z)$ are large enough, then this inequality will hold. For example, suppose that $p = \frac{1}{2}$ and that you prefer $\$z$ to both A and B. The lower bound on your estimate of the extent to which the actual value of L exceeds the actual value of R will be greater than zero just in case $2 \cdot V(\$z) > V(\$x) + V(\$y)$. And, because you prefer $\$z$, by itself, to B, $V(\$z) \geq V(\$y)$; and, because you prefer $\$z$, by itself, to A, $V(\$z) \geq V(\$x)$. So, $2 \cdot V(\$z) > V(\$x) + V(\$y)$. Therefore, according to **Actual Value Decision Theory**, you should prefer option L to option R. Although it's just as likely that your options are on a par as it is that the actual value of L exceeds the actual value of R, the extent to which the actual value of L might exceed the actual value of R is large enough to outweigh the elasticity of the parity between the outcomes in the other states.

$p = ?$	*verdict*
0	$L \bowtie R$
$\frac{1}{2}$	$L \succ R$
1	$L \succ R$

The interesting cases are when $0 < p < \frac{1}{2}$. For that range of credences, whether you are rationally required to take L over R depends on the extent to which your preference for \$$z$ over \$0 is greater than the extent to which you regard prizes A and B as on a par. And we needn't expect there to be a precise fact of the matter about this.

2.5.5 Relaxing the Unrealistic Assumption

As mentioned above, it's unrealistic to assume that $\lceil \mathcal{CV}_K(\phi, \psi) \rceil$ and $\lfloor \mathcal{CV}_K(\phi, \psi) \rfloor$ are well-defined. There needn't be a *least* upper bound, nor a *greatest* lower bound, on the extent to which the value of outcome $(\phi \wedge K)$ exceeds the value of outcome $(\psi \wedge K)$ if you regard these outcomes as on a par.

It's *not* unrealistic to assume, however, that you can place *some* (upper and lower) bounds on $\mathcal{CV}_K(\phi, \psi)$. In order to relax the unrealistic assumption, then, replace $\lceil \mathcal{CV}_K(\phi, \psi) \rceil$ in the formulation above with *some* upper bound, and replace $\lfloor \mathcal{CV}_K(\phi, \psi) \rfloor$ with *some* lower bound. If your estimate of the actual value comparisons between ϕ and ψ *using these bounds* is greater than zero, then rationality requires you to prefer ϕ to ψ. Why? If this estimate is greater then zero, then there's some lower bound on the estimate of the extent to which ϕ's actual value might exceed ψ's which is greater than zero. But if this lower bound is greater than zero, then, the *greatest* lower bound on your estimate of the extent to which ϕ's actual value might exceed ψ's — if it were to exist — would, also, be greater than zero.[32] In this manner, we can relax the unrealistic assumption that there are precise bounds on the extent to which outcomes are on par, while retaining sufficient conditions for when rationality requires you to prefer one option to another.

[32] Alternatively, because $\lfloor \mathcal{CV}_K(\phi, \psi) \rfloor = -\lceil \mathcal{CV}_K(\psi, \phi) \rceil$, if your estimate of the extent to which the actual value of ψ might exceed the actual value of ϕ using some upper bounds is *less* than zero, then you should prefer ϕ to ψ. This is because, if there's an upper bound on your estimate of the extent to which ψ's actual value might exceed ϕ's that's less than zero, then the *least* upper bound — if it were to exist — would, also, be less than zero.

> *Sufficient Conditions for Preferring ϕ to ψ.*
>
> If your lower bound estimate of the extent to which ϕ's actual value might exceed ϕ's actual value is greater than zero, then rationality requires you to prefer ϕ to ψ.

We can work towards providing necessary conditions for preferring one option to another (as well as sufficient conditions for regarding two options as on a par) by looking at the upper and lower bounds on $\lceil \mathcal{CV}_K(\phi, \psi) \rceil$ and $\lfloor \mathcal{CV}_K(\phi, \psi) \rfloor$ themselves. If you prefer outcome $(\psi \wedge K)$ sweetened by $\$u$ to outcome $(\phi \wedge K)$, then $V(\$u)$ is an upper bound on $\lceil \mathcal{CV}_K(\phi, \psi) \rceil$. And, if you regard $(\psi \wedge K)$ sweetened by $\$l$ as on par with $(\phi \wedge K)$, then $V(\$l)$ is a lower bound on $\lceil \mathcal{CV}_K(\phi, \psi) \rceil$. In other words, the least upper bound on the extent to which the value of $(\phi \wedge K)$ exceeds the value of $(\psi \wedge K)$ — were it to exist — can be approached from above or below.

> *Necessary Conditions for Preferring ϕ to ψ.*
>
> Rationality requires you to prefer ϕ to ψ only if your lower-bound-on-the-least-upper-bound estimate of the extent to which ψ's actual value might exceed ϕ's actual value is less than zero.

If the least upper bound on your estimate of the extent to which ψ's actual value might exceed ϕ's actual value — were it to exist — is less than zero, then your lower-bound-on-the-least-upper-bound estimate must also be less than zero. Because, according to **Actual Value Decision Theory**, you should prefer ϕ to ψ only when $\textsc{Estimate}\big[\lceil \mathcal{CV}_{@}(\psi, \phi) \rceil\big] < 0$, it follows that you should prefer ϕ to ψ only if the lower-bound-on-the-least-upper-bound estimate is less than zero. In a similar fashion — by making use of the upper and lower bounds that you can place on $\lceil \mathcal{CV}_K(\phi, \psi) \rceil$ and $\lfloor \mathcal{CV}_K(\phi, \psi) \rfloor$ — we can provide a sufficient condition for regarding your options as on a par.

> *Sufficient Conditions for Parity Between ϕ and ψ.*
>
> If (i) your upper-bound-on-the-greatest-lower-bound estimate of the extent to which ϕ's actual value might exceed ψ's actual value is less than zero, and (ii) your lower-bound-on-the-least-upper-bound estimate of the extent to which ϕ's actual value might exceed ψ's actual value is greater than zero, then you ought to regard ϕ and ψ as on a par.

We have, then, sufficient conditions for preferring one option to another and sufficient conditions for regarding the two as on a par. But there's no guarantee that these sufficient conditions will be met in all cases — it will depend on the (upper and lower, and upper-on-lower and lower-on-upper) bounds that you place on the value-comparisons between the outcomes of your options. If none of the sufficient conditions are met, then **Actual Value Decision Theory** is silent: it says nothing at all about what rationality requires or permits you to do.

What do I mean by *"silent"*? Again, think of a decision theory like a helpful advisor who, given information about your perspective and your aims, issues recommendations about what you rationally ought to do. The advisor might say, of some particular option, that you are rationally required to take it. Or, it might say that there are several options, each of which it is rationally permissible to take. But, also, it might remain silent: it might say nothing at all about what rationality requires, or permits, you to do. If your advisor isn't given enough information about your perspective and your aims, we shouldn't expect her to be able to help you. If you are unable to place informative enough (upper and lower, and upper-on-lower and lower-on-upper, etc.) bounds on the extent to which the outcomes of your options are on a par, **Actual Value Decision Theory**, like the helpful advisor, can't help you. There's simply no fact of the matter about what rationality requires of you. And that, I think, is exactly right.

2.6 Conclusion

In the previous chapter I argued that causal decision theory is equivalent to a particular way of cashing out the **Actual Value Conception**, but that evidential decision theory is inconsistent with it. Moreover, I argued that the arguments causal decision theorists have given for Two-Boxing in the Newcomb Problem implicitly gain their plausibility from the **Actual Value Conception**. In this chapter, I've demonstrated that the same idea also underlies a different, seemingly unrelated, issue: how to evaluate your options when you lack complete preferences. I've argued that *Supervaluational Expected Utility Theory* is, like evidential decision theory, inconsistent with the **Actual Value Conception.** Furthermore, the arguments that have been deployed against *Supervaluational Expected Utility Theory*, much like the arguments (which, not coincidentally, even have the same names) deployed against One-Boxing in the Newcomb Problem, implicitly gain their support from the **Actual Value Conception.**

Now, it's true that one can, without contradiction, endorse *both* Two-Boxing in the Newcomb Problem *and* preferring the Larger box to the Regular box in **Vacation Boxes**. Causal decision theory is not inconsistent with *Supervaluational Expected Utility Theory*; one could endorse both views if one wanted to. However, what I hope to have shown is that doing so lacks motivation. The reasons for favoring causal decision theory over evidential decision theory in the Newcomb Problem are also reasons to reject *Supervaluational Expected Utility Theory.*

Although I've argued against several of the considerations in favor of *Supervaluational Expected Utility Theory*, aside from showing that it's inconsistent with the **Actual Value Conception**, I've given little independent reason to reject it. I've said next to nothing about why one shouldn't simply abandon the **Actual Value Conception** in favor of *Supervaluational Expected Utility Theory.*

Before moving on to the next chapter, however, I do want to present an argument, not against *Supervaluational Expected Utility Theory* itself, but against it conjoined with evidential decision theory: call it *Supervaluational Evidential Expected Utility Theory.* This view holds that you should prefer an option ϕ to an option ψ when, and only when, every utility-function in your representor ranks the expected utility of ϕ ahead of the expected utility of ψ, where these expected utilities are evidential expected utilities (that is, they're calculated with respect to your *conditional* credences). The argument makes use of the following example.

> **The Newcomb Vacation Box.** There is a single box in front of you. The box either contains prize A or prize B. You have two options: you can take the box for free, or you can pay \$1 for it. But there's a twist. Whether the box contains prize A or prize B was determined by a super-reliable predictor. The predictor

placed A in the box if she predicted that you'd take the box for free; she placed B in the box if she predicted you'd pay the dollar.

	PREDICTS: "FREE"	PREDICTS: "PAY"
take for free	A	B
pay \$1	A^-	B^-

According to *Supervaluational Evidential Expected Utility Theory*, it's permissible to pay \$1 for the box even though you could take the box for free. Here's why. Suppose the predictor is 99% reliable. It's permissible to pay \$1 if there is a utility-function in the representor such that paying \$1 is ranked ahead of taking the box for free. Consider the utility-function in your representor, u, which ranks the prizes as follows: $u(B) > u(B^-) > u(A) > u(A-)$.

$$.01 \cdot u(A^-) + .99 \cdot u(B^-) > .99 \cdot u(A) + .01 \cdot u(B)$$

$$.99\Big(u(B^-) - u(A)\Big) > .01\Big(u(B) - u(A^-)\Big)$$

$$\frac{u(B^-) - u(A)}{u(B) - u(A^-)} > \frac{1}{99}$$

$$\frac{u(B) - u(A) - u(\$1)}{u(B) - u(A) + u(\$1)} > \frac{1}{99}$$

So long as there is a utility-function in your set such that $u(B) - u(A) > \frac{100}{98} \approx 1.0204$, there will be a utility-function that ranks paying the \$1 ahead of taking the box for free, given that the expected utilities are calculated evidentially. If there is no utility-function in your set like this, then every utility-function in your set must rank A sweetened with $\$1\frac{2}{98}$, or more, ahead of B; in which case, you must prefer $(A+\$1\frac{2}{98})$ to B, meaning that A and B, while on a par, are only slightly so. Therefore, as long as the parity between A and B is more elastic than that, *Supervaluational Evidential Expected Utility* recommends paying the dollar over taking the box for free.

But it seems (to me anyway) absurd to pay \$1 for the box when you could take it for free! As in the Newcomb Problem, you are in a position to know that the actual value of taking the box for free exceeds the actual value of paying a $1 for it. Unlike in the Newcomb Problem, however, the followers of *Supervaluational Evidential Expected Utility Theory* aren't likely to leave the room "richer" than their non-evidential counterparts.

This isn't a knockdown argument, of course. A dyed-in-the-wool evidentialist will be happy to say that it's okay to pay $1 for the box (even though you know that doing so has less actual value than taking the box for free). But, if taking it to be permissible to pay the $1 for the box makes you feel uncomfortable, you should either reject evidential decision theory or reject *Supervaluational Expected Utility Theory.* If you reject evidential decision theory, in favor of causal decision theory, though, then you also have good reason to reject *Supervaluational Expected Utility Theory.* Conversely, if you reject *Supervaluational Expected Utility Theory,* in favor of the alternative offered here, then you also have good reason to reject evidential decision theory.

To sum up, I've argued that the **Actual Value Conception** is incompatible with *Supervaluational Expected Utility Theory.* The latter sometimes recommends having instrumental preferences that, according to the former, you aren't rationally required to have.

Chapter 3

Reasons, Rationality, and Agglomeration

3.1 Introduction

In the previous chapters, I developed a picture of instrumental rationality — the **Actual Value Conception** — and argued that it is incompatible with both evidential decision theory and *Supervaluational Expected Utility Theory.* I presented a decision theory for agents with incomplete preferences that's motivated by the **Actual Value Conception**, and showed, among other things, that it is a generalization of causal decision theory.

As things stand, we have two competing pictures of instrumental rationality — the **Actual Value Conception** and *Prospectism* — and, for each picture, a corresponding decision theory for agents with incomplete preferences.

Picture of Instrumental Rationality	Decision Theory
Actual Value Conception	*Actual Value Decision Theory*
Prospectism	*Supervaluational Expected Utility Theory*

I think the **Actual Value Conception** is correct, but I've offered little in the way of argument for it. In this chapter, I will defend the view from some arguments against it and draw out some of its interesting consequences. I'll address two arguments against the **Actual Value Conception.**[1]

The first is the Most Reason Argument. It goes like this: in **Vacation Boxes**, you have more reason to prefer taking the Larger box over the Regular box; and, being rational involves doing what you have the most reason to do; and so, rationality requires you to prefer the Larger box to the Regular box, contrary to what **Actual Value Decision Theory** recommends.

The second is the Diachronic Agglomeration Argument. It holds that there are diachronic decision-problems about which **Actual Value Decision Theory** will say, for each action in a sequence of actions, that you are rationally required to prefer it to its alternatives, but that you are not rationally required to prefer performing that sequence of actions to *its* alternatives.

Although each of these arguments against **Actual Value Decision Theory** is powerful, I will argue that none of them are fatal. These arguments do, however, bring to light some interesting consequences of taking the **Actual Value Conception** seriously. In particular, (1) we need to give up the idea that rationality consists in doing what you have the most reason to do; and (2) we need to allow you to have rational non-transitive instrumental preferences.

3.2 Most Reason Argument

The *Most Reason Argument*, presented in [Hare, 2010, 2013], concludes that in **Vacation Boxes** you are rationally required to prefer the Larger box (L) to the Regular box (R). If it's sound, then **Actual Value Decision Theory**, which holds that you are *not* rationally required to prefer L to R, is false.

The argument goes like this: you have a reason to prefer L over R, but you have no reason to prefer R over L; furthermore, rationality requires that you do what you have the most reason to do; so, you are rationally required to

[1] In the appendix, I briefly discuss an additional argument — the In-the-Long-Run Argument: **Actual Value Decision Theory** won't always require you to prefer one option to another even when, by your own lights, you are almost certain that the former will do better "in the long run" than the latter.

prefer L to R.

You do know that *there are reasons* to prefer R over L, but you don't specifically know what those reasons are; so, you don't *have* a reason for preferring R over L. The reason that you have for taking L over R is that you'll get a dollar. Because you have no reason to prefer R to L and you have some reason to prefer L to R, you have *most* reason to prefer L to R.

REASON ARGUMENT

P1 You have a reason to prefer taking the Larger box over taking the Regular box.

P2 You have no reason to prefer taking the Regular box over taking the Larger box.

P3 If you have a reason to prefer ϕ over ψ, and you have no reason to prefer ψ over ϕ, then you have *more* reason to prefer ϕ to ψ than you do to prefer ψ to ϕ.

P4 [*More Reason Principle*] If you have more reason to prefer ϕ to ψ than you do to prefer ψ to ϕ, then rationality requires you to prefer ϕ to ψ.

C You are rationally required to prefer taking the Larger box to taking the Regular box.

In order to asses **P1** and **P2**, we need to say more about what it is to *have a reason* to prefer one thing to another.

3.2.1 Having Reasons: defending P1

I'm going to defend premise **P1** of the *Reason Argument* — which says that you have a reason to prefer taking the Larger box over taking the Regular box — by clarifying what it is, in the relevant sense, to *have a reason.*

There's a great deal of controversy about what it is for something to be a reason, and what it takes for you to have them.[2] I will try to avoid controversy

[2] Some philosophers think that the notion of a *reason* is unanalyzable ([Scanlon, 2014], [Raz, 1999]). Others, like Broome [2007, 2013] for example, understand reasons to be facts that, in some sense, *explain* why it is that you ought to do what you ought to do in a

as far as possible (although, as we'll see, it won't be easy to be completely neutral) by saying no more about what it is to "have a reason" than is necessary to present the most plausible version of the REASON ARGUMENT.

Very roughly: a reason for you to ϕ is a consideration that "counts in favor" of you ϕing.[3] If $\mathbf{r}$ is a reason for you to ϕ, then $\mathbf{r}$ provides some (*prima facie*) *justification* for ϕing. There are reasons for you to ϕ just in case there are some facts that speak in favor of you ϕing. In order for one of the reasons that *there are* for you to ϕ to be a reason you *have* to ϕ, you must be aware of it and disposed to be motivated by it.[4] Think of the reasons that you have to ϕ as those considerations that you could cite as "pros" for ϕing were you to make a list of "pros & cons" about what to do.

More specifically, what is it for you to have a reason to prefer ϕ to ψ? Let's say that a fact $\mathbf{r}$ is a reason for you to prefer ϕ to ψ just in case:

(1) $\mathbf{r}$ is of the form $\ulcorner$if ϕ, then p & if ψ, then $\neg p\urcorner$ and

(2) You consider p to be a good thing.[5]

particular situation. Still others, like Setiya [2007, 2014] and Smith [1987, 1995], analyze reasons in terms of *ideal deliberation:* roughly, a reason to ϕ is a fact such that (i) were you to be aware of it and (ii) ideally rational (and, perhaps, ideal in some other respects as well), it would provide you with some motivation to ϕ.

[3] This is a very rough characterization of what it is for something to be a *normative* reason. We can draw a distinction between two different kinds of reasons: *motivating* reasons and *normative* reasons. Motivating reasons are the reasons *why* you did one thing and not another. They are, roughly, the considerations that, as a matter of psychological fact, motivated you to do what you did. Normative reasons, on the other hand, provide support for doing one thing over another. Motivating reasons play a role in explaining your behavior (or attitudes), whereas normative reasons play a role in explaining how you *ought* to behave (or what attitudes you *ought* to adopt).

[4] This rough characterization of what it is to *have* a reason — namely, that *there be* a reason of which you are aware, and disposed to be motivated by — is by no means uncontroversial. Schroeder [2008], for example, argues that something can be a reason you *have* without it being a reason *there is.* The argument appeals to examples in which you have false (but reasonable) beliefs, the contents of which *would* be a reason for you if only it were true. (See [Lord, 2010] for a rebuttal). Because the REASONS ARGUMENT is being applied to a case in which you have no (relevant) false beliefs, I will be ignoring this complication from here on out.

[5] What is it to "consider p to be a good thing"? Clause (2) is meant to capture the idea that the reason $\mathbf{r}$ counts in *favor* of ϕing. The reader should feel free to understand clause (2), or to revise it, in whatever way they most prefer so long as it captures the same idea. For example, the reader may substitute 'good' for 'valuable' or 'desirable', or instead say 'You'd welcome the news that p', or 'You prefer p to $\neg p$', etc. Some philosophers (notably, Scanlon [1998]) endorse the so-called *Buck-Passing Account of Value*, according to which evaluative notions — like, *good* or *valuable* or *desirable* — are to be understood as the property of having other ("first-order") properties that provide reasons to hold certain

A reason $\mathbf{r}$ is a reason that you *have* just in case:

(3) You are in a position to be aware that $\mathbf{r}$ is a reason for you.[6]

Why clause (1)? In order for $\mathbf{r}$ to be a reason for you to prefer ϕ to ψ, it has to be a consideration that counts more strongly in favor of ϕ than it does for ψ. If you'll give me a high-five no matter what I do, that won't provide me with a reason to prefer one of my options to another. Only considerations that distinguish between ϕ and ψ can be reasons for me to prefer one over the other.

Why clause (2)? Reasons are considerations that count in favor of something. Clause (2) — that you take p to be something good — is meant to capture the idea that $\mathbf{r}$ speaks in favor of, and not against, preferring ϕ to ψ.[7]

Premise **P1** says that you have a reason to prefer taking the Larger box over taking the Regular box. Is there a consideration of which you're aware that counts in favor of you taking the Larger box over the Regular box? Yes — if you take the Larger box, you will get $1; you won't get the dollar if you take the Regular box; you (*ceteris paribus*) consider getting $1 to be a good thing; and you are in a position to know all this. Therefore, the fact that

pro-attitudes toward the object being evaluated. (For example, if you consider p to be a good thing, there are features had by p that provide you with *reasons* to value it). If the *Buck-Passing Account* is correct, the characterization being offered here looks to be problematically circular — reasons are analyzed in terms of the good, and the good is analyzed in terms of reasons, and around and around we go. For all I know, the *Buck-Passing Account* is correct (although, see [Bykvist, 2009], [Zimmerman, 2007] for some reasons to think it isn't). Let this not detain us, however, as nothing much will end up turning on the specific formulation of clause (2).

[6] The account being sketched here bears more than a passing resemblance to the account in [Bales et al., 2014]. Instead of (2), they say: you prefer p to $\neg p$. They go on to point out that, because p picks out a sub-maximal state of affairs, there are various ways of cashing out what it is to "prefer" p to $\neg p$, settling on an account according to which you prefer p to $\neg p$ just in case you prefer every p-world to the $\neg p$-world that is, in all respects expect for the truth or falsity of p, exactly like it.

[7] A quick point of clarification. It might be objected that, while (1) and (2) might provide sufficient conditions for a fact to be a reason to prefer one thing to another, they aren't necessary conditions. I might have all sorts of reasons to prefer ϕ to ψ, none of which having much to do with ϕ or ψ themselves. For example, suppose that a brilliant neuroscientist, who's developed a fool-proof way of determining people's preferences, offers to pay me a million dollars if I prefer ϕ to ψ. The fact that I will get a million dollars if I prefer ϕ to ψ (and that I won't get a million dollars if I don't) seems very much like a reason I have to adopt, if I can, this preference. (Although, I'm inclined to say that this is a reason for you to *want* to prefer ϕ to ψ rather than a reason to prefer ϕ to ψ itself). We should understand (1) and (2) as characterizing your *object*-given, as opposed to *state*-given, reasons [Parfit, 2001].

you'll get a dollar if you take the Larger box but won't get a dollar if you take the Regular box is a reason you have for preferring the Larger box to the Regular box.

Before moving on, let me quickly address one way that premise **P1** could be challenged: one might complain that, in this case, it's less-than-obvious that you should consider getting a $1 to be a good thing; and, therefore, the fact that you'll get a dollar if you take the Larger box and won't get a dollar if you take the Regular box is *not* a reason for you to prefer the Larger box to the Regular box.

Here's the idea. Even though getting a dollar is *usually* good, it isn't always. For example, suppose that you know you'll get a dollar if you ϕ and that if you ψ you won't. And, furthermore, suppose you know that getting an additional dollar will bump you into a higher tax bracket (or, more fancifully, that there's a madman on the loose attacking all and only those who have very recently received a dollar). In a case like this, the fact that you'll get a dollar if you ϕ and won't if you ψ doesn't seem like a reason for you to ϕ rather than ψ (rather, it seems like a reason to ψ rather than ϕ). It's not a reason because, in these cases, you don't prefer getting the dollar to not getting it — you don't consider getting the dollar to be a good thing.

Is **Vacation Boxes** a case in which you should consider getting a dollar to be a good thing? Do you prefer getting the dollar to not getting it? Well, you don't prefer prize A plus a dollar to prize B, and you don't prefer prize B plus a dollar to prize A. And you know that either the Larger box contains A and the Regular box B, or *vice versa.* So, you know that, in this case, each of the two ways of getting a dollar aren't preferable to the corresponding ways of not getting a dollar.[8] But, even still, the fact that you'll get a dollar if you

[8] Bales et al. [2014] make a similar point in response to the REASONS ARGUMENT. They argue that if what it is to prefer p to $\neg p$ (in the relevant sense) is to prefer every p-world to every $\neg p$-world, then premise **P1** is false — getting a dollar doesn't provide you with a reason to prefer L to R because there are some ways of getting a dollar that you don't prefer to some of the ways of not getting a dollar. However, as they note, this isn't a very plausible account of what it is to prefer p to $\neg p$ (in the relevant sense) because, if it were true, reasons would be hard to come by: logical space is vast and, for nearly any p, there will be some very excellent $\neg p$-worlds and some very not-excellent p-worlds. They go on to offer a more plausible analysis: you prefer p to $\neg p$ just in case, *ceteris paribus*, you all else equal prefer p to $\neg p$. You *all else equal* prefer p to $\neg p$ just in case each p-world is preferred to the $\neg p$-world that is, in all relevant respects aside from the truth of p, exactly like it. On this analysis, premise **P1** is true. However, Bales et al. [2014] still think the

take it *is* a consideration in favor of taking the Larger box. Although you don't prefer *each possible way* of getting a dollar to *every way* of not getting a dollar, you *do* prefer, *ceteris paribus*, getting a dollar to not getting one *holding all else equal.* And that's enough. Getting a dollar, in this case, is a good thing, and you should regard it as such. After all, in deciding between, e.g., A^+ and B, the fact that you'll get a dollar if you take the former but not the latter surely is a reason to prefer the former to the latter, even if it's not a decisive reason. Likewise, the fact that you'll get a dollar if you take the Larger box but not the Regular box, is a reason to prefer the Larger box to the Regular box.

Let's turn to premise **P2.** Is it true that you have no reason to prefer taking the Regular box over taking the Larger box?

3.2.2 Having No Reason: assessing P2

The second premise of the REASONS ARGUMENT says that you have no reason to prefer taking the Regular box over taking the Larger box. In other words, there are no considerations of which you are aware that tell in favor of taking the Regular box over the Larger box. I will argue that premise **P2** is false: you, in fact, *do* have a reason to prefer taking the Regular box over taking the Larger box. But, because of the nature of this reason, the spirit of the REASON ARGUMENT can be resuscitated.

Why think that premise **P2** is true? Roughly, the idea is that, although there are reasons to prefer taking the Regular box to the Larger box, because of your ignorance concerning which prize is in which box, you fail to *have* a reason to prefer taking the Regular box to the Larger box.

It's true that as a matter of fact *there are* reasons for you to take the Regular box rather than the Larger box. Suppose, for example, that the coin landed heads: the Larger box contains A^+ and the Regular box contains B. Now consider a feature had by B not had by A^+ — a proposition true in outcome

argument is defective because "*I know that other things are not equal.* Indeed, I know that I will only obtain the additional dollar at the cost of forgoing a good of great value to me." [Emphasis in the original]. It's not obvious to me what about this makes the argument defective.

B that's not true in outcome A^+ — that you consider to be a good thing. Here's an example:

> q = *if I take the Regular box, I can take a relaxing nap in the sun & if I take the Larger box, I cannot take a relaxing nap in the sun.*

It's true — again, supposing that as a matter of fact the coin landed heads — that q is a reason for you to take the Regular box over the Larger box. But because you don't know how the coin actually landed, while q *is* a reason for you, it's not a reason you *have:* you aren't in a position to know that q is a reason for you to take the Regular box over the Larger box. You, of course, *are* in a position to know that *there are* reasons for you to take the Regular box over the Larger box, but none of these reasons are reasons you *have.*

Moreover, it doesn't follow from the fact that you know that *there is* a reason to favor ϕ over ψ that you, thereby, *have* a reason to favor ϕ over ψ. If you know that there is a reason to prefer ϕ over ψ, then it's true that, for every possible reason **p, q, r,** etc., you know that *either* p is a reason to prefer ϕ over ψ, *or* q is a reason to prefer ϕ over ψ, *or* ..., but a disjunction of propositions of the form ⌜ *p is a reason to prefer ϕ over ψ* ⌝ is not itself a reason. This is because, even if p is reason for you to prefer ϕ over ψ, the proposition that *p is reason for you to prefer ϕ over ψ* is not itself a reason to prefer ϕ over ψ. Why? Because if p is a reason for you to prefer ϕ to ψ, then p will be a reason for you to prefer ϕ to ψ whether or not ψ is true. Therefore, the fact that p is a reason will not itself be a reason for you to prefer ϕ to ψ because it fails to distinguish between ϕ and ψ, in violation of clause (1).[9]

What are some other candidate reasons that you might be said to have for preferring the Regular box? You know that if you take the Regular box, you might get prize A. But that cannot be a consideration that speaks in favor

[9] Some philosophers, notably Schroeder [2009], have defended the idea that if there is a reason for you to ϕ, then the fact that there is a reason for you to ϕ is *itself* a reason for you to ϕ. If this idea is correct, and you know — as you do in **Vacation Boxes** — that there are reasons for you to prefer taking the the Regular box to the Larger box, then don't you have reason to prefer taking the Regular box to the Larger box after all? Not necessarily. Even if the fact that there are reasons to take the Regular box is, itself, a reason for you to take the Regular box, it's not clear that it provides you with a reason to prefer taking the Regular box *over taking the Larger box* given that there are also reasons to take the Larger box.

of taking the Regular box *rather than* the Larger box because you also know that by taking the Larger box, you might get prize A. Similarly for prize B: you know that if you take the Regular box, you might get prize B; but you know this is also true if you take the Larger box. You know that if you take the Regular box, you'll get a prize that was inside a regular-sized box; and that won't be true if you take the Larger box. However, you don't care about that — you don't consider *getting a prize that was inside a regular-sized box* to be a good thing (it's neither here nor there, goodness-wise) — so it's not a consideration that speaks *in favor* of taking the Regular box. What about the fact that you know there are uniquely good things about taking the Regular box? Again, that, too, isn't a reason to prefer taking it over taking the Larger box because you also know there are uniquely good things about taking the Larger box.

It doesn't seem like there are *any* considerations of which you are aware that, both, speak in favor of taking the Regular box and distinguish between taking it and taking the Larger box.

This conclusion can be resisted, however. Let's introduce a name 'the R-prize' that, by stipulation, picks out whichever of the two prizes is, as a matter of actual fact, in the Regular box. (So, for example, if the coin landed heads, then 'the R-prize' picks out prize B; and if the coin landed tails, then 'the R-prize' picks out prize A.) Here, then, is a consideration of which you are aware that counts in favor of taking the Regular box over the Larger box:

> $\mathbf{r}_R$ = *if I take the Regular box, I'll get the R-prize & if I take the Larger box, I won't get the R-prize.*

You consider getting the R-prize to be a good thing. You are in a position to know all this. So, **r** is a reason you have to prefer taking the Regular box to taking the Larger box. And, thus, premise **P2** is false.

As stated, the REASON ARGUMENT is unsound because its second premise is false. However, its *spirit* lives on. Although you (in some sense) have a reason to prefer taking the Regular box to taking the Larger box, there is a perfectly analogous reason that you have to prefer taking the Larger box to taking the Regular box — namely: that if you take the Larger box, you'll get the L-prize; but if you take the Regular box, you won't get the L-prize (where 'the L-prize' is a name you introduce to pick out whichever prize is in the Larger box).

> $\mathtt{r}_L$ = *if I take the Larger box, I'll get the* L*-prize & if I take the Regular box, I won't get the* L*-prize.*

It appears as though *this* reason perfectly balances the other. Furthermore, the dollar provides you with an additional reason to prefer the Larger box. And if you have more reason to prefer one thing to another, you are rationally required to adopt that preference.

More Reason Argument

P1* You have more reason to prefer the Larger box to the Regular box than you do to prefer the Regular box to the Larger box.

P4 [*More Reason Principle*] If you have more reason to prefer ϕ to ψ than you do to prefer ψ to ϕ, then rationality requires you to prefer ϕ to ψ.

C You are rationally required to prefer taking the Larger box over taking the Regular box.

But is premise **P1*** true? Do you really have *more* reason to prefer the Larger box to the Regular box?

Perfectly Balanced Reasons. It's true that you have, in some sense, *two* reasons to prefer the Larger box to the Regular box (i.e., that you will get the L-prize, and that you will get a dollar) and only *one* reason to prefer the Regular box to the Larger box (i.e., that you will get the R-prize). But it's not obvious that the reasons in favor of taking the Larger box *outweigh* the reasons for taking the Regular box. Here's an example to bring out the unobviousness. Consider a choice between prize B and prize A plus a dollar. You have reasons to prefer B to A^+ (e.g., the way the sand will feel beneath your toes, the soothing song of the ocean's tides, bottomless Mai Tai's at the Tiki Bar, etc.), but you also have reasons to prefer A^+ to B (e.g., the crisp mountain air, the adrenaline rush of ripping down a black diamond, bottomless hot chocolate at the lodge, etc.). Furthermore, you have an additional reason to take A^+ (you'll get a dollar). But it's not the case that you have *more* reason to prefer A^+ to B than you do to prefer B to A^+. Reasons are to be *weighed* not merely *counted.* And, on balance, your reasons do not weigh more heavily toward A^+ than toward B, and *vice versa.*

However, you *would* have more reason to prefer taking the Larger box to taking the Regular box than to prefer the Regular box to the Larger box if $\mathbf{r}_R$ (your R-prize reasons) and $\mathbf{r}_L$ (your $\mathtt{L} - \mathit{prize}$ reasons) are perfectly balanced. What is it for a reason to perfectly balance another? Here's a suggestion:[10]

> [Perfectly Balanced Reasons]
>
> A reason $\mathbf{r}_1$ for preferring ϕ to ψ *perfectly balances* a reason $\mathbf{r}_2$ for preferring ψ to ϕ just in case it would be a rational mistake to act on the basis of $\mathbf{r}_1$ while in full recognition of $\mathbf{r}_2$.

Two reasons are, let's say, *incommensurable* when neither reason is stronger than the other and it needn't be a mistake to act on the basis of the one while in full recognition of the other. What does it mean for it to be "a rational mistake" to act on the basis of reason while in full recognition of the other? If $\mathbf{r}_1$ and $\mathbf{r}_2$ are perfectly balanced, then $\mathbf{r}_1$ cannot *justify* your choosing ϕ over ψ. Here's an example. Suppose that if you ϕ you'll get a dollar, and if you ψ you'll get a different dollar.

> Let $\mathbf{r}_1$ = *if I ϕ, I will get dollar #1 & if I ψ, I will not get dollar #1.*
>
> Let $\mathbf{r}_2$ = *if I ψ, I will get dollar #2 & if I ϕ, I will not get dollar #2.*

Reason $\mathbf{r}_1$ is a reason for you to prefer ϕ to ψ and reason $\mathbf{r}_2$ is a reason for you to prefer ψ to ϕ. Suppose you decide to ϕ. If what ultimately motivated you to do so (or, if your justification for doing so) is $\mathbf{r}_1$, then you've made a rational mistake. You, we are assuming, are completely *indifferent* between getting dollar #1 and getting #2, so it would be irrational of you to be moved by consideration $\mathbf{r}_1$ when you are fully aware of consideration $\mathbf{r}_2$. On the other hand, consider choosing between prize A and prize B.

[10] This idea is inspired by the discussion of parity and choice in [Chang, 2009].

> Let $\mathbf{r}_A$ = *if I choose A, I will get to enjoy the crisp mountain air & if I choose B, I will not get to enjoy the crisp mountain air.*

> Let $\mathbf{r}_B$ = *if I choose B, I will get to enjoy the way the sand feels beneath my toes & if I choose A, I will not get to enjoy the way the sand feels beneath my toes.*

Suppose you choose to receive B over A. If what ultimately motivates you to do so is $\mathbf{r}_B$, have you thereby made a rational mistake?[11] Not necessarily. You could, recognizing that your reasons don't weigh more heavily in one direction than the other, choose to treat $\mathbf{r}_B$ as decisive by focusing in on its distinctive characteristics.[12] There needn't be anything irrational about deciding to care more strongly about enjoying the sand beneath your toes than enjoying the crisp mountain air; and so it needn't be a rational mistake for you to act on the basis of $\mathbf{r}_B$ while in full recognition of $\mathbf{r}_A$.

Would it be rational to be moved by the thought *"if I take the Regular box, then I will get the R-prize"* while fully aware that if you take the Larger box, you will get the L-prize? If the two reasons are incommensurable, then it would be; if, on the other hand, the two reasons are perfectly balanced, it would not be rational to be moved in this way.

According to proponents of the Most Reasons Argument, like Hare [2010, 2015], it wouldn't be rational of you to be moved by this consideration. Why? You don't know whether 'the R-prize' picks out B or A, so you aren't in a position to focus in on the various distinctive features of the R-prize that make it choiceworthy. From your current vantage point, what are the features of the R-prize that could recommend it over the L-prize? You don't know enough about the R-prize for there to be any such features.[13]

[11] I'm presupposing, for the sake of the example, that all of the considerations in favor of choosing A don't outweigh the considerations in favor of choosing B, and that all of the considerations in favor of choosing B don't outweigh the considerations in favor of choosing A.

[12] Chang [2009] would describe this in terms of exercising one's rational agency — when your given reasons "run out," you can *will* a consideration to be a reason, or exercise your will to transform one of your given reasons into a new *volitional* reason. Doing so involves putting your agency behind one of the considerations that counts in favor of one of your options. You, roughly, take that consideration to be particularly important to you, and, in so doing, it becomes something which helps make you the distinctive person that you are.

[13] This raises an interesting question about how acquainted one must be with the features that count in favor of one's options in order for those features to provide reasons that it

It would be a rational mistake to decide on taking the Regular box because it contains the R-prize when you are fully aware that the Larger box contains the L-prize. From your vantage point, you have no rational grounds on which to distinguish the two. Put differently: you have *no reason* to prefer the R-prize to the L-prize. Therefore, the consideration *"if I take the Regular box, I will get the R-prize & if I take the Larger box, I will not get the R-prize"* is perfectly balanced by the consideration *"if I take the Larger box, I will get the L-prize & if I take the Regular box, I will get the L-prize."*

Furthermore, because those two reasons are perfectly balanced and you have an additional reason to prefer the Larger box to the Regular box, you have *more reason* to prefer the Larger box to the Regular box than you do to prefer the Regular box to the Larger box. So, if the *More Reason Principle* is correct, you are rationally required to prefer the Larger box to the Regular box. Is the *More Reason Principle* correct?

3.2.3 Against the *More Reason Principle*

The *More Reason Principle* says:

> If you have more reason to prefer ϕ to ψ than you do to prefer ψ to ϕ, then rationality requires you to prefer ϕ to ψ.

wouldn't be a rational mistake to act on. In other words, (following [Chang, 2009]) how acquainted with a consideration must one be in order to will it to be a reason? Suppose that you've never been to the beach and that you've never been skiing. You've never felt the sand beneath your toes, or the crisp mountain air in your lungs. Would it, then, be a rational mistake to take B over A on the grounds that there'll be sand at the beach but not on the slopes? Or, perhaps you've experienced both, a long time ago, and can now no longer remember what they're like.

Or, imagine that you've been abducted by aliens. They take you to their home planet, where you'll live out the rest of your days. They offer you a choice about how to spend the rest of your life. You can choose to become a Gazingaborp or, alternatively, to become a Mikaelsour. They explain that if you become a Gazingaborp, you'll get plenty of baggilums. And, that if you instead opt to become a Mikaelsour, although you'll have to put up with a fair share of yeakizros, you'll be rewarded with as many farfanudles as you'd like. Would it be a rational mistake to choose to be a Gazingaborp over a Mikaelsour on the grounds that you'll get plenty of baggilums given that you fully recognize that by choosing, instead, to become a Mikaelsour you'll be rewarded with farfanudles?

How much do I have to know about a consideration — how acquainted with it must I be — in order to put my agency behind it? This is an interesting question to which I have no answer.

Rationality, according to this principle, is about correctly responding to the reasons that you have.

I disagree. Although what you have most reason to do and what rationality requires of you will often coincide, it's possible for the two to come apart. In particular: it's not true that rationality requires you to prefer ϕ to ψ whenever you have more reason to prefer ϕ to ψ than you do to prefer ψ to ϕ. If you know that the reasons you have are not all the reasons that there are, it might be okay to exercise some *rational humility:* to defer to what you know about the reasons *there are*, irrespective of what you *have* the most reason to do.

Here's the rough idea. You know that your better-informed self has various reasons to prefer taking the Regular box to taking the Larger box. In fact, you know that your better-informed self has *sufficient* reason take the Regular box over the Larger box — that is, the reasons had by your better-informed self are, on balance, weighty enough to render choosing the Regular box rationally permissible. You recognize this. It's not that, in virtue of recognizing this, you thereby come to *have* an additional reason to prefer the Regular box to the Larger box — it's still the case that *you have* more reason to prefer latter to the former.[14] But, you know that your better-informed self, in virtue of being better-informed, has *better* reasons than you do. And it's okay to (as far as possible) adopt the attitudes, which are informed by these better reasons, of your better-informed self.

Rational Humility: It's not always true that if you have more reason to prefer ϕ to ψ than you do to prefer ψ to ϕ, then you are rationally required to prefer ϕ to ψ.

Being rational isn't about doing what you have the most reason to do; rather, it's about aiming to do what there is most reason for you to do.

[14] Recognizing that your better-informed self has sufficient reason to ϕ amounts to knowing that there is sufficient reason to ϕ. And, as we saw in §2.2, it doesn't follow from the fact that you know *there is* reason to ϕ that you, thereby, *have* a reason to ϕ. That being said, some philosophers (see Schroeder [2009], for example) think: *the fact that there is a reason for you to ϕ* can, itself, be a reason for you to ϕ, albeit a derivative one.

Here's an example to help motivate **Rational Humility** against the *prima facie* implausibility of rejecting the *More Reason Principle:*

> **(Not) Eating Humble Pie.**[15] You come home to find a tasty looking pie cooling on your kitchen counter. You're hungry, the pie looks delicious, and you'd very much like to eat it. Before digging in, however, you notice that appended to the pie is a note from your mother — whom you correctly trust to have your best interests at heart — which reads: "DO NOT EAT." You know that your mother would only leave such a note if she knew that there was a decisive reason for you to not eat the pie. You, therefore, come to know that there is a decisive reason for you to avoid eating the pie. You do not, however, know what this decisive reason is. It might be that the pie was meant for your sister, or it might be that the pie is full of deadly poison, or it might be that the pie is not a pie at all but rather your father who fell victim to some kind of arcane sorcery, or it might be any number of other possibilities, you know not what. Whatever the case may be, you know that *there is* a decisive reason against eating the pie, but that this isn't a reason you *have*, and that the reasons you *do* have tell in favor of eating it.

In this story, it wouldn't be irrational of you to refrain from eating the pie. (In fact, it might be irrational of you *not* to do so). You know that there is a decisive reason for you to not eat the pie, and so, on that basis, you choose to not eat it. But, arguably, this is a case in which you have most reason to eat the pie (at least given the account of what it is to have a reason, sketched above). And, if all that is right, then this is a case in which rationality doesn't require you to do what you have the most reason to do; instead, you should defer to your mother, who, in virtue of being better-informed, is a guide to what *there is* most reason for you to do.

If you find that example unconvincing, allow me to make one final, more modest point, concerning the dialectic. If you're sympathetic to the **Actual Value Conception** — and, in particular, to the idea that if you know how

[15] This example is nearly structurally identical to the kinds of cases that Schroeder [2009] appeals to in order to motivate the claim, mentioned in the previous footnote, that existential facts concerning what reasons there are can, themselves, be reasons.

to align your preferences over your options with the facts concerning those options' actual values (in the manner described by the Regulative Ideal), you are rationally required to do so — then you already have good reason to be skeptical of the *More Reason Principle.* So the MORE REASON ARGUMENT, rather than providing a completely independent reason to reject **Actual Value Decision Theory**, appeals to a principle that is antecedently at odds with the picture of rationality that underlies the proposal it's targeting. That's not to say that the MORE REASON ARGUMENT is, thereby, unsound; rather, the thought is that, by using the *More Reason Principle* as a premise, the argument fails to further the dialectic.

There are two ways to bring out why someone who's sympathetic to the **Actual Value Conception** should, antecedently, regard the *Most Reason Principle* with suspicion. The first way, hinted at above, involves drawing an analogy between **Rational Humility** and the Regulative Ideal of Instrumental Rationality. The second way involves drawing an analogy between the *Most Reason Principle* and the PROSPECTOR PRINCIPLE.

Rational Humility & the Regulative Ideal. A central motivation for the **Actual Value Conception** is that, ideally, you should align your preferences over your options to the facts concerning those options' actual values. So, if you're in a position to know that ϕ has more actual value than ψ, you should prefer ϕ to ψ. There is a connection between the facts concerning your options' actual values and the facts concerning what there is most reason for you to prefer: roughly, if ϕ has more actual value than ψ, then there is more reason for you to prefer ϕ to ψ than to prefer ψ to ϕ. Because of this connection, we can understand **Rational Humility** as giving rise to a Regulative Ideal concerning, not the actual values of your options, but the reasons that there are for you to prefer one option to another. Ideally, were you omniscient, you would prefer ϕ to ψ when, and only when, that is what there is most reason for you to do. Of course, you aren't always (or, even usually) in a position to know what there is most reason to do. But, if you were to know that there isn't more reason to prefer ϕ to ψ, then you aren't rationally required to prefer ϕ to ψ.

> **The Regulative Ideal of Preferring for a Reason:** "Aim to prefer ϕ to ψ when, and only when, there is more reason to prefer ϕ to ψ than there is to prefer ψ to ϕ."
>
> $$\phi \succ \psi \quad \text{when, and only when} \quad \textsc{Reason}(\phi \succ \psi) > \textsc{Reason}(\psi \succ \phi)$$

This Regulative Ideal is analogous to, if not a consequence of, the Regulative Ideal of Instrumental Rationality, which underlies the **Actual Value Conception.** But accepting the ideal involves denying that you should always do what you have the most reason to do. And so, if you're sympathetic to the **Actual Value Conception**, you should reject the *More Reason Principle.* Sometimes, you are in a position to know that there are strong enough reasons to justify doing one thing, strong enough reasons to justify doing another, but fail to know enough about what, specifically, those reasons are so as to count as *"having"* them. In such cases, even if the reasons you *have* point in one direction, it's okay to go in another. Rationality is not about doing what you have the most reason to do; rather, being rational is about *aiming* to do what *there is* most reason to do.

Reasons & Prospects. Here's another (somewhat more speculative) reason that those of us who are sympathetic to the **Actual Value Conception** should regard the *More Reason Principle* with suspicion: given our characterization of what it is to have a reason, there appears to be a very tight connection between it and the PROSPECTOR PRINCIPLE. In fact, I hereby offer the following conjecture:

> **Reason/Prospect Conjecture**
>
> You have more reason to prefer ϕ to ψ than you do to prefer ψ to ϕ if and only if you regard the prospects associated with ϕing to be better than the prospects associated with ψing.

I suspect that this is true. Although I cannot prove it, I can offer some reasons for thinking it might be true. But, first, notice that *if* it's true, then the *More Reason Principle* is equivalent to the PROSPECTOR PRINCIPLE.[16]

[16] Recall what these two principles say:

[PROSPECTOR PRINCIPLE]

But the truth of the Prospector Principle is, as we saw in the previous chapter, the very thing at issue. So denying the *More Reason Principle* comes at no extra cost — the principle is merely *Prospectism* in sheep's clothing.

Why think the conjecture is true? What's the relationship between the reasons you have to prefer one option to another and those options' corresponding prospects? Consider the Larger box and the the Regular box, as an example. The More Reason Argument worked by, first, establishing that reasons you have for preferring the Larger box *without the added dollar* to the Regular box are perfectly balanced by your reasons for preferring the Regular box to the Larger box *without the added dollar* — everything that could be said in favor of preferring the one could, either, also be said in favor of preferring the other or was perfectly balanced by something that could be said in favor of preferring the other. Similarly, both of these options — taking the Larger box without the added dollar, or taking the Regular box — correspond to the same prospects: $\{\langle\frac{1}{2},A\rangle,\langle\frac{1}{2},B\rangle\}$. This isn't a coincidence. The prospects associated with an option capture all of the features, of which you are in a position to know, concerning how good taking that option might be. The reasons you have, also, correspond to the features, of which you are in a position to know, that concern how good taking that option might be. So, we should expect that if two options each correspond to the same prospects, you will have *equal* reason to prefer one to the other. Adding the dollar back to the Larger box provides you with an additional reason to prefer the Larger box, and (again, not coincidentally) improves the prospects that are associated with taking the Larger box. The fact that the prospects associated with taking the Larger box are *better* than the prospects associated with taking the Regular box, and the fact that you have more reason to prefer taking the Larger box over the Regular box than you do to prefer taking the Regular box to the Larger box, essentially, come to the same

> You should prefer ϕing to ψing if and only if you regard the prospects associated with ϕing to be better than the prospects associated with ψing.
>
> [*More Reason Principle*]
>
> Rationality requires you to prefer ϕ to ψ if and only if you have more reason to prefer ϕ to ψ than you do to prefer ψ to ϕ.

(Note that we made use of only the right-to-left direction of this principle in the More Reason Argument). If the conjecture is true, then these two principles are clearly equivalent.

thing. This far from *proves* the **Reason/Prospect Conjecture**, of course, but it does, I think, suggest that it, or something in the vicinity, is true.

Summing up: if you're sympathetic to the **Actual Value Conception**, you should reject the *More Reason Principle*: being rational is not about doing what you *have* the most reason to do; rather, it involves aiming to do what there is most reason to do. And, because of the connection between the principle and *Prospectism*, rejecting it ultimately comes at no extra cost.

But if the *More Reason Principle* is false, why does it seem true? Allow me to suggest two possibilities. First, even if the principle is false, it appears to hold in most cases. In general, doing what you *have* the most reason to do will coincide with aiming to do what *there is* most reason to do. And so, it's easy to run the two together. The second possibility is a bit more speculative: *If* you care about being able to easily justify your choices (under perhaps possibly antagonistic conditions), it might also make sense to value standing in close relation to reasons of a certain kind. Situations in which you anticipate having to justify your behavior to others, in particular, might inspire a strong interest in being able to articulate convincing rationales for your actions. And it might be that citing a specific non-"disjunctive" reason is typically more convincing than pointing to the fact that you know there is some reason or other.[17] Insofar, then, as you care about being able to marshall a compelling defense of your choices to others — and, for what it is worth, I think we *do* probably care about such things (and with good reason given our social nature) — you might be rationally compelled to choose the Larger box over the Regular box. But, supposing (as we've tacitly been doing) that you *only* care about the prizes in the boxes, and not the ease with which you could justify your choice to others, you aren't rationally required to choose the Larger box.

So much for the MORE REASON ARGUMENT. Let's now turn to the next argument against **Actual Value Decision Theory.**

[17] For example, Simonson and Nowlis [2000] and Brad M. Barber and Odean [2003] found that much of the "Reason-based choice" behavior (described by Eldar Shafir and Tversky [1993]) is amplified when decisions are made in groups, or when individuals are made to expect that they will have to justify their decisions to others. See, also, [Lerner and Tetlock, 1999] for a discussion of how "accountability" affects decision-making.

3.3 The Agglomeration Argument

In this section, we will look at another argument against **Actual Value Decision Theory.** The argument appeals to the fact that **Actual Value Decision Theory** offers recommendations that violate an agglomeration principle governing rational requirement. However, as I'll go on to argue, *Supervaluational Expected Utility Theory* violates a related agglomeration principle governing, not rational requirement, but rational permission. The lesson we should draw is this: when your preferences are incomplete, violations of this sort should be expected. There mere fact that **Actual Value Decision Theory** violates an agglomeration principle is not sufficient reason to reject it.

The argument goes like this. If **Actual Value Decision Theory** is correct, then there will be sequences of actions such that (i) you are rationally required to take each of the actions in the sequence, but (ii) you are not rationally required to perform the sequence itself. In other words, **Actual Value Decision Theory** violates the following *agglomeration* principle.[18]

> [WEAK AGGLOMERATION (REQUIREMENT)]
>
> For any sequence of actions $\langle \phi_1, \phi_2, \ldots, \phi_n \rangle$, if you are rationally required to take ϕ_1 (irrespective of what else you might do), and you are rationally required to take ϕ_2 (irrespective of what else you might do), ..., then you are rationally required to perform the sequence $\langle \phi_1, \phi_2, \ldots, \phi_n \rangle$.

And, so the objection continues, in order for a decision theory to be adequate, it should not issue advice that violates WEAK AGGLOMERATION (REQUIREMENT).[19] If a decision theory violates this agglomeration princi-

[18] [Hare, 2015]

[19] What makes WEAK AGGLOMERATION (REQUIREMENT) *"weak"*? The parenthetical clauses, in the antecedent of the principle, that say "irrespective of what else you might do" are what distinguish the weak version from the strong. Proponents of *Actualism*, like [Frank Jackson, 1986] for example, can accept weak but not strong agglomeration. The example of Professor Procrastinate brings this out. Professor Procrastinate is deciding between *accept*ing an invitation to write a book review and *declining* the invitation. If Professor Procrastinate *accepts*, she will then have to choose, either, to *write* the review or to *put it off* indefinitely. It would be best, all things considered, for her to *accept* and then *write*. And that's what she ought to do. However, Professor Procrastinate is self-aware enough to know that if she *accepts*, she'll procrastinate: she won't write the

ple, we should reject it. Therefore, we should reject **Actual Value Decision Theory.**

The Weak Agglomeration (Requirement) Argument

P1 **Actual Value Decision Theory** violates Weak Agglomeration (Requirement).

P2 If a decision theory violates Weak Agglomeration (Requirement), it is inadequate and should be rejected.

C **Actual Value Decision Theory** is inadequate and should be rejected.

I think we should reject premise **P2.**[20] But first, let's go over how **Actual Value Decision Theory** violates Weak Agglomeration (Requirement).

3.3.1 Actual Value Decision Theory Violates Agglomeration

Consider the following two decision-problems:

	Heads	Tails
L_1	A^+	0
R_1	0	A

	Heads	Tails
L_2	0	B^+
R_2	B	0

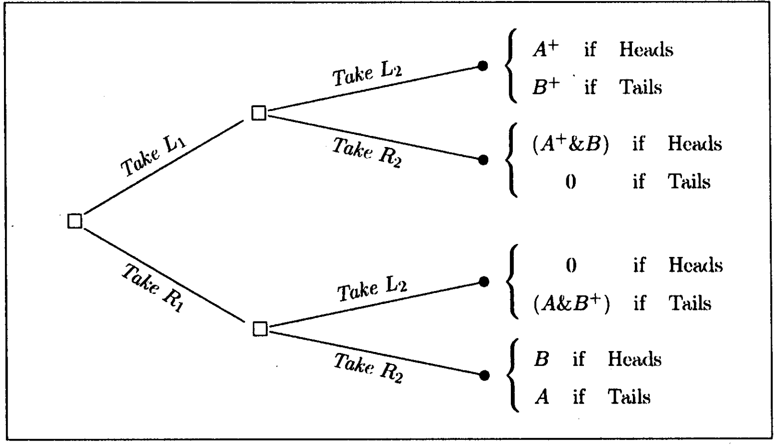
Imagine that in **Vacation Boxes** you are offered the following deal. You are to guess which of the two boxes, the Larger or the Regular, contains

review. But it's better to decline the invitation than to *accept* and *put it off*. So, according to *Actualism*, she ought to *decline*. So, she both ought to *decline* and ought to perform the sequence ⟨*accept*, *write*⟩. However, this case doesn't violate Weak Agglomeration (Requirement) (although, it does violate the strong version) because it's not the case that she ought to decline the invitation *irrespective of what else she might do:* if she won't write the review, she should decline the invitation; but she shouldn't decline the invitation if she can bring herself to write the review.

[20] More accurately, as will become clear, I think the argument fallaciously equivocates between two distinct ways of understanding the agglomeration principle. On one way of understanding the principle, the first premise is false. On the other way of understanding the principle, the second premise is false. If we understand the principle so as to make *both* premises true, then the argument is invalid.

prize A. If you guess correctly, you win it (plus whatever else is in the box). Because it's equally likely that A is in the Larger box as it is that it's in the Regular box, the fact that you'll win a dollar if you correctly guess that it is in the Larger box provides you with a decisive reason to guess that it's in the Larger box. In other words, you should prefer option L_1 to option R_1. Similar reasoning says that you should also prefer L_2 to R_2. Furthermore, suppose you face the two decision-problems in sequence: first, you are asked to guess which box contains prize A, then you are asked to guess which box contains prize B. Irrespective of what you'll guess in round 2, you are rationally required to guess that A is in the Larger box; and, irrespective of what you guessed in round 1, you are rationally required in round 2 to guess that B is in the Larger box.[21]

Agglomeration Failure?



The sequence $\langle L_1, L_2\rangle$ is just like taking the Larger box: you get A^+ if heads, and B^+ if tails. And the sequence $\langle R_1, R_2\rangle$ is just like taking the Regular box: you get B if heads, and A if tails. According to **Actual Value Decision Theory**, though, you are *not* rationally required to prefer taking the Larger box to taking the Regular box, even though you *are* rationally required to prefer L_1 to R_1 and to prefer L_2 to R_2.

[21] Nothing is meant to turn on the fact that you are making "guesses" (and that, by performing L_1 and L_2, it's guaranteed that one of your "guesses" is incorrect). This is but one way of dramatizing the decision-problem that interests us. It would be fine, instead, to imagine being offered two choices between two different gambles, all of which turn on the result of the same coin toss.

Why are you rationally required to prefer L_1 to R_1 and to prefer L_2 to R_2?[22] According to **Actual Value Decision Theory**, you are rationally required to prefer L_1 to R_1 just in case your estimate of the extent to which L_1's actual value exceeds R_1's is greater than your estimate of the extent to which R_1's actual value exceeds L_1's. Your estimates of these comparisons depends on what you believe you will do when faced with the choice between L_2 and R_2. There are eight relevant dependency hypotheses, each corresponding to one of the four ways your future choice might depend on your current choice and one of the two ways the coin might have landed.[23]

Ways Your Future Choice Might Depend on Your Current Choice:

$$\begin{array}{cccc}
\overbrace{L_1 \mathrel{\Box\!\!\rightarrow} L_2}^{K_1} & \overbrace{L_1 \mathrel{\Box\!\!\rightarrow} L_2}^{K_2} & \overbrace{L_1 \mathrel{\Box\!\!\rightarrow} R_2}^{K_3} & \overbrace{L_1 \mathrel{\Box\!\!\rightarrow} R_2}^{K_4} \\
R_1 \mathrel{\Box\!\!\rightarrow} L_2 & R_1 \mathrel{\Box\!\!\rightarrow} R_2 & R_1 \mathrel{\Box\!\!\rightarrow} L_2 & R_1 \mathrel{\Box\!\!\rightarrow} R_2
\end{array}$$

[22] Here's an answer that, while too quick to be correct, is instructive. Consider the choice between L_1 and R_1. If the coin landed heads, then the actual value of performing L_1 exceeds the actual value of performing R_1 to an extent that is proportional to the "value" you assign to A^+. On the other hand, if the coin landed tails, then the actual value of performing R_1 exceeds the actual value of performing L_1, but only to an extent proportional to the "value" you assign to A. Therefore, because you prefer A^+ to A, the extent to which the actual value of L_1 might exceed the actual value of R_1 is greater than the extent to which the actual value of R_1 might exceed the actual value of L_1. The coin is fair, so, from your perspective, it's equally likely that the L_1's actual value exceeds R_1's as it is that R_1's exceeds L_1. So, your estimate of the comparison of actual values between your two options should rank L_1 ahead of R_1. Consequently, you are rationally required to prefer L_1 to R_1, and are therefore rationally required to take L_1. Analogous reasoning will lead us to conclude that you are rationally required to prefer L_2 to R_2, and are therefore rationally required to take L_2.

Although this line-of-thought is mostly correct, it's too quick. It misdescribes the outcomes of your decision-problems. You know, when deciding between L_1 and R_1, that you'll soon face a choice between L_2 and R_2. And what rationality requires of you *now* depends on what you believe you will do *later*. Similarly, when you are deciding between L_2 and R_2, you have already taken either L_1 or R_1, which affects the outcome your (once future, but now current) choice might bring about. A more careful statement of the argument is presented in the main text.

[23] This is a bit of an oversimplification. At the risk of being overly pedantic, I should note that there are actually many more (relevant) dependency hypotheses — in particular, ones according to which your current and/or future choices causally influence how the coin landed. However, because you know the coin toss is independent of your choices (it already landed!), you ought to place no credence in these dependency hypotheses. For ease of presentation, then, I've ignored them.

Your Initial Decision

	K_1		K_2		K_3		K_4	
	Heads	Tails	Heads	Tails	Heads	Tails	Heads	Tails
L_1	A^+	B^+	A^+	B^+	$A^+ \wedge B$	0	$A^+ \wedge B$	0
R_1	0	$A \wedge B^+$	B	A	0	$A \wedge B^+$	B	A

Whether rationality requires you to take L_1 over R_1 depends on your credences in these eight dependency hypotheses. For example, if you place enough credence in K_2, you should regard your options as on a par and, thus, you are not rationally required to take L_1 over R_1; or, if you are certain that K_3 is actual, then you should be indifferent between the two and, again, you are therefore not required to take L_1 over R_1. What should your credences in these hypotheses look like? Here's an argument for placing all your credence in K_1 — the hypothesis according to which you will choose L_2 over R_2 irrespective of what you decide to do now. First, suppose that you took option L_1. If so, then your choice between L_2 and R_2 is, effectively, the choice between (g_i) a gamble that pays out A^+ if heads and B^+ if tails and (g_{ii}) a gamble that pays out *both* A^+ and B if heads and nothing if tails. Assuming (for the sake of the objection) that you regard the values of A and B (and the dollar) to all be independent of each other, you should prefer the former gamble to the latter. Here's why. If the coin landed heads, then the actual value of g_{ii} exceeds the actual value of g_i to an extent proportional to the value of receiving prize B; if the coin landed tails, then the actual value of g_i exceeds the actual value of g_{ii} to an extent proportional to the value of receiving B *plus* a dollar. Because you prefer B^+ to B, the extent to which the actual value of g_i might exceed the actual value of g_{ii} is greater than the extent to which the actual value of g_{ii} might exceed the actual value of g_i. Because you know the coin is fair, you should think it equally likely that g_{ii}'s actual value exceeds g_i's as that g_i's exceeds g_{ii}'s. Therefore, your estimate of the comparison in actual value between the two should come down in favor of g_i over g_{ii}; and so, supposing that you took L_1 over R_1, according to **Actual Value Decision Theory**, you are rationally required to take L_2 over R_2. Now suppose, instead, that you took R_1 rather than L_1. If so, then your choice between L_2 and R_2 is, effectively, the choice between (g_{iii}) a gamble that pays out nothing if heads and *both* A and B^+ if tails and (g_{iv}) a gamble that pays out B if heads and A if tails. An argument analogous to the one just given delivers the result that you should prefer g_{iii} to g_{iv}. And so, supposing you took R_1 over L_1, you are rationally required to take L_2 over R_2. It follows, then, that irrespective of the choice you make between L_1 and R_2, if you're rational, you will choose L_2 over R_2.

Supposing you take L_1 ...

	Heads	Tails
L_2	A^+	B^+
R_2	$A^+ \wedge B$	0

Supposing you take R_1 ...

	Heads	Tails
L_2	0	$A \wedge B^+$
R_2	B	A

By reasoning "backwards" from what you would choose to do (assuming you're rational) at the second round of the decision-problem, we can conclude that, no matter which of the two options you choose during the first round, you will go on to choose L_2 over R_2. In other words, you are in a position to be rationally certain that were you to choose L_1 you would go on to choose L_2 and were you to choose R_1 you would go on to choose L_2. That's what hypothesis K_1 says. So, you should be maximally confident that K_1 is actual. So, the first round decision-problem between options L_1 and R_1 can be simplified, by "deleting" those states to which you assign no credence:

K_1

	Heads	Tails
L_1	A^+	B^+
R_1	0	$A \wedge B^+$

The extent to which L_1 might (if the coin landed heads) do better than R_1 exceeds the extent to which R_1 might (if the coin landed tails) do better than L_1. (Why? Because, given that you prefer A^+ to A, your preference for A^+ over 0 should be stronger than your preference for $(A \wedge B^+)$ over B^+). Because you know the coin is fair, you should regard both possibilities to be equally likely. Therefore, you are rationally required to choose L_1 over R_1.

Putting all of the pieces together, according to **Actual Value Decision Theory**: (i) you are rationally required to take L_1 over R_1, and you are rationally required to take L_2 over R_2, but (ii) you are *not* rationally required to perform the sequence $\langle L_1, L_2\rangle$ over the sequence $\langle R_1, R_2\rangle$. And those recommendations violate WEAK AGGLOMERATION (REQUIREMENT).

It's worth being explicit about some of the assumptions that were made in the service of defending premise **P1**. First, we assumed that you valued A, B, and the dollar in a particular way: namely, that you value all three *independently* in that having any one of them doesn't add to or take away from the value of having any of the others. Without this assumption, we can no longer say, for example, that your preference for A^+ over 0 is stronger than your preference for $(A \wedge B^+)$ over B^+. This is a strong and, at least in regards to things like vacations, implausible assumption. However, the argument only requires there to be *at least one* situation in which WEAK AGGLOMERATION (REQUIREMENT) is violated. So, while it's implausible that you value vacations independently, it's not implausible that there are some goods such that you, both, value those goods independently and regard them as being on a par. Second, the argument presupposed a particular way of thinking about sequential decision-making: what you are rationally required to do at one time is (partially) determined by what you believe you will do later on. If you can predict what you will choose to do later on, then certain plans, or sequences of choices, are treated as *infeasible* from your current perspective. This is often called *Sophisticated Choice.* But there are other ways of thinking about sequential decision-making that conflict with *Sophisticated Choice.*[24] For now, I will merely flag this as an assumption.

Before turning our attention to premise **P2** — which, recall, says that a decision theory must satisfy WEAK AGGLOMERATION (REQUIREMENT) in order to be adequate — I want to first show that *Supervaluational Expected Utility Theory*, also, violates an agglomeration principle. This serves an important dialectic function: if the defense given for premise **P2** can just as easily apply to this agglomeration principle (which I believe it can), then

[24] Traditionally, there are three methods for choosing over time: myopically, sophisticatedly, and resolutely. Myopic Choice holds that, at each time, you should perform the available action that is part of the overall sequence (or plan, or strategy) of actions that you most prefer *at that time.* Sophisticated Choice says that you should, first, determine which sequences of actions are *feasible* given the preferences your future-self will have; then, choose the action that you regard as best, given the predictions made about what you will do in the future. Resolute Choice says that you should identify the sequence of actions you most prefer, and then *bind* yourself to performing that sequence of actions.

neither decision theory is adequate. The fact that my proposal violates an agglomeration principle isn't reason to reject it in favor of *Supervaluational Expected Utility Theory* if it, too, violates an agglomeration principle.

3.3.2 Diachronic Problems for *Supervalutional Expected Utility Theory*

As we've seen, **Actual Value Decision Theory** violates a weak agglomeration principle for rational requirements. To my knowledge, *Supervaluational Expected Utility Theory* doesn't violate WEAK AGGLOMERATION (REQUIREMENT).[25] However, there is a different agglomeration principle that it does violate.

> [WEAK AGGLOMERATION (PERMISSION)]
>
> For any sequence of actions $\langle \phi_1, \phi_2, \ldots, \phi_n \rangle$, if you are rationally permitted to take ϕ_1 (irrespective of what else you might do), and you are rationally permitted to take ϕ_2 (irrespective of what else you might do), ..., then it's rationally permissible to perform the sequence $\langle \phi_1, \phi_2, \ldots, \phi_n \rangle$.

Here's a situation in which *Supervaluational Expected Utility Theory* violates WEAK AGGLOMERATION (PERMISSION). Consider the following two decision-problems:

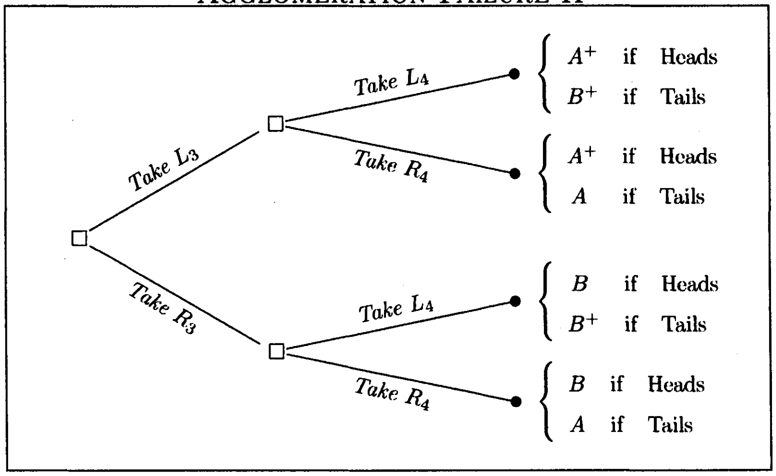
[25] Here's a reason to think that there can be no situations in which *Supervaluational Expected Utility Theory* violates the principle. For simplicity, assume that you are facing two decision-problems in sequence: first, a choice between ϕ_1 and ψ_1 and, then, a choice between ϕ_2 and ψ_2. Furthermore, suppose that you are required to choose ϕ_1 over ψ_1 and ϕ_2 over ψ_2. If so, that must mean that every utility-function u in your set is such that $\sum_K Cr(K) \cdot u(\phi_1 \wedge K) > \sum_K Cr(K) \cdot u(\psi_1 \wedge K)$ and $\sum_K Cr(K) \cdot u(\phi_2 \wedge K) > \sum_K Cr(K) \cdot u(\phi_2 \wedge K)$. Also assume that, for each state K_i, you value the goods in outcome $(\phi_1 \wedge K_i)$ independently of the goods in outcome $(\phi_2 \wedge K_i)$, and likewise for ψ_1 and ψ_2, so that, for every utility-function u in the set, $u(\phi_1 \wedge \phi_2 \wedge K_i) = u(\phi_1 \wedge K_i) + u(\phi_2 \wedge K_i)$ (and likewise for ψ_1 and ψ_2). Because of the inequalities, above, it must also be true that $\sum_K Cr(K) \cdot u(\phi_1 \wedge K) + \sum_K Cr(K) \cdot u(\phi_2 \wedge K) > \sum_K Cr(K) \cdot u(\psi_1 \wedge K) + \sum_K Cr(K) \cdot u(\phi_2 \wedge K)$, which holds just in case $\sum_K Cr(K) \cdot (u(\phi_1 \wedge K) + u(\phi_2 \wedge K)) > \sum_K Cr(K) \cdot (u(\psi_1 \wedge K) + u(\psi_2 \wedge K))$. And, given our assumption of independence, $\sum_K Cr(K) \cdot u(\phi_1 \wedge \phi_2 \wedge K) > \sum_K Cr(K) \cdot u(\psi_1 \wedge \psi_2 \wedge K)$. And so every utility-function in the set will rank the sequence $\langle \phi_1, \phi_2 \rangle$ ahead of $\langle \psi_1, \psi_2 \rangle$.

	HEADS	TAILS
L_3	A^+	0
R_3	B	0

	HEADS	TAILS
L_4	0	B^+
R_4	0	A

You are first offered a choice between option L_3 and option R_4. After making that decision, you will be offered a choice between option L_4 and option R_4. According to *Supervaluational Expected Utility Theory*, it's permissible for you to choose R_3 over L_3, and it's permissible for you to choose R_4 over L_4. According to WEAK AGGLOMERATION (PERMISSION), it should also be permissible for you to perform the sequence $\langle R_3, R_4\rangle$ over the sequence $\langle L_3, L_4\rangle$. But the prospects associated with performing the sequence $\langle R_3, R_4\rangle$ as the same as those associated with taking the Regular box, and the prospects associated with the sequence $\langle L_3, L_4\rangle$ are the same as those associated with taking the Larger box. So, according to *Supervaluational Expected Utility Theory*, you are rationally *required* to perform the sequence $\langle L_3, L_4\rangle$ over the sequence $\langle R_3, R_4\rangle$.

AGGLOMERATION FAILURE II



But how can it be rationally permissible to choose R_3 (irrespective of what else you might do) and rationally permissible to choose R_4 (irrespective of what you earlier did), and yet not permissible to perform the sequence $\langle R_3, R_4\rangle$?[26]

[26] There is a further diachronic problem for *Supervaluational Expected Utility Theory*. A problem it doesn't share with **Actual Value Decision Theory**, and a problem that cannot be easily circumvented by distinguishing between *performing* a sequence of actions and *choosing to perform* a sequence of actions. The problem is this: *Supervaluational*

3.3.3 Assessing Premise P2: Why Not Violate Agglomeration?

Let's grant that **Actual Value Decision Theory** violates WEAK AGGLOMERATION (REQUIREMENT). Why should this concern us? Why think that a decision theory that violates the principle is inadequate?

Packages of Unfollowable Advice?

Hare [2015] defends a version of WEAK AGGLOMERATION (REQUIREMENT) (concerning the moral "ought") on the following grounds:

> I think [WEAK AGGLOMERATION (REQUIREMENT) is] compelling because it is the job of a moral theory to give you a package of advice, and a package of advice that is *followable* in the following sense: there is something that you can do such that, if you were

Expected Utility Theory will sometimes recommend avoiding relevant cost-free information. In fact, sometimes, it will recommend *paying* to avoid getting relevant information. The case presented above can be transformed into such a situation. Imagine that, at time t_1, you are given the opportunity to learn whether the coin landed heads or tails (and, thus, to learn which box contains what). Or, you can pay a very small amount of money to remain ignorant. After either learning or paying to avoid learning, you are to choose between taking the Larger box and taking the Regular box. If you decided to pay to remain ignorant, then, according to *Supervaluational Expected Utility Theory*, you are rationally required to take the Larger box. If you decide to learn which box contains what, you don't know which box you will ultimately decide to take — if you learn that the coin landed heads, it's permissible to take either; and, likewise, if you learn that the coin landed tails. The prospects associated with learning which box contains what are worse than the prospects associated with paying to avoid this information. So, at time t_1, you should pay to remain ignorant.

Given the close connection between this case and the Reflection Argument (discussed in the previous chapter), this result is perhaps not too surprising. Interestingly, it's alleged that, in some cases, evidential decision theory also recommends avoiding (and even paying to avoid) cost-free information [Skyrms, 1990]. Consider a variation of the Newcomb Problem in which you are, first, given the opportunity to learn whether there is money in the opaque box. Evidentialists, it's alleged, will pay to remain ignorant. They will reason as follows: "If I learn that there's money in the opaque box, I will take both boxes; if I learn that there's no money in the opaque box, I'll take both; so, no matter what I learn, I'll take both boxes. But the predictor is very reliable. So, if I choose to learn, it's very likely that the predictor predicted I'd take both boxes, and put nothing in the opaque box. On the other hand, if I choose to remain ignorant and then take only the one box, it's very likely that the predictor predicted I'd take only the one, and put a million dollars in it. So the expected value of remaining ignorant vastly exceeds the expected value of learning." I take this to be further confirmation that there is a tight connection between One-Boxing vs Two-Boxing, on the one hand, and decision-making in the face of parity, on the other.

> to do it, then you would have done everything that the theory says you ought to do. Any theory that violates [WEAK AGGLOMERATION (REQUIREMENT)] does not give us a package of advice that is followable in this sense. (pg. 13)

Although this argument appeals to the job of a *moral* theory, it's clear that an analogous point can be made about the job of a theory of *rationality:* it, too, should only issue requirements that can be followed.[27] Why do theories that violate WEAK AGGLOMERATION (REQUIREMENT) fail at issuing packages of advice that are followable? Suppose that a theory says the following: you are rationally required to choose option ϕ_1 (irrespective of whatever else you might do), and you are rationally required to choose option ϕ_2 (irrespective of whatever else you did or might do), and ..., and you are rationally required to choose option ϕ_n (irrespective of whatever else you did), but you are also rationally required to perform some sequence of actions other than $\langle \phi_1, \phi_2, \ldots, \phi_n \rangle$. This package of advice is not followable: there is nothing you can do such that, were you to do it, you would have done everything that the theory advises you to do. The only way for you to satisfy the requirement of performing a sequence *other* than $\langle \phi_1, \phi_2, \ldots, \phi_n \rangle$ is for you to not choose one of the options constituting the sequence. But, according to the theory, each of the options constituting the sequence is one you are rationally required to choose. So there is no way for you to follow all of the advice the theory offers.

But note two things. First, **Actual Value Decision Theory** violates WEAK AGGLOMERATION (REQUIREMENT) because it says (i) that you are rationally required to choose L_1 over R_1 and that you are rationally required to choose L_2 over R_2, and (ii) that you are *not* rationally required to choose to perform the sequence $\langle L_1, L_2 \rangle$ over the sequence $\langle R_1, R_2 \rangle$. But it *does not* say that there is some sequence other than $\langle L_1, L_2 \rangle$ such that you are rationally required to perform it. In other words, according to **Actual Value**

[27] In fact, this argument for WEAK AGGLOMERATION (REQUIREMENT) is, perhaps, even more compelling for the "ought" of rationality than the moral "ought." If a moral theory violates the principle, then, in some situations, there will be nothing you can do such that, if you were to do it, you would have done everything that the moral theory says you ought to do. If there are genuine moral dilemmas — situations in which there are multiple incompatible things you ought to do — then there will be situations in which there is nothing you can do to make it such that you've done everything you morally ought to do. But, while there might be genuine moral dilemmas, it's less plausible that there are genuine *rational* dilemmas.

Decision Theory, it is *permissible*, but not rationally required, to choose to perform $\langle L_1, L_2\rangle$. Second, a package of advice which says that you're required to choose ϕ_1, that you're required to choose ϕ_2, ..., that you're required to choose ϕ_n, and that it's *permissible* to perform the sequence $\langle \phi_1, \phi_2, \ldots, \phi_n\rangle$ *is* followable in the relevant sense: there is something that you can do — namely, perform each of the actions in the sequence $\langle \phi_1, \phi_2, \ldots, \phi_n\rangle$ — such that, were you to do it, you would have done everything required of you. This suggests that Hare [2015]'s argument for WEAK AGGLOMERATION (REQUIREMENT) really supports a weaker agglomeration principle.

> [WEAKER AGGLOMERATION (REQUIREMENT)]
>
> For any sequence of actions $\langle \phi_1, \phi_2, \ldots, \phi_n\rangle$, if you are rationally required to take ϕ_1 (irrespective of what else you might do), and you are rationally required to take ϕ_2 (irrespective of what else you might do), ..., then it's rationally permissible to perform the sequence $\langle \phi_1, \phi_2, \ldots, \phi_n\rangle$.

Actual Value Decision Theory violates WEAK AGGLOMERATION (REQUIREMENT) but it does not violate WEAKER AGGLOMERATION (REQUIREMENT). And it is the latter, not the former, that needs to be maintained if we want to ensure that our theory of rationality only issues packages of followable advice.

Distinctions without a Difference

One might object that the package of advice **Actual Value Decision Theory** issues in this case — namely, that you're required to choose L_1, that you're required to choose L_2, but that you're not required to choose $\langle L_1, L_2\rangle$ — while followable in the relevant sense, is nevertheless, at the very least, *odd*. How can there be any room between, on the one hand, being required to choose L_1 (irrespective of what else you might do) and being required to choose L_2 (irrespective of what else you might do), and, on the other hand, being required to choose to do *both* L_1 and L_2? If you're required to choose L_2 (irrespective of what else you might do), then you are required to choose L_2 conditional on having already chosen L_1. But, by choosing L_2 after having already chosen L_1 (which is something you were required to do), you thereby make it the case that you've performed the sequence $\langle L_1, L_2 \rangle$. If you choose L_1 and then choose L_2, nothing more needs to be done in order to perform the sequence. Given that you're required to choose L_1, and that you're required to then choose L_2, and that, by choosing to do both of these things, you've thereby performed the sequence $\langle L_1, L_2 \rangle$, aren't you also thereby *required* to perform the sequence? If the rules of the house require you to put your dirty dishes in the sink, and if they also require you to wash any dishes that are in the sink, don't the house rules *thereby* require you to wash your dirty dishes?

In the service of making clear just how odd this is, consider the following case.

> **Non-Sticky Buttons.** Before you are two buttons, one marked $\mathbf{L_1}$, the other marked $\mathbf{L_2}$. You are to, first, decide whether or not to push the $\mathbf{L_1}$-Button, and then decide whether or not to push the $\mathbf{L_2}$-Button. Pushing the $\mathbf{L_1}$-Button amounts to taking option L_1. Not pushing amounts to taking option R_1. Similarly, pushing the $\mathbf{L_2}$-Button amounts to taking L_2, and not pushing amounts to taking R_2.

What should you do? If you don't push *both* buttons, you've made a rational error. You should push the $\mathbf{L_2}$-Button, irrespective of whether you pushed the the $\mathbf{L_1}$-Button; and you should push the the $\mathbf{L_1}$-Button, irrespective of whether you will push the the $\mathbf{L_1}$-Button. So you should, first, push the $\mathbf{L_1}$-Button and then push the the $\mathbf{L_2}$-Button. If you are rational, you will end up performing the sequence $\langle L_1, L_2 \rangle$.

> **Sticky Buttons.** Same as before, except for the following. You learn that a piece of tape has been placed on the two buttons so that if one is pushed, so is the other.

Now what should you do? In **Non-Sticky Buttons**, you had to push both buttons on pain of irrationality. In this case, however, it's rationally permissible to refrain from pushing the buttons. This is odd. Before you knew about the tape, you were rationally required to push both buttons. And, because you are a rational person, that's what you planned on doing. Shouldn't the added tape merely make your decision in **Sticky Buttons** more convenient? You were going to push both buttons anyway, adding the tape simply reduces your workload — now, instead of having to push *two* buttons, the tape allows you to accomplish the same thing by simply pushing one!

According to **Actual Value Decision Theory**, however, the added tape doesn't make your decision *easier* — it makes it, in some sense, *harder.* Without the added tape, it was clear what needed to be done: you should push both buttons. But *with* the tape, you now must deliberate about what to do: it's permissible to push both buttons, but it's also permissible to push neither. How could a measly piece of tape make such a profound difference?

The tape matters because, by sticking the two buttons together, it changes what choices you are able to make. The outcome of rational deliberation is *choice*, not mere behavior or actions or performances. The requirements of rationality apply, foremost, to the choices we make, not to the actions we perform or the things that we do. (Or, rather, insofar as rationality *does* apply to actions or performances or the like, it does so only *derivatively:* in general, when we choose to perform some action, our choice results in us performing that action. But it is the *choices* we make — not the actions that, in general, result from making the choice — that are the subjects of rational evaluation.) Because rationality concerns the choices we make, by changing the choices you are able to make, the added tape can make a difference to what you are rationally required, or permitted, to choose to do.

In **Sticky Buttons**, there are only two things you can choose to do: you can choose to push both buttons, or you can choose to push neither. The former is, in terms of payoffs, equivalent to performing the sequence $\langle L_1, L_2 \rangle$; while the latter is, in terms of payoffs, equivalent to performing the sequence $\langle R_1, R_2 \rangle$. According to **Actual Value Decision Theory**, neither of these choices is rationally required. In terms of the choices you can make, **Non-Sticky Buttons** differs from **Sticky Buttons**. Because the two buttons, unencumbered by tape, must be pushed independently (if pushed at all), we need to distinguish between the choices you have at time t_1 and the choices you have a time t_2. At time t_1, you can either choose to push the $\mathbf{L_1}$-Button or choose to not push it. Whether these are the *only* choices you can make at time t_1 depends on whether or not you have, at that time, the ability to *commit* yourself to push, or to not push, the button at time t_2. If you can bind yourself in this way, then, at time t_1, you also have the choice to perform the sequence $\langle L_1, L_2\rangle$, the choice to perform the sequence $\langle L_1, R_2\rangle$, the choice to perform the sequence $\langle R_1, L_2\rangle$, and the choice to perform the sequence $\langle R_1, R_2\rangle$. Suppose for the moment, though, that you cannot bind yourself at time t_1 — the only choices you can make are whether or not to push the $\mathbf{L_1}$-Button. According to **Actual Value Decision Theory**, you are rationally required to push it. At time t_2, there are two more choices you can make: you can choose to push the $\mathbf{L_2}$-Button or choose not to. Again, **Actual Value Decision Theory** says that you are rationally required to push. At the end of the process, you will have performed the sequence of actions corresponding to $\langle L_1, L_2\rangle$. But notice that at no time — neither time t_1 nor time t_2 — was performing that sequence an option for you. Although you performed the sequence, you didn't *choose* to perform it. The requirements of rationality apply only to what you can *choose* to do, so there is no conflict between what **Actual Value Decision Theory** says is required of you in **Non-Sticky Buttons** and what it says is required of you in **Sticky Buttons**. It's true that if you're rational, then, in **Non-Sticky Buttons**, you will perform the sequence $\langle L_1, L_2\rangle$, and it's also true that if you're rational, then, in **Sticky Buttons**, you needn't perform the sequence $\langle L_1, L_2\rangle$. And that's because performing the sequence is something you can choose to do in the latter case, but it is not something you can choose to do in the former. Furthermore, because the tape changes what your available choices are, it is *not* odd for the requirements of rationality to differ between the two cases.

The distinction between what you do and what you *choose* to do can help us assess the agglomeration principles.

(1) You are rationally required to perform the sequence.

  a. You are rationally required to be such that you've performed the sequence.

  b. You are rationally required to *choose* to perform the sequence.

(2) It's rationally permissible to perform the sequence.

  a. It's rationally permissible to be such that you've performed the sequence.

  b. It's rationally permissible to *choose* to perform the sequence.

If there are some actions you're required to perform, then (1a): you are, thereby, rationally required to be such that you've performed that sequence of actions. However, it needn't be true that if there are some actions you're required to perform, then (1b): you are rationally required to *choose* to perform that sequence. As we've seen, performing a sequence of actions is not something that you'll always be in a position to choose to do. Each action in the sequence is an available option that you, at one point in time or another, could choose to do; however, there needn't be any point in time at which the *sequence itself* is something you could choose to do. That is to say: while ϕ_1 is an option for you at time t_1, and option ϕ_2 is an option for you at time t_2, $\ldots$, and ϕ_n is an option for you at time t_n, there needn't be any particular time at which performing the sequence $\langle\phi_1, \phi_2, \ldots, \phi_n\rangle$ is an option for you. If something isn't an option for you, then rationality cannot require you to choose it.

If we understand the consequent of WEAK AGGLOMERATION (REQUIREMENT) in terms of (1b), the principle is false. It won't always be true that you are rationally required to *choose* to perform a sequence of rationally required actions simply because you won't always be in a position to choose to perform such a sequence. A sequence of actions is an option for you only if there's some time at which you can choose to perform it. And, unless you can choose, at the very beginning, to *commit* yourself to perform the sequence, there will be no time at which performing the sequence is a rationally evaluable option for you. On the other hand, if we understand the consequent of WEAK AGGLOMERATION (REQUIREMENT) in terms of (1a), it isn't violated by **Actual Value Decision Theory**.

Furthermore, proponents of *Supervaluational Expected Utility Theory* can, and should, make use of the very similar distinction made between (2a) and (2b) — given that, as we saw in §3.2, their view, too, violates an agglomeration principle.

3.3.4 A Lingering Worry for Actual Value Decision Theory: non-transitive preferences?

There is yet another worry, however, for **Actual Value Decision Theory.** It looks like you should prefer the sequence $\langle L_1, L_2\rangle$ to the sequence $\langle L_1, R_2\rangle$, and you should prefer the sequence $\langle L_1, R_2\rangle$ to the sequence $\langle R_1, R_2\rangle$. But, according to **Actual Value Decision Theory**, you are not required to prefer $\langle L_1, L_2\rangle$ to $\langle R_1, R_2\rangle$. So, the view appears to violate the transitivity of preference.

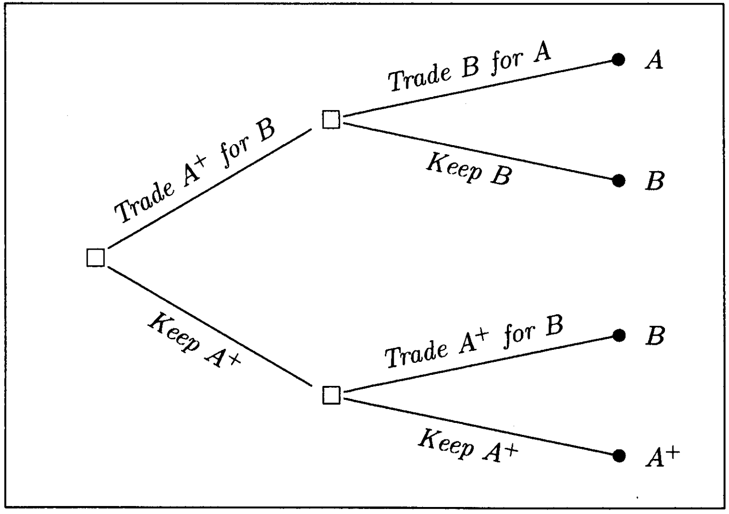
$$\begin{aligned}\langle L_1, L_2\rangle &\succ \langle L_1, R_2\rangle\\ \langle L_1, R_2\rangle &\succ \langle R_1, R_2\rangle\\ \langle L_1, L_2\rangle &\not\succ \langle R_1, R_2\rangle\end{aligned}$$

Why is this a worry? It's not inexplicable why transitivity fails in this case. In comparing $\langle L_1, L_2\rangle$ to $\langle L_1, R_2\rangle$, you are justified in thinking that the actual value of the former might exceed the actual value of the latter to a greater extent than you think the actual value of latter might exceed the actual value of the former. The same goes for the comparison between $\langle L_1, R_2\rangle$ and $\langle R_1, R_2\rangle$. However, in comparing $\langle L_1, L_2\rangle$ to $\langle R_1, R_2\rangle$, you are rationally certain that neither option's actual value exceeds the other.

Although **Actual Value Decision Theory** can tell a story that *explains* why your evaluations of options might sometimes fail to be transitive, aren't there decisive pragmatic reasons to adopt transitive preferences? In particular, if your preferences aren't transitive, can't you be money pumped? In a sense, yes. Here's a way to bring this out. Suppose we give you the Regular box. You don't know which prize it contains. We offer you a trade. For a small fee \$$\epsilon$, you can exchange the Regular box for the Lopsided box, which either contains A, B, and a dollar (if the coin landed heads) or nothing (if the coin landed tails). You should make the trade. You prefer getting A, B, and a dollar to getting B alone more strongly than you prefer A to nothing. Then, we offer you to trade in the Lopsided box for the Larger box. Again, you should pay a small fee to take the deal. Finally, we offer you the opportunity to trade in the Larger box for the Regular box. According to **Actual Value Decision Theory**, it's rationally permissible for you to make this trade. But if you do, you end up where you began minus some small fees!

Note, however, that this is a *weak* money-pump. There are some trades you ought to make and a trade it's *permissible* for you to make such that, if you make all of them, you are guaranteed to end up worse off, by your own lights, than you would have had you not made all the trades. **Actual Value Decision Theory**, then, doesn't *compel* you to make yourself worse off than you could've been; rather, it merely fails to *protect* you from making yourself worse off in this way. That might seem like meager comfort — at least until we notice that if you have incomplete preferences, you're vulnerable to these kind of weak money pumps anyway. This is a problem for *Supervaluational Expected Utility Theory* too. Suppose you start out with A^+. We offer you the opportunity to trade it in for B. It's permissible for you to make the trade. Then, we offer you the opportunity to trade B for A. Again, it's permissible for you to make the trade. But, if you make both trades, you will end up with A, which is something you strictly disprefer to A^+. You've, effectively, been (weakly) money pumped of a dollar.

(WEAK) MONEY PUMP FOR INCOMPLETE PREFERENCES



If you have incomplete preferences, you are vulnerable to being (weakly) money pumped. You are not, however, necessarily vulnerable to being *strongly* money pumped. Furthermore, having non-transitive — but nevertheless *acyclic* — preferences leaves you vulnerable to being weakly, but not strongly, money pumped. So the mere fact that **Actual Value Decision Theory** won't protect you from being weakly money pumped doesn't, by itself, provide us with a reason to reject it.

There is a lingering worry for **Actual Value Decision Theory**, however. As mentioned, it's not merely that the view fails to protect you from being weakly money pumped; rather, it recommends adopting *non*-transitive instrumental preferences. According to **Actual Value Decision Theory**, you should prefer the Larger box to the Lopsided box, and you should prefer the Lopsided box to the Regular box, but you are not rationally required to prefer the Larger box to the Regular box. Absent some further compelling reasons to think that non-transitive instrumental preferences are *ipso facto* irrational, I'm inclined to accept the consequences: being instrumentally rational sometimes involves having non-transitive instrumental preferences.

3.4 Conclusion

This chapter defended **Actual Value Decision Theory** against two different objections: the Reasons Argument and the Agglomeration Argument.

In defending the view from these objections, we arrived at some interesting consequences of taking the **Actual Value Conception** seriously. First, being instrumentally rational isn't entirely about correctly responding to the reasons you *have*; rather, you should aim, as best you can, to do what *there is* most reason to do. Second, being instrumentally rational might sometimes involve evaluating your options non-transitively: in some cases, you should have non-transitive instrumental preferences.

Appendix A

The Actual Value Conception & Causal Decision Theory

In this section, I prove the following claim:

> *Your unconditional estimate of the extent to which the actual value of ϕ exceeds the actual value of ψ is greater than your unconditional estimate of the extent to which the actual value of ψ exceeds the actual value of ϕ* if and only if *the causal expected utility of ϕ is greater than the causal expected utility of ψ.*

Recall our notion of *actual value:*

$$V_{@}(\phi) = V(\phi \wedge K_{@})$$

The proposition that $V_{@}(\phi) = v$, then, is equivalent to the proposition that $V(\phi \wedge K_{@}) = v$. In turn, *that* proposition is equivalent to the following disjunction of conjunctions:

$$\bigvee_{K_i} \Big(V(\phi \wedge K_i) = v \wedge K_i \Big)$$

Similarly, the proposition that $\mathcal{CV}_{@}(\phi, \psi) = v^*$ is equivalent to the following disjunction of conjunctions:

$$\bigvee_{K_i} \Big(\mathcal{CV}_{K_i}(\phi, \psi) = v^* \wedge K_i \Big)$$

Because the dependency hypotheses, in virtue of being a partition, are mutually exclusive and mutually exhaustive, exactly one such K holds (which we've been calling $K_{@}$). Furthermore, because the dependency hypotheses are mutually exclusive, each of this disjunction's disjuncts are mutually exclusive. Consequently, your credence that $\mathcal{CV}_{@}(\phi, \psi) = v^*$ can be expressed as a sum of your credences in each of the disjuncts.

$$Cr\Big(\mathcal{CV}_{@}(\phi, \psi) = v^*\Big) = \sum_K Cr\Big(\mathcal{CV}_K(\phi, \psi) = v^* \wedge K\Big)$$

Furthermore, assume (as we've implicitly been doing) that you are *self-aware:* you know how the values you assign to the various possible outcomes compare to one another. In other words, we take your credences in propositions of the form $\mathcal{CV}_K(\phi, \psi) = v^*$ to be maximally opinionated and accurate:

$$Cr\Big(\mathcal{CV}_K(\phi, \psi) = v^*\Big) = \begin{cases} 1 & \text{if } \mathcal{CV}_K(\phi, \psi) = v^* \\ 0 & \text{otherwise} \end{cases}$$

Therefore, your credence that *both* $\mathcal{CV}_K(\phi, \psi) = v^*$ *and* K should, likewise, be zero when $\mathcal{CV}_K(\phi, \psi) \neq v^*$ but equal $Cr(K)$ when $\mathcal{CV}_K(\phi, \psi) = v^*$.

$$Cr\Big(\mathcal{CV}_K(\phi, \psi) = v^* \wedge K\Big) = \begin{cases} Cr(K) & \text{if } \mathcal{CV}_K(\phi, \psi) = v^* \\ 0 & \text{otherwise} \end{cases}$$

Let $\mathbb{1}[q]$ be an *indicator function* that returns 1 if q is true and 0 if q is false.

$$\mathbb{1}[q] = \begin{cases} 1 & \text{if } q \\ 0 & \text{otherwise} \end{cases}$$

Using this indicator function, we can express your credences in propositions about the actual values of your options in terms of your credences in dependency hypotheses. In particular,

$$Cr\Big(\mathcal{CV}_{@}\left(\phi,\psi\right)=v^{*}\Big)=\sum_{K}\mathbb{1}\left[\mathcal{CV}_{K}\left(\phi,\psi\right)=v^{*}\right]\cdot Cr\left(K\right)$$

In other words, your credence that the extent to which the actual value of ϕ exceeds the actual value of ψ is v^* should be equal to the sum of your credences in the dependency hypotheses in which the difference in value between the outcome of ϕ and the outcome of ψ is v^*.

This allows us to rewrite ACTUAL VALUE ESTIMATE in terms of your credences in dependency hypotheses, as follows:

$$\begin{aligned}\text{ESTIMATE}\Big[\mathcal{CV}_{@}\left(\phi,\psi\right)\Big]&=\sum_{v^*}Cr(\mathcal{CV}_{@}\left(\phi,\psi\right)=v^{*})\cdot v^{*}\\&=\sum_{v^*}\Big(\sum_{K}\mathbb{1}\left[\mathcal{CV}_{K}\left(\phi,\psi\right)=v^{*}\right]\cdot Cr\left(K\right)\Big)\cdot v^{*}\end{aligned}$$

For each possible value v^*, the term $\sum_K \mathbb{1}\left[\mathcal{CV}_K\left(\phi,\psi\right)=v^*\right]\cdot Cr\left(K\right)\cdot v^*$ equals $\sum_K Cr(K)\cdot\mathcal{CV}_K\left(\phi,\psi\right)$ if $\mathcal{CV}_K\left(\phi,\psi\right)=v^*$ and, otherwise, it equals zero.

And, so,

$$\begin{aligned}\text{ESTIMATE}\Big[\mathcal{CV}_{@}\left(\phi,\psi\right)\Big]&=\sum_{v^*}\sum_{K}\mathbb{1}\left[\mathcal{CV}_{K}\left(\phi,\psi\right)=v^{*}\right]\cdot Cr\left(K\right)\cdot\mathcal{CV}_{K}\left(\phi,\psi\right)\\&=\sum_{K}Cr\left(K\right)\cdot\mathcal{CV}_{K}\left(\phi,\psi\right)\cdot\sum_{v^*}\mathbb{1}\left[\mathcal{CV}_{K}\left(\phi,\psi\right)=v^{*}\right]\end{aligned}$$

Furthermore, because, for each dependency hypothesis K, there is *exactly one* possible value v^* such that $\mathcal{CV}_K\left(\phi,\psi\right)=v^*$,

$$\sum_{v}\mathbb{1}\left[\mathcal{CV}_{K}\left(\phi,\psi\right)=v^{*}\right]=1$$

Therefore, the (unconditional) estimate of the extent to which the actual value of ϕ exceeds the actual value of ψ is equal to the difference between ϕ's and ψ's causal expected utilities:[1]

$$\begin{aligned}
\textsc{Estimate}\Big[\mathcal{CV}_{@}(\phi,\psi)\Big] &= \sum_{v^*} Cr\big(\mathcal{CV}_{@}(\phi,\psi) = v^*\big)\cdot v^* \\
&= \sum_{K} Cr(K)\cdot \mathcal{CV}_{K}(\phi,\psi) \\
&= \sum_{K} Cr(K)\cdot \big(V(\phi\wedge K) - V(\psi\wedge K)\big) \\
&= \sum_{K} Cr(K)\cdot V(\phi\wedge K) - \sum_{K} Cr(K)\cdot V(\psi\wedge K)
\end{aligned}$$

According to the **Actual Value Conception,** you should prefer option ϕ to option ψ if and only if $\textsc{Estimate}\Big[\mathcal{CV}_{@}(\phi,\psi)\Big] > \textsc{Estimate}\Big[\mathcal{CV}_{@}(\psi,\phi)\Big]$.

$$\textsc{Estimate}\Big[\mathcal{CV}_{@}(\phi,\psi)\Big] > \textsc{Estimate}\Big[\mathcal{CV}_{@}(\psi,\phi)\Big]$$

$$\sum_{K} Cr(K)\cdot V(\phi\wedge K) - \sum_{K} Cr(K)\cdot V(\psi\wedge K) > \sum_{K} Cr(K)\cdot V(\psi\wedge K) - \sum_{K} Cr(K)\cdot V(\phi\wedge K)$$

$$2\cdot \sum_{K} Cr(K)\cdot V(\phi\wedge K) > 2\cdot \sum_{K} Cr(K)\cdot V(\psi\wedge K)$$

$$\sum_{K} Cr(K)\cdot V(\phi\wedge K) > \sum_{K} Cr(K)\cdot V(\psi\wedge K)$$

$$U(\phi) > U(\psi)$$

So, the **Actual Value Conception** entails causal decision theory: you should prefer ϕ to ψ if and only if the causal expected utility of ϕ is greater

[1] Or, rather, this equivalence holds when the actual values of your options are well-defined, so that $\mathcal{CV}_{K}(\phi,\psi) = V(\phi\wedge K) - V(\psi\wedge K)$.

than the causal expected utility of ψ.

THIS PAGE INTENTIONALLY LEFT BLANK

Appendix B

Benchmark Decision Theory & The Actual Value Conception

Ralph Wedgwood [Wedgwood, 2013] defends a decision theory that, much like evidential decision theory, uses *conditional probabilities* but that, much unlike evidential decision theory, conforms to the **Actual Value Conception** by measuring the value of an option in a state *comparatively*.

> **Benchmark Decision Theory**
> You should prefer an option ϕ to an option ψ just in case
>
> $$\sum_K Cr(K \mid \phi) \cdot (V(\phi \wedge K) - b_K) > \sum_K Cr(K \mid \psi) \cdot (V(\psi \wedge K) - b_K)$$
>
> Where b_K is a "benchmark" value in state K. Wedgwood suggests that, when there are only two options under consideration, we can take b_K to be the *average* of the values of the outcomes of ϕ and ψ in K:
>
> $$b_K = \frac{V(\phi \wedge K) + V(\psi \wedge K)}{2}$$

Let's write $V_B(\phi)$ to denote the "benchmark" expected value of ϕ. (That is: $V_B(\phi) = \sum_K Cr(K \mid \phi) \cdot (V(\phi \wedge K) - b_K)$.) When you choosing only between two options, Wedgwood [2013] recommends identifying b_K, the benchmark value in state K, with the average of the values of the outcomes in K. However, when there are three or more options under consideration, a more complicated procedure is necessary to generate an appropriate "benchmark." Wedgwood [2013] offers a couple suggestions for how this procedure might go. The argument in this section, however, pertains only to what benchmark decision theory says in the simple two-option case.

According to Benchmark Decision Theory, you should prefer ϕ to ψ (when those are the only two options under consideration) just in case:

$$\sum_K Cr(K \mid \phi) \cdot \Big(V(\phi \wedge K) - avg\left(V_K(\phi), V_K(\psi)\right)\Big) > \sum_K Cr(K \mid \psi) \cdot \Big(V(\psi \wedge K) - avg\left(V_K(\phi), V_K(\psi)\right)\Big)$$

$$\sum_K Cr(K \mid \phi) \cdot \left(\frac{V(\phi \wedge K) - V(\psi \wedge K)}{2}\right) > \sum_K Cr(K \mid \psi) \cdot \left(\frac{V(\psi \wedge K) - V(\phi \wedge K)}{2}\right)$$

$$\sum_K Cr(K \mid \phi) \cdot \Big(V(\phi \wedge K) - V(\psi \wedge K)\Big) > \sum_K Cr(K \mid \psi) \cdot \Big(V(\psi \wedge K) - V(\phi \wedge K)\Big)$$

$$\sum_K Cr(K \mid \phi) \cdot \mathcal{CV}_K(\phi, \psi) > \sum_K Cr(K \mid \psi) \cdot \mathcal{CV}_K(\psi, \phi)$$

Furthermore, $\sum_K Cr(K \mid \phi) \cdot \mathcal{CV}_K(\phi, \psi)$ is equivalent to your *conditional estimate* of the extent which the actual value of option ϕ exceeds the actual value of ψ.[1] And so, if your estimates are conditional estimates, the "benchmark value" of an option (at least when there are only two options under consideration) equals your estimate of the extent to which that option's actual value exceeds the actual value of the other option under consideration.

Wedgwood's benchmark decision theory conforms to the **Actual Value Conception**. It uses conditional, rather than unconditional, estimates. Causal decision theory, as we've seen, also conforms to the **Actual Value Conception**. Evidential decision theory, on the other hand, does not.

[1] The proof is analogous to the one presented in the previous section.

Appendix C

The Actual Value Conception & Credence Preference Coherence

C.1 Causal Decision Theory Entails Credence Preference Coherence

According to the unconditional version of the **Actual Value Conception,** you should prefer ϕ to ψ if and only if $\sum_K Cr(K) \cdot \mathcal{CV}_K(\phi, \psi) > 0$. In order to show that the principle follows, we'll assume that $Cr\big(V_{@}(\phi) > V_{@}(\psi)\big) = 0$ and, then, show that $\sum_K Cr(K) \cdot \mathcal{CV}_K(\phi, \psi) \not> 0$.

Assume that $Cr\big(V_{@}(\phi) > V_{@}(\psi)\big) = 0$.

$$Cr\big(V_{@}(\phi) > V_{@}(\psi)\big) = \sum_K \mathbb{1}\big[V(\phi \wedge K) > V(\psi \wedge K)\big] \cdot Cr(K)$$

So, if $Cr\big(V_{@}(\phi) > V_{@}(\psi)\big) = 0$, then $\sum_K \mathbb{1}\big[V(\phi \wedge K) > V(\psi \wedge K)\big] \cdot Cr(K) = 0$.

And, $\sum_K \mathbb{1}\big[V(\phi \wedge K) > V(\psi \wedge K)\big] \cdot Cr\,(K) = 0$ just in case, for each dependency hypothesis K, either:

1. $\mathbb{1}\big[V(\phi \wedge K) > V(\psi \wedge K)\big] = 0$, or

2. $Cr(K) = 0$, (or both). For each K, if $\mathbb{1}\big[V(\phi \wedge K) > V(\psi \wedge K)\big] = 0$, then $\mathcal{CV}_K\,(\phi, \psi) \not> 0$.

And so, $Cr(K) \cdot \mathcal{CV}_K\,(\phi, \psi) \not> 0$. Also, for each K, if $Cr(K) = 0$, then $Cr(K) \cdot \mathcal{CV}_K\,(\phi, \psi) = 0$.

Therefore, for every dependency hypothesis K, $Cr(K) \cdot \mathcal{CV}_K\,(\phi, \psi) \not> 0$. And thus, $\sum_K Cr(K) \cdot \mathcal{CV}_K\,(\phi, \psi) \not> 0$.

Therefore, if $Cr\big(V_@(\phi) > V_@(\psi)\big) = 0$, then $\sum_K Cr(K) \cdot \mathcal{CV}_K\,(\phi, \psi) \not> 0$. And, so, according to the **Actual Value Conception,** you should not prefer option ϕ to option ψ.

C.2 Benchmark Decision Theory Entails Credence Preference Coherence

I will show that if $Cr\big(V_@(\phi) > V_@(\psi)\big) = 0$, then Wedgwood's Benchmark Decision Theory will say that you shouldn't strictly prefer ϕ-ing to ψ-ing.

> **Claim:** If $Cr\big(V_@(\phi) > V_@(\psi)\big) = 0$, then $V_B(\phi) \not> V_B(\psi)$

First, recall that if $Cr\big(V_@(\phi) > V_@(\psi)\big) = 0$, then, for all dependency hypotheses K,

$$\mathbb{1}\big[V(\phi \wedge K) > V(\psi \wedge K)\big] \cdot Cr\,(K) = 0$$

If $\mathbb{1}\big[V(\phi \wedge K) > V(\psi \wedge K)\big] \cdot Cr\,(K) = 0$, then either $\mathbb{1}\big[V(\phi \wedge K) > V(\psi \wedge K)\big] = 0$, or $Cr\,(K) = 0$, or both.

1. If $\mathbb{1}\big[V(\phi \wedge K) > V(\psi \wedge K)\big] = 0$, then $V(\phi \wedge K) - V(\psi \wedge K) \leq 0$, or

2. If $Cr\,(K) = 0$, then $Cr(K \mid \phi) = Cr(K \mid \psi) = 0$.

Second, if we, following Wedgwood, take b_K to be the average of the values of the outcomes of ϕ and ψ in K, then the benchmark expected value of an option ψ can be rewritten as follows:[1]

$$\begin{aligned} V_B(\phi) &= \sum_K Cr(K \mid \phi) \cdot \big(V(\phi \wedge K) - b_K\big) \\ &= \sum_K Cr(K \mid \phi) \cdot \left(V(\phi \wedge K) - \frac{V(\phi \wedge K) + V(\psi \wedge K)}{2}\right) \\ &= \sum_K Cr(K \mid \phi) \cdot \left(\frac{V(\phi \wedge K) - V(\psi \wedge K)}{2}\right) \end{aligned}$$

Finally, $V_B(\phi) > V_B(\psi)$ just in case $V_B(\phi) - V_B(\psi) > 0$.

$$\sum_K Cr(K \mid \phi) \cdot \left(\frac{V(\phi \wedge K) - V(\psi \wedge K)}{2}\right) - \sum_K Cr(K \mid \psi) \cdot \left(\frac{V(\psi \wedge K) - V(\phi \wedge K)}{2}\right) > 0$$

$$\sum_K \left(\frac{V(\phi \wedge K) - V(\psi \wedge K)}{2}\right) \cdot \Big(Cr(K \mid \phi) + Cr(K \mid \psi)\Big) > 0$$

$$\sum_K \Big(V(\phi \wedge K) - V(\psi \wedge K)\Big) \cdot \Big(Cr(K \mid \phi) + Cr(K \mid \psi)\Big) > 0$$

As established above, if $Cr\big(V_{@}(\phi) > V_{@}(\psi)\big) = 0$, then, for all K, either, $V(\phi \wedge K) - V(\psi \wedge K) \leq 0$, or $Cr(K \mid \phi) = Cr(K \mid \psi) = 0$, or both. This means that, for all K,

$$\Big(V(\phi \wedge K) - V(\psi \wedge K)\Big) \cdot \Big(Cr(K \mid \phi) + Cr(K \mid \psi)\Big) \leq 0$$

[1] The benchmark value in K needn't be the unweighted average of the values of the outcomes in K in order for the proof to go through. Any weighted average — just so long as the same weights are used in every K — will work just as well.

Therefore,

$$\sum_K \Big(V(\phi \wedge K) - V(\psi \wedge K)\Big) \cdot \Big(Cr(K \mid \phi) + Cr(K \mid \psi)\Big) \leq 0$$

So, if $Cr\big(V_{@}(\phi) > V_{@}(\psi)\big) = 0$, then $V_B(\phi) \not> V_B(\psi)$.

Appendix D

Strong Competitiveness & the Actual Value Conception

Bales et al. [2014] argues against *Supervaluational Expected Utility Theory* by appealing to a plausible dominance-type principle. They don't offer an alternative decision theory that coheres with their dominance principle. However, they do point in the direction of developing an alternative by suggesting a further principle, called STRONG COMPETITIVENESS.

> [STRONG COMPETITIVENESS]
>
> If one or more actions are competitive, and other actions are not competitive, it is rationally required to perform a competitive action.

I think this is false. Consider the following decision problem. You have two options: ϕ and ψ. A lottery ticket will be randomly drawn from a pool of 100. If you take option ϕ, then if ticket #100 is selected, you will get A^+ as a prize; otherwise, you will get B as a prize. If you take option ψ, then no matter which ticket it selected, you will get prize A.

	TICKETS #1-#99	TICKET #100
ϕ	B	A^+
ψ	A	A

Option ϕ is competitive — for every the world could be (no matter which ticket is drawn) its consequences are no worse than the consequences of ψ. Option ψ, however, is not competitive — if ticket #100 is selected, then the consequences of ψ are worse than the consequences of ϕ. Nevertheless, it doesn't seem like option ϕ is rationally required in this case. There is a 99% chance that if you take option ϕ you will get B, which you regard as no better or worse than A. But you are not *indifferent* between B and A, either. They are on a par. Even though there is a small chance of getting something you prefer by taking ϕ, it's not obvious to me that this is enough to outweigh the parity between the much more likelier outcomes.

Furthermore, if we accept a weak kind of "averaging" principle, then Bales et al. [2014]'s two principles are inconsistent.

> [Weak Averaging]
>
> For any options ϕ, ψ, if, for all states K, $\phi \succ (\psi \wedge K)$ and $(\phi \wedge K) \succ \psi$, then $\phi \succ \psi$.

This principle says that if you prefer taking option ϕ to any, and all, the outcomes that might result from taking ψ, and you prefer all the outcomes that might result from taking ϕ to taking ψ, then you should prefer ϕ to ψ. The principle is plausible, even in the face of parity, for the following reasons: (1) if you prefer taking option ϕ to each of the outcomes that might result from taking ψ, then should you prefer ϕ to the (degenerate) gamble that pays out ψ's *best* outcome in every state; so, you'd prefer to take your chances with ϕ than receive ψ's best outcome. (2) If you prefer all the outcomes that might result from taking ϕ to taking ψ, then should prefer the (degenerate) gamble that pays out ϕ's *worst* outcome in every state to ψ; so, you'd prefer receiving ϕ's worst outcome to taking your chances with ψ. Furthermore, you shouldn't strictly prefer an option to its *best* outcome, and you shouldn't strictly prefer an option's *worst* outcome to that option itself. (That's what makes "Averaging" an appropriate name for this principle: the value you assign to an option should, in some sense, be "in between" the values you assign to its best and worst potential outcomes).

Let ϕ be the gamble that corresponds to taking the Larger box — that is: the gamble that pays out A^+ if heads and B^+ if tails. And let ψ be the gamble corresponding to taking the Regular box — that is: the gamble that pays out B if heads and A if tails. If Strong Competitiveness is right, then $\phi \succ (\psi \wedge H)$, $\phi \succ (\psi \wedge T)$, $(\phi \wedge H) \succ \psi$, $(\phi \wedge T) \succ \psi$.

	Heads	Tails
ϕ	A^+	B^+
$(\psi \wedge H)$	B	B
$(\psi \wedge T)$	A	A

	Heads	Tails
ψ	B	A
$(\phi \wedge H)$	A^+	A^+
$(\phi \wedge T)$	B^+	B^+

By Weak Averaging, then, $\phi \succ \psi$. But, relative to each other, both gambles are competitive. So both are rationally permissible and, hence, the former cannot be strictly preferred to the latter. I think this suggests that Strong Competitiveness is too strong.

Actual Value Decision Theory, also, conflicts with Strong Competitiveness. That's good because the principle, as the example above hopefully illustrates, is false.

THIS PAGE INTENTIONALLY LEFT BLANK

Appendix E

The In-the-Long-Run Argument

Here is one last argument against **Actual Value Decision Theory.** In-the-Long-Run Arguments, which have been used to support *Expected Utility Theory,* generally appeal to what rationality would require of you *were* you to face a particular decision-problem an infinite number of times. They, then, draw a conclusion about what rationality requires of you in the one-off version of the decision-problem. (See [Briggs, 2014] for a general overview).

In particular, these arguments appeal to the following mathematical fact:

> In the long run — as the number of trials goes to infinity — the average value of an option converges (almost surely) to its expected value.

They, then, make the following claim: Even if you're not facing the same decision again and again an infinite number of times, you should still choose the option that *would* do the best, on average, in the long run. Together, these two entail that you should maximize expected utility.

A structurally analogous argument can be given for taking the Larger box over the Regular box. If it's sound, then, because **Actual Value Decision Theory** entails that you are not required to take the Larger box over the Regular box, we should reject **Actual Value Decision Theory.** I will argue that the argument does not succeed, and that, in its most plausible form, it is version of the Agglomeration Argument.

Let's, first, see how the argument is meant to go. Suppose that you are facing the decision between the Larger box and the Regular box an infinite number of times. It's tempting to think that the policy of always choosing the Larger box will, on average, net you more value than the policy of always choosing the Regular box. Imagine, for example, that you've faced the decision one hundred times; and suppose (somewhat unrealistically, but for the ease of presentation) that the coin has landed heads and tails an equal number of times. If you take the Larger box every time, you'll end up with fifty A^+s and fifty B^+s; if you take the Regular box every time, you'll end up with fifty As and fifty Bs. You do prefer the former package of goods to the latter. The average value of each option can be found by dividing the value of each package by one hundred. It's not clear, given your incomplete preferences, what the values of the two packages are. But no matter which specific values are assigned, the average value of the former will be higher than the average value of the second.

But that argument can't be right. We shouldn't be comparing the values of the packages of goods you stand to accrue divided by the number of trials; rather, we should compare the sum of the values of the goods won on each occasion divided by the number of trails. Otherwise, as pointed out in [Buchak, 2013], the In-the-Long-Run Argument would support maximizing expected *dollar value* rather than expected *utility.*

> **Example:** Suppose a fair coin will be flipped. If you take Option 1, then you get \$4 no matter what. If you take Option 2, you get \$10 if the coin lands heads and \$0 otherwise. After, say, 100 trials you'll (roughly) have \$500 if you take Option 2, and only \$400 if you take Option 1. But suppose, also, that for you the value of money marginally decreases — e.g., \$10 isn't twice as good as \$5. It might very well be, then, that the expected *utility* of Option 1 is greater than that of Option 2's. We should *not* look at how much money you would have after n trails; rather, we should look at how much *utility* you would have. After 100 trials, Option 1 will net you $100 \cdot u(\$4)$ and Option 2 will net you (roughly) $50 \cdot u(\$0) + 50 \cdot u(\$10)$.

Any argument which entails that you are rationally required to maximize expected dollar value is surely false. But the version of the In-the-Long Argument presented above — which appeals to the fact that you would prefer the bundle of vacation prizes that would result from the policy of always taking the Larger box to the bundle of prizes that would result from the policy of always taking the Regular box — *does* entail that you would be required to maximized expected dollar value. Therefore, it cannot be right.

What about a version of the In-the-Long-Run Argument that compares the sum of the values of the goods won on each occasion (as opposed to value of the total package of goods that would result)? Could such an argument be used to show that rationality requires you to prefer taking the Larger box to the Regular box? No. Because you have incomplete preferences, you cannot assign (precise) values to the potential outcomes of your options. Because you cannot assign real-numbered values to the potential outcomes, we cannot take the sum of these values. It is unclear, then, that the policy of always taking the Larger box would do better *on average*, in the long run, than the policy of taking the Regular box. It longer makes sense to talk about the policy that would do better "on average" because "the average" is not well-defined. Furthermore, in each trial, whatever prize you would get from taking the Larger box is not something you prefer to the prize you would have received had you, that trial, taken the Regular box instead. And so it's not just *unclear* that the policy of always taking the Larger box would do better on average, in the long run, the policy of always taking the Regular box; rather, it seems to be false. Here's a way to bring this out. After each trial, suppose, we ask you if you are unequivocally happy about how it turned out. You never are. So, at least in this sense, the sweetened option does no better on average in the long run than the unsweetened one.

Nevertheless, it is still is true that, in the long run, the policy of always taking the Larger box would result in a package of goods that you do prefer to the package that would result from the policy of always taking the Regular box. By always taking the Larger box, you will almost surely end up with a bundle of prizes that you strictly prefer to the bundle of prizes you would get were you to always take the Regular box. However, according to **Actual Value Decision Theory**, on each occasion, it is rationally permissible to take the Regular box rather than the Larger box. But notice that on this way of cashing out the "In-the-Long-Run" idea, it is a cousin of the Weak Agglomeration Argument, presented in §3. That section discussed a case in which there was a sequence of actions such that it was permissible to perform the sequence but impermissible to perform any of the individual actions constituting it. Here, we have a sequence of actions such that it is permissible to perform each of the actions in the sequence, but impermissible to perform the sequence itself. Even still, what was said about the Agglomeration Argument there equally well applies here. If you are able to *choose* to adopt the policy of always taking the Larger box, you are rationally required to do so. However, if, on each occasion, you are only able to choose between taking the Larger box or the Regular box *on that occasion*, it's possible for you to end up performing a sequence other than the one corresponding to the policy of always taking the Larger box without making any irrational choices.

The In-the-Long Argument fails to establish that you are rationally required to prefer taking the Larger box to taking the Regular box.

Bibliography

F. Anscombe and R. Aumann. A definition of subjective probability. *Annals of Mathematical Statistics*, pages 199–205, 1963.

Francis J Anscombe and Robert Aumann. A definition of subjective probability.

R. Aumann. Utility theory without the completeness axiom. *Econometrica*, 30(3):445–462, 1962.

A. Bales, D. Cohen, and Toby Handfield. Decision theory of agents with incomplete preferences. *Australasian Journal of Philosophy*, 92(3):453–470, 2014.

Chip Heath Brad M. Barber and Terrance Odean. Good reasons sell: Reason-based choice among group and individual investors in the stock market. *Management Science*, 49(12):1636–1652, 2003.

Rachel Briggs. Normative theories of rational choice: Expected utility, Aug 2014. URL `http://plato.stanford.edu/entries/rationality-normative-utility/`.

John Broome. Is incommensurability vagueness? In Ruth Chang, editor, *Incommensurability, Incomparability, and Practical Reason*. Harvard University Press, 1997.

John Broome. Does rationality consist in responding correctly to reasons? *Journal of Moral Philosophy*, 4(3):349–374, 2007.

John Broome. *Rationality Through Reasoning*. Wiley-Blackwell, 2013.

Lara Buchak. *Risk and Rationality*. Oxford University Press, 2013.

K Bykvist. No good fit: Why the fitting attitude analysis of value fails. *Mind*, 118(469):1–30, 2009.

Ruth Chang. The possibility of parity. *Ethics*, 112(4):659–688, 2002.

Ruth Chang. Parity, interval value, and choice. *Ethics*, 115(2):331–350, 2005.

Ruth Chang. Voluntarist reasons and the sources of normativity. In D. Sobel and D. Wall, editors, *Reasons for Action*, pages 243–271. Oxford University Press, 2009.

J. Dubra, F. Maccheroni, and E. Ok. Expected utility theory without the completeness axiom. *Journal of Economic Theory*, 115(1):118–133, 2004.

Itamar Simonson Eldar Shafir and Amos Tversky. Reason-based choice. *Cognition*, 49:11–36, 1993.

Ozgur Evren and Efe Ok. On the multi-utility representation of preference relations. *Journal of Mathematical Economics*, 47:554–563, 2011.

Robert Pargetter Frank Jackson. Oughts, options, and actualism. *Philosophical Review*, 95:233–255, 1986.

Tsogbadral Galaabaatar and Edi Karni. Subjective expected utility theory with incomplete preferences. *Econometrica*, 81(1):255–284, 2013.

J. Gert. Value and parity. *Ethics*, 114(3):492–510, 2004.

Allan Gibbard and William Harper. Counterfactuals and two kinds of expected utility. In JJ. Leach C.A. Hooker and E.F. McClennen, editors, *Foundations and Applications of Decision Theory, Vol I.* Dordrecht, 1978.

I.J. Good. Rational decisions. *Journal of the Royal Statistical Society*, B (14):107–114, 1952.

Caspar Hare. Take the sugar. *Analysis*, 70(2):237–247, 2010.

Caspar Hare. *The Limits of Kindness.* Oxford University Press, 2013.

Caspar Hare. Should we wish well to all? *The Philosophical Review*, Forthcoming, 2015.

H. Herzberger. Ordinal preference and rational choice. *Econometrica*, 41: 187–237, 1973.

Frank Jackson. Decision-theoretic consequentialism and the nearest and dearest objection. *Ethics*, 101:461–82, 1991.

William James. The will to believe. In *The Will to Believe and Other Essays in Popular Philosophy.* Dover Publications, New York, 1956 1896.

Richard Jeffrey. *The Logic of Decision.* University of Chicago Press, 1983.

James Joyce. *The Foundations of Causal Decision Theory.* Cambridge University Press, 1999.

James Joyce. A defense of imprecise credences in inference and decision making. *Philosophical Perspectives*, 24(1):281–323, 2010.

Nico Kolodny and John McFarlane. Ifs and oughts. *The Journal of Philosophy*, 108(3):115–143, 2010.

J.S. Lerner and P.E. Tetlock. Accounting for the effects of accountability. *Psychological Bulletin*, 125:255–275, 1999.

Isaac Levi. Imprecision and indeterminacy in probability judgment. *Philosophy of Science*, 36:331–340, 1985.

Isaac Levi. *Hard Choices.* Cambridge University Press, Cambridge, 1986.

Isaac Levi. Value commitmnts, value conflict, and the separability of belief and value. *Philosophy of Science*, 66:509–533, December 1999.

Isaac Levi. Symposium on cognitive rationality: Part i minimal rationality. *Mind and Society*, 5:199–211, August 2006.

Isaac Levi. Why rational agents should not be liberal maximizers. *Canadian Journal of Philosophy*, 38(Supplementary Vol 34):1–17, 2008.

David Lewis. Causal decision theory. *Australasian Journal of Philosophy*, 59 (1):5–30, 1981.

Errol Lord. Having reasons and the factoring account. *Philosophical Studies*, 149(3):283–296, 2010.

Robert Nau. The shape of incomplete preferences. *The Annals of Statistics*, pages 2430–2448, 2006.

E. Ok, P. Ortoleva, and Gil Riella. Incomplete preferences under uncertainty: Indecisiveness in beliefs versus tastes. *Econometrica*, 80(4):1791–1808, 2012.

Derek Parfit. Rationality and reasons. In Dan Egonsson et al., editor, *Exploring Practical Philosophy*, pages 17–39. Ashgate, 2001.

Richard Pettigrew. Risk, rationality and expected utility theory. *Canadian Journal of Philosophy*, 45(5-6):798–826, 2015.

Wlodek Rabinowicz. Incommensurability meets risk. unpublished, July 2016.

J. Raz. Value incommensurability: Some preliminaries. *Proceedings in the Aristotelian Society*, 86:117–134, 1985.

Joseph Raz. *Engaging Reason.* Oxford University Press, Oxford, 1999.

Michael D. Resnik. *Choices: an Introduction to Decision Theory.* University of Minnesota Press, 1987.

Susanna Rinard. A decision theory for imprecise credences. *Philosopher's Imprint*, 15(7):1–16, 2015.

Leonard J Savage. *The Foundations of Statistics.* Wiley, 1954.

T.M. Scanlon. *What We Owe to Each Other.* Harvard University Press, 1998.

T.M. Scanlon. *Being Realistic about Reasons.* Oxford University Press, 2014.

Miriam Schoenfield. Decision making in the face of parity. *Philosophical Perspectives*, 28(1):263–277, 2014.

Mark Schroeder. Having reasons. *Philosophical Studies*, 139(1):57–71, 2008.

Mark Schroeder. Buck-passers' negative thesis. *Philosophical Explorations*, 12(3):341–7, 2009.

Teddy Seidenfeld, M.J. Schervish, and J.B. Kadane. Decisions without ordering. In W. Sieg, editor, *Acting and Reflecting: the interdisciplinary turn*, pages 143–170. Kluwer Academic Publishing, 1990.

Teddy Seidenfeld, M.J. Schervish, and J.B. Kadane. A representation of partially ordered preferences. *The Annals of Statistics*, 23(6):2168–2217, 1995.

Amartya Sen. Incompleteness and reasoned choice. *Synthese*, 140(1/2):43–59, May 2004.

Kieran Setiya. *Reasons without Rationalism.* Princeton University Press, Princeton, NJ, 2007.

Kieran Setiya. What is a reason to act? *Philosophical Studies*, 167:221–235, 2014.

Itamar Simonson and Stephen Nowlis. The role of explanations and need for uniqueness in consumer decision making: Unconventional choices based on reasons. *Journal of Consumer Research*, 27, 2000.

Brian Skyrms. The value of knowledge. *Minnesota Studies in the Philosophy of Science*, 14:245–66, 1990.

Michael Smith. The humean theory of motivation. *Mind*, 96(381):36–61, 1987.

Michael Smith. Internal reasons. *Philosophy and Phenomenological Research*, 55(1):109–131, 1995.

Jordan Howard Sobel. Probability, chance and choice: A theory of rational agency. unpublished, May 1978.

Jack Spencer and Ian Wells. An argument for two-boxing. manuscript, June 2016.

Robert Stalnaker. Letter to david lewis. In Robert Stalnaker William Harper and Glenn Pearce, editors, *Ifs: Conditionals, Belief, Decision, Chance, and Time*. Dordrecht, 1981.

J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

Brian Weatherson. Decision making with imprecise probabilities. manuscript, November 2008.

Ralph Wedgwood. Gandalf's solution to the newcomb problem. *Synthese*, 190(14):2643–75, 2013.

Michael Zimmerman. The good and the right. *Utilitas*, 19(3):326–53, 2007.