Evidence and Rationalization

Ian Wells

Forthcoming in *Philosophical Studies*

Abstract

Suppose that you have to take a test tomorrow but you do not want to study. Unfortunately you should study, since you care about passing and you expect to pass only if you study. Is there anything you can do to make it the case that you should not study? Is there any way for you to 'rationalize' slacking off? I suggest that such rationalization is impossible. Then I show that if evidential decision theory is true, rationalization is not only possible but sometimes advisable.

1

Suppose that you have to take a test tomorrow but you do not want to study. Unfortunately you should study, since you care about passing and you expect to pass only if you study. Is there anything you can do to make it the case that you should not study? Is there any way for you to 'rationalize' slacking off?

If you could somehow stop caring about passing the test, then you would be under no rational pressure to study. In general, changing what one cares about is a way of changing what one rationally ought to do. But changing what one cares about is not always an option. It's not as easy as saying "I don't care if I fail." Let's assume that in this case you're not able to kick the desire to pass. At least not within the next 24 hours.

What else might you do? You could drink a bottle of mouthwash. Doing that would make studying irrational, since having drunk the mouthwash you should rather go to the hospital. (You care about passing the test, but much more about keeping your liver.) Then again, drinking the mouthwash would itself be irrational. So it wouldn't really accomplish what you sought in the first place: to avoid studying without thereby doing anything irrational. Nor would it allow you to slack off the rest of the night. By drinking the mouthwash you simply trade one unpleasant obligation (studying) for another (a trip to the emergency room).

If you had some way of forgetting about the test—for example, by taking a memory-erasing pill—then you could rationalize slacking off, since having taken the pill you would no longer have any reason to study. But like drinking the mouthwash, taking the pill would be irrational. For you now expect that it would result in your not studying and failing the test—an undesirable outcome.

What if you could take a different kind of pill, one that would cause you to know everything you need to know to pass the test? In that case you could rationalize slacking off, since having taken the pill you would no longer need to study. And unlike taking the memory-erasing pill, taking this pill would be rational, since you expect that it would result in your passing the test. So it seems that by taking the pill you can tailor the demands of rationality to your liking, without thereby doing anything irrational.

But taking a knowledge-inducing pill is a suspiciously easy way to rationalize slacking off. There is a simple diagnosis of why that is. The knowledge induced by the pill—consisting of facts about the subject matter of the test—does not play an essential role in the explanation of why, after taking the pill, it is rational for you to slack off. After all, simply *believing* that you took the pill (whether or not you actually took it) is enough to rationalize slacking off.

Let's assume that there is nothing you can do such that simply believing that you did it would rationalize slacking off. In particular, you don't have access to any knowledge-inducing pills. Are you therefore stuck with the obligation to study, or is there something else you could do to bypass that obligation?

A natural thought is that you might investigate whether studying really will help you pass the test. For if the investigation led you to sufficiently doubt the connection between studying and passing, then slacking off would become rationally permissible. If, on the other hand, the investigation reinforced the connection, then you would once again find yourself obligated to study.

What kind of investigation might you undertake? Suppose that many students just like you have faced decisions just like yours in the past. There is a trustworthy record book documenting, for each student, whether the student studied and whether the student passed. The book aggregates the data, giving overall pass rates for studying and not studying. Before reading the book, you believe that the pass rate for studying is high while the pass rate for not studying is low. So you expect the book to reflect these estimates. But it could be that the book says that the overall pass rates for studying and not studying are equal. If it does, then after reading the book it would be rational for you to slack off.

Before saying anything more about the record book investigation, let me make two clarifications. First, it's no surprise that we are sometimes able to rationalize a choice without seeing the rationalization coming. Suppose that you have the option of buying a bet that pays if you develop lung cancer before age 70. At present you believe that you don't have lung cancer, that lung cancer doesn't run in your family, and that you've never smoked and never will. So you shouldn't buy the bet. But now imagine that, before deciding whether to buy the bet, you have the option of getting a CAT scan. Since you believe that you don't have cancer, you expect that the scan will reveal nothing ominous. But, to your dismay, the scan reveals a malignant tumor. Now you have done something—getting the CAT scan—that made it the case that you should buy the bet. So in a sense you have rationalized buying the bet. But the credit should really go—not to you—but to the world, for setting you straight on your unfortunate condition. Although it was your investigation that ended up

rationalizing the choice, you played no part in orchestrating the rationalization.

Second, it's no surprise that we are sometimes able to rationalize a choice without being 100% certain, in advance, that the rationalization is coming. Suppose that you must pick one of two planes to board: plane A or plane B. Each plane has 50 passengers, and you know that of the 100 total passengers exactly one is carrying explosives. You have no idea which plane the bomber is on. But you have the opportunity to perform an experiment before making your choice: first, you choose a plane to investigate; then, one passenger from the chosen plane is selected at random and scanned for explosives; then, you're shown the results of the scan. Suppose that you wisely decide to perform the experiment. Before seeing the results of the experiment, you're 99% confident that the scan will show nothing unusual on the selected passenger. And you know that if that happens, you will have a very slight reason to prefer boarding the plane from which the subject of the experiment was selected. So you're 99% confident that investigating plane A will rationalize boarding plane A. And you're 99% confident that investigating plane B will rationalize boarding plane B. As it happens, you choose to investigate plane A and the scan shows nothing unusual, as expected. So by investigating plane A, you have rationalized boarding plane A, and you have done so in a way that you expected, with 99% confidence, to rationalize boarding plane A. Still, you don't deserve all the credit for the rationalization. You needed a little help from the world. For if the test had come up positive for explosives, then you would have had extremely strong reason to board plane B and your attempt at rationalizing boarding plane A would have failed.

Returning to your decision of whether to study, it's now clear why reading the record book is not a satisfying way of rationalizing slacking off. Although reading the book could rationalize slacking off, you're not confident, let alone certain, that it will. (After all, if you were confident that the record book contained evidence against the connection between studying and passing, then it wouldn't be rational for you to study in the first place.) What you seek when you seek to rationalize slacking off is not just to do something that might rationalize slacking off, but rather to do something that you can foresee with certainty will rationalize slacking off. You want to guarantee the rationalization of your preferred choice.

But so far we have seen no reason to think that you can. Indeed, just thinking abstractly about the matter, the kind of guaranteed rationalization described above is hard to countenance. If rationality was such that one could manipulate its demands in foreseeable ways, then it would seem too easy to be rational. Rationality—the subjective, practical flavor of rationality under discussion—is supposed to be within our ken, in some important sense, but it is not supposed to be completely under our control. Just as the correct rule of rational action cannot be "do whatever is objectively best," so too it cannot be "do whatever you want." A good theory of practical rationality should strike a balance between these two extremes. But a theory that permitted rationalization would seem to tread too closely to the "do whatever you want" extreme. And in so doing, such a theory would seem to be stripped of any

normative force. How could we feel pressure to follow the demands of the theory, when *we* decided what those demands are? Perhaps it is hard to find examples of rationalization because rationalization is impossible.

2

If rationalization is impossible, then a recently defended theory of practical rationality, known as *evidential decision theory*, stands refuted.[1] For this theory permits rationalization. In §4 I will describe an example in which evidential decision theory permits rationalization.[2] In this section and the next, I will introduce the theory, its main rival, and the model in which both are formulated.

The main element of the model is a quadruple $\langle \mathcal{A}, \mathcal{S}, p, u \rangle$, called a *decision problem*, in which $\mathcal{A}$ is a finite partition of propositions $a_1, \ldots, a_m$ representing the *actions* available to a particular agent at a particular time, $\mathcal{S}$ is a finite partition of propositions $s_1, \ldots, s_n$ representing the possible *states* of the world upon which the consequences of the actions depend, p is a probability function representing the agent's *degrees of belief* at the specified time, and u is a utility function representing the agent's *non-instrumental desires* at the time. The actions $a \in \mathcal{A}$ and states $s \in \mathcal{S}$ are chosen so that each conjunction as picks out a unique *consequence*, fully specifying how things stand with respect to everything that matters to the agent. Such consequences form the domain of the agent's utility function.

Various decision theories can be formulated within this model. Given a decision problem $\langle \mathcal{A}, \mathcal{S}, p, u \rangle$, the *simple expected utility* of an action $a_i \in \mathcal{A}$ is identified with a weighted sum, for each state, of the utility of taking the action in a world in that state, weighted by the probability of the state obtaining:

$$SEU(a_i) = \sum_{j=1}^{n} p(s_j) u(a_i s_j).$$

Simple decision theory (SDT) enjoins agents to choose among the $a \in \mathcal{A}$ so as to maximize SEU.[3]

The inadequacy of SDT is well known and easy to illustrate. Suppose that Al is deciding whether to smoke. He believes that smoking has a strong tendency to cause cancer and, while he finds some pleasure in smoking, that pleasure is

[1] The theory originated with Jeffrey (1965) and has been most recently and extensively defended by Ahmed (2014).

[2] To forestall potential confusion about the structure of my argument, let me clarify some phrases used in this paragraph. By "rationalization is impossible," I mean that it is never rational to manipulate the demands of rationality in the way elucidated in §1. By "evidential decision theory permits rationalization," I mean that it is sometimes rational, according to evidential decision theory, to manipulate the demands of rationality in such a way.

[3] This theory is sometimes associated with Savage (1954). However, Savage intended his states to form a privileged partition—essentially a partition of dependency hypothesis, a la Lewis (1981). So Savage's theory is best understood as an early version of causal decision theory, for which the smoking problem does not arise. Thanks to [removed for blind review] for clarification on this point.

dwarfed by the displeasure he associates with cancer. Given what he believes and desires, Al should not smoke.

But SDT advises that Al smoke. Let $\mathcal{S}$ include the proposition that Al gets cancer and its negation. Let $-\gamma$ be the large amount of disutility associated with cancer, let $\delta > 0$ be the small amount of utility associated with smoking, and set an arbitrary zero where both cancer and smoking are absent.

	$cancer$	$\neg cancer$
$smoke$	$-\gamma + \delta$	δ
$\neg smoke$	$-\gamma$	0

Table 1: the smoking problem.

If x is Al's degree of belief in the proposition that he will get cancer, then his simple expected utilities are related by the following equation:

$$\begin{aligned} SEU(smoke) &= x(-\gamma + \delta) + (1 - x)\delta \\ &= -x\gamma + \delta > -x\gamma = SEU(\neg smoke). \end{aligned}$$

Hence, SDT enjoins Al to smoke, in spite of his belief that smoking causes cancer and his strong desire to avoid cancer.

To handle this kind of problem, Jeffrey (1965) proposed a different definition of expected utility. Whereas simple expected utility weights utilities by unconditional probabilities in states, Jeffrey's definition (EEU) weights utilities by conditional probabilities in states, conditional on actions:

$$EEU(a_i) = \sum_{j=1}^{n} p(s_j|a_i)u(a_i s_j).$$

Applying Jeffrey's definition to the smoking problem:

$$\begin{aligned} EEU(smoke) &= p(cancer|smoke)(-\gamma + \delta) + (1 - p(cancer|smoke))\delta \\ &= -p(cancer|smoke)\gamma + \delta. \\ EEU(\neg smoke) &= -p(cancer|\neg smoke)\gamma. \end{aligned}$$

Therefore, not smoking maximizes EEU iff the difference

$$p(cancer|smoke) - p(cancer|\neg smoke)$$

exceeds the fraction δ/γ. In other words, not smoking maximizes EEU iff Al regards smoking as sufficiently strong evidence of cancer—a condition plausibly satisfied by the description of the smoking problem. (The required strength of the evidential connection lowers as γ increases and δ decreases.)

Since Jeffrey's definition uses conditional probabilities of states on actions, and since the differences between these conditional probabilities measure the extent to which the agent regards actions as evidence of states, Jeffrey's brand

of expected utility has come to be known as *evidential expected utility*, and the theory enjoining agents to maximize EEU has come to be known as *evidential decision theory* (EDT).

Although EDT gives rational advice in the smoking problem, some believe that it does so only incidentally. For these theorists, Al is irrational to smoke because he believes that smoking *causes* cancer, not because he regards smoking as evidence of cancer. The shift in emphasis makes no difference in the smoking problem because smoking is regarded both as a cause and as evidence of cancer. But it makes a difference elsewhere, such as in the notorious Newcomb problem:[4]

> *Newcomb*: There is a transparent box containing \$1,000 and an opaque box containing either \$1,000,000 (*full*) or nothing (*empty*). Ted has two options: he can take just the opaque box (*onebox*) or he can take both boxes (*twobox*). The content of the opaque box was determined yesterday by a reliable predictor. The opaque box contains \$1,000,000 iff the predictor predicted that Ted would take just the opaque box.

	$full$	$empty$
$twobox$	1,001,000	1,000
$onebox$	1,000	0

Table 2: *Newcomb*.

Supposing for simplicity that Ted's utilities are linear and increasing in dollars, the evidential expected utilities of his options in *Newcomb* are:

$$\begin{aligned} EEU(twobox) &= p(full|twobox)(1,001,000) + (1 - p(full|twobox))(1,000) \\ &= p(full|twobox)(1,000,000) + 1,000. \\ EEU(onebox) &= p(full|onebox)(1,000,000). \end{aligned}$$

Therefore, one-boxing maximizes EEU iff the difference

$$p(full|onebox) - p(full|twobox)$$

exceeds the fraction $1/1,000$. In other words, EDT advises one-boxing so long as Ted regards one-boxing as at least a little evidence that the opaque box contains \$1,000,000—a condition satisfied by the description of *Newcomb*.

Those who reject EDT's diagnosis of the smoking problem also find fault in its treatment of *Newcomb*. They maintain that Ted should take both boxes, since he knows that his actions have no causal effect on the content of the opaque box, and since—no matter what the opaque box contains—two-boxing nets \$1,000 more than one-boxing.[5]

Many of those who reject EDT accept an alternative decision theory. As formulated by Lewis (1981), the alternative theory is essentially a return to

[4] Attributed to physicist William Newcomb, *Newcomb* was popularized by Nozick (1969).

[5] See Spencer and Wells (Forthcoming) for a more detailed defense of two-boxing.

SDT, with one caveat. Whereas in SDT the set of states can be any partition of logical space, Lewis requires that the states be *(causal) dependency hypotheses*. A dependency hypothesis, for an agent at a time, is a proposition fully specifying how things the agent cares about do or do not depend causally on the agent's present actions.[6] It is a hypothesis about the causal structure of the world, as it pertains to the decision. Lewis proves that, necessarily, the dependency hypotheses for an agent are causally independent of the actions between which the agent is deciding. So, for example, the proposition that the opaque box contains $1,000,000 is a dependency hypothesis for Ted in *Newcomb*, whereas getting cancer is not a dependency hypothesis for Al in the smoking problem.

Replacing states with a partition of dependency hypotheses $\mathcal{C} = \{c_1, \ldots, c_n\}$, we can characterize a third kind of expected utility:

$$CEU(a_i) = \sum_{j=1}^{n} p(c_j)u(a_i c_j).$$

Since the concept of a dependency hypothesis is causal by definition, this kind of expected utility has come to be known as *causal expected utility*, and the theory enjoining agents to maximize CEU has come to be known as *causal decision theory* (CDT).

CDT directly opposes EDT in *Newcomb*. If x is Ted's degree of belief in the proposition that the opaque box contains $1,000,000, then his causal expected utilities are related by the following equation:

$$\begin{aligned} CEU(twobox) &= x(1,001,000) + (1-x)1,000 \\ &= 1,000,000x + 1,000 > 1,000,000x = CEU(onebox). \end{aligned}$$

Hence, CDT enjoins Ted to take both boxes.

3

Each of the decision problems considered so far is non-sequential, in the sense that there is just one set of options and the choice between the members of that set occurs at a single time. A sequential decision problem, on the other hand, has at least two sets of options corresponding to two different times at which a choice must be made.

Sequential decisions are common. Often when we make a decision it is just one move in a chain of subsequent decisions. One particularly common kind of sequential decision problem involves evidence gathering. Often we decide to ask a question, make an observation, look up something on the internet or perform

[6] In order to account for decision problems in which the objective chance of an action yielding a particular consequence is neither 0 nor 1, we would need to alter the framework slightly, removing the stipulation that each ac entails a unique consequence and requiring that the c specify objective conditional chances of consequences on actions. However, the decision problems discussed in this paper require no such alteration, so we will work with the simpler albeit less general framework sketched above.

an experiment before proceeding further with our lives. We saw an example of this in §1, with the decision of whether to read the pass-fail data before deciding whether to study.

The problem that I will present in the next section—to illustrate the possibility of rationalization under EDT—is a sequential problem involving evidence gathering. It will take a little work to see exactly how to apply EDT and CDT to this kind of problem. The purpose of this section is to extend the framework of §2 so that we can more easily apply the theories presented in that section to the problem of the next section.

Start with a simple two-stage sequential decision problem where the options include gathering more information before acting.

> *Simple Problem.* There are two opaque boxes before you: A and B. One contains $100; the other is empty. You're 75% confident that A contains $100. You may take either box, but not both. Alternatively, you may look inside A and then take a box.

	$fullA$	$fullB$
A	100	0
B	0	100

Table 3: *Simple Problem.*

In the *Simple Problem* your first set of options includes looking in A, taking A straightaway, or taking B straightaway. We already know how to calculate the expected utilities of the latter two options and it is clear that the expected utility of taking A straightaway (75) exceeds the expected utility of taking B straightaway (25). The question is whether the expected utility of taking A straightaway also exceeds that of looking in A before choosing a box.

Let us assume that, if you look in A, you will act rationally thereafter, in the sense that you will update your degrees of belief by conditionalizing on the truth about what is in the box, and that after updating you will choose the option that maximizes expected utility relative to your updated degrees of belief. So, for example, if you see that A contains $100, your new expected utility for taking A will be 100, and you will take A. And if you see that A contains nothing, your new expected utility for taking B will be 100, and you will take B. So in either case, you will choose an action with expected utility 100, and you can be certain of this. So the expected utility of looking in A is itself 100, i.e. greater than that of taking A straightaway. Hence, the uniquely rational choice is to look in A before choosing a box.

To generalize the informal reasoning above, we take as given a partition of propositions $\mathcal{E} = \{e_1, \ldots, e_m\}$ representing the possible pieces of evidence that you might learn by making a particular observation. We then define the expected utility of using a particular piece of evidence $e_k \in \mathcal{E}$ to inform your decision (call this act use_{e_k}) as the expected utility of the action that maximizes expected utility relative to your updated-on-e_k degrees of belief. This definition

yields causalist and evidentialist formulae:

$$CEU(use_{e_k}) = \mathtt{max}_i \sum_{j=1}^{n} p(c_j|e_k)u(a_i c_j). \tag{1}$$

$$EEU(use_{e_k}) = \mathtt{max}_i \sum_{j=1}^{n} p(s_j|e_k a_i)u(a_i s_j). \tag{2}$$

Next we define the expected utility of gathering and using the evidence gathered (call this action *look*) as a weighted sum, for each possible piece of evidence, of the expected utility of using that piece of evidence, weighted by the probability that the piece of evidence is true (i.e. that it will be the piece of evidence gathered). This definition also yields causalist and evidentialist formulae:

$$CEU(look) = \sum_{k=1}^{m} p(e_k)CEU(use_{e_k}). \tag{3}$$

$$EEU(look) = \sum_{k=1}^{m} p(e_k|look)EEU(use_{e_k}). \tag{4}$$

Applying these formulae to the *Simple Problem*, it is straightforward to confirm that looking in A maximizes both CEU and EEU.

There is an alternative version of *Newcomb* that has garnered some attention in the literature.[7] This alternative version may seem to supply a case in which EDT permits rationalization. In fact, it does not. But it is instructive to see why it does not. Here is the problem:

> *Viewcomb*: Everything is the same as in *Newcomb*, only now Ted has the option of looking inside the opaque box before making his decision.

According to EDT, Ted should one-box straightaway. For suppose that Ted looks in the box and sees that it is full. Then the act that maximizes EEU relative to his updated degrees of belief is two-boxing, and its EEU is 1,001,000. Suppose on the other hand that Ted sees that the box is empty. Then the act that maximizes EEU relative to his updated degrees of belief is again two-boxing, although its EEU in this case is 1,000. Hence, by (2),

$$EEU(use_{e_{full}}) = 1,001,000; \text{ and,}$$
$$EEU(use_{e_{empty}}) = 1,000.$$

Plugging these values into (4), we have:

$$\begin{aligned} EEU(look) &= p(full|look)EEU(use_{e_{full}}) + p(empty|look)EEU(use_{e_{empty}}) \\ &= p(full|look)(1,001,000) + p(empty|look)(1,000) \end{aligned}$$

[7] See, for example, Gibbard and Harper (1978), Adams and Rosenkrantz (1980), Skyrms (1990), Arntzenius (2008), Meacham (2010), Ahmed (2014), Hedden (2015) and Wells (Forthcoming).

Notice that, for Ted, *look* entails two-boxing, since he is certain that he will two-box no matter what he learns by looking. Plausibly, then, $p(full|look) = p(full|twobox)$ and $p(empty|look) = 1 - p(full|twobox)$. Supposing for concreteness that the predictor is believed to be 60% reliable, we have:

$$\begin{aligned} EEU(look) &= (.4)(1,001,000) + (.6)(1,000) \\ &= 401,000. \end{aligned}$$

Note also that

$$\texttt{max}_i\ EEU(a_i) = EEU(onebox) = (.6)(1,000,000) = 600,000.$$

Hence,

$$EEU(look) = 401,000 < 600,000 = \texttt{max}_i\ EEU(a_i).$$

Hence, EDT recommends that Ted one-box straightaway.

Note that in *Viewcomb* Ted can change EDT's recommendation as he pleases. Although at the outset EDT recommends that Ted one-box, it is within Ted's power to look in the opaque box, and he can be certain, in advance, that if he looks in the box, EDT will thereafter recommend that he two-box. We thus *seem* to have a case in which EDT countenances a rationalization of the kind discussed in §1.

But we do not. The reason is that EDT does not permit looking in the box. From an evidentialist perspective, looking in the box (to avoid one-boxing) is just like drinking the mouthwash or taking the memory-erasing pill (to avoid studying). In each case, one is able to change the demands of rationality in a foreseeable way, but only by first doing something irrational. There is nothing odd about such irrational rationalizations. The odd rationalization is that which is itself rational. Our question is whether there is a theory of rationality that *permits* rationalization.

4

There is.[8] Consider:

[8] *The Switch Problem* is a modification of a problem called *Newcomb Coin Toss*, presented recently in Wells (Forthcoming). In both problems, the probabilistic relations are such that if the agent gathers a certain piece of evidence then, no matter what she learns, evidential decision theory will require her to make a decision that it does not antecedently require her to make. Moreover, in both problems, the agent is in a position to know this in advance of gathering the evidence. Now, there is a minor difference between *The Switch Problem* and *Newcomb Coin Toss*: in *The Switch Problem*, EDT permits the gathering of evidence, whereas in *Newcomb Coin Toss*, it does not. For this reason, *Newcomb Coin Toss* showcases EDT's violation of Good's Theorem (see note 13), while *The Switch Problem* showcases no such violation. However, this difference can be easily erased by increasing the cost of not observing the light in *Newcomb Coin Toss*. Thanks to an anonymous reviewer for clarification on this point. Nevertheless, there is a major difference between the two cases, and it can be stated rather precisely. Let us say that a probability function P instantiates *Simpson's paradox* just if there are propositions X, Y and Z such that:

The Switch Problem: There are two opaque boxes, A and B. One contains \$100. The other is empty. Sue may take A ($TakeA$) or B ($TakeB$) but not both. Additionally, there are two colored switches, one red and one green, blocked from Sue's view. Each switch is either on or off. Before choosing a box, Sue may look at the red switch ($LookR$) or the green switch ($LookG$) but not both. The statuses of the switches and the contents of the boxes were determined in advance, by a predictor, in the following way.

If the predictor predicted that Sue would take A ($PredA$), she tossed a fair coin, put \$100 in A (InA) if it landed heads (H), and tossed another coin. If the second coin landed heads, she flipped both switches on (RG). If it landed tails, she flipped just the green switch on ($\neg RG$). Alternatively, if the first coin landed tails (T), she put \$100 in B (InB), and tossed another coin. If the second coin landed heads, she flipped just the green switch on. If it landed tails, she flipped both switches off ($\neg R\neg G$).

If the predictor predicted that Sue would take B ($PredB$), she tossed a coin, put \$100 in B if it landed heads, and tossed another coin. If the second coin landed heads, she flipped both switches on. If it landed tails, she flipped just the red switch on ($R\neg G$). Alternatively, if the first coin landed tails, the predictor put \$100 in A, and tossed a second coin. If it landed heads, she flipped just the red switch on. If it landed tails, she flipped both switches off.

Sue is fully aware of the foregoing details, which are summarized in table 4 and figure 1.

	$PredA$				$PredB$			
	InA		InB				InA	
	RG	$\neg RG$	$\neg RG$	$\neg R\neg G$	RG	$R\neg G$	$R\neg G$	$\neg R\neg G$
$TakeA$	100	100	0	0	0	0	100	100
$TakeB$	0	0	100	100	100	100	0	0

$$P(X \mid YZ) > P(X \mid \neg YZ),$$
$$P(X \mid Y\neg Z) > P(X \mid \neg Y\neg Z), \text{ yet}$$
$$P(X \mid Y) \leq P(X \mid \neg Y).$$

In *The Switch Problem*, for fixed X and Y, there is a Z satisfying the above inequalities, and also a Z' satisfying their reversal. *The Switch Problem* thus contains two instances of Simpson's paradox. *Newcomb Coin Toss*, like the original *Newcomb* problem, contains only one. This difference is significant. Whereas an agent facing *Newcomb Coin Toss* can gather evidence so as to ensure that EDT will give *one particular* piece of advice (i.e. she can look at the light so as to ensure that EDT will advise buying the box), an agent facing *The Switch Problem* can gather evidence so as to ensure that EDT will give *either of two contradictory* pieces of advice. This seems to me to aggravate the case against EDT considerably, as discussed in §5.

Table 4: *The Switch Problem.*

Here is how to interpret table 4. For each cell containing a number, the number in the cell represents the payoff of choosing the option that is directly left of the cell, at a world in which each proposition directly above the cell is true. So, for example, the number 100 in the top-left cell represents the payoff of taking box A at a world in which each proposition directly above the cell is true, i.e. a world in which the predictor predicted that A would be taken, put $100 in A, and flipped both switches on.

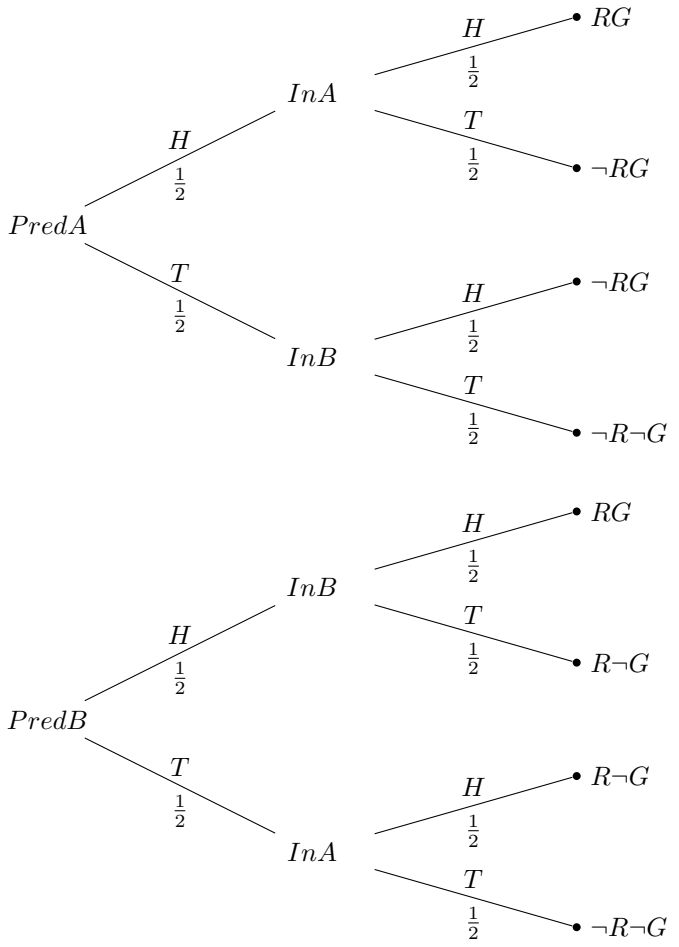

Figure 1: Probability trees representing the probabilistic relations in *The Switch Problem.*

In *The Switch Problem*, EDT permits Sue to manipulate rationality in foreseeably contradictory ways. By this I mean the following:

Claim 1. Sue is certain that if she looks at the red switch, EDT will require that she take box A.
Claim 2. Sue is certain that if she looks at the green switch, EDT will require that she take box B.
Claim 3. EDT permits Sue to look at either switch.

To prove these claims I will make two simplifying assumptions, each of which may be relaxed without loss. First, I will assume that Sue believes the predictor to be 100% reliable.[9] Second, I will assume that Sue's utilities are linear and increasing in dollars. I will also carry over the 'transparency' assumption from the discussion of evidence gathering in §3. That is, I will assume that Sue is certain that if she looks at a switch, she will act rationally thereafter, in the sense that she will update her degrees of belief by conditionalizing on the truth about the switch, and that after updating she will choose the option that maximizes expected utility relative to her updated degrees of belief.[10] Let p be Sue's probability function before she decides whether to look at a switch. For any proposition e, let p_e be Sue's credence function after conditionalizing on e: $p_e(.) = p(.|e)$.

To prove **Claim 1**, it suffices to prove the embedded conditional from premises of which Sue is certain. Suppose that Sue looks at the red switch. Then, either she learns R or she learns $\neg R$.

Suppose that she learns R. Then her new probability function is p_R. Note that there is only one possibility in which the red switch is on and the predictor predicted that Sue would take box A; and, in that possibility, \$100 is in box A. Moreover, conditional on her taking box A, Sue is certain that the predictor predicted that she take A. Hence, $p_R(InA|TakeA) = 1$. Note also that there are three equiprobable possibilities in which the red switch is on and the predictor predicted that Sue would take box B; and, in two of them, \$100 is in box B. Hence, $p_R(InB|TakeB) = 2/3$. Hence, after learning R, Sue's EEU of taking box A is 100 while her EEU of taking box B is $100(2/3)$. Hence, EDT requires that Sue take box A.

Suppose, on the other hand, that Sue learns $\neg R$. Then her new credence function is $p_{\neg R}$. Note that there are three equiprobable possibilities in which the red switch is off and the predictor predicted that Sue would take box A; and, in one of them, \$100 is in A. Hence, $p_{\neg R}(InA|TakeA) = 1/3$. Note also that there is only one possibility in which the red switch is off and the predictor predicted that Sue would take box B; and, in that possibility, \$100 is in A. Hence, $p_{\neg R}(InB|TakeB) = 0$. Hence, after learning $\neg R$, Sue's EEU of taking box A is $100(1/3)$ while her EEU of taking box B is 0. Hence, EDT requires that Sue take box A.

[9]This assumption may lead to Sue assigning zero probability to some of her available actions, in which case the evidential expected utilities of those actions would be undefined. To avoid this, we may assume instead that the probability that the predictor is mistaken is non-zero but negligible. Thanks to [removed for blind review] for clarification on this point.

[10]This assumption may also lead to Sue assigning zero probability to some of her available actions. As before, we can avoid this by assuming instead that irrational actions get negligible positive probability.

Hence, if Sue looks at the red switch, then, no matter what she learns, EDT will require that she take box A. The only premise is that Sue will conditionalize on the truth about whether the switch is on. By the transparency assumption, Sue is certain of this. Hence, Sue is certain of the conclusion. **Claim 1** follows.

To prove **Claim 2**, Suppose that Sue looks at the green switch. Then, either she learns G or she learns $\neg G$. Suppose that she learns G. Then her new probability function is p_G. Note that there is only one possibility in which the green switch is on and the predictor predicted that Sue would take box B; and, in that possibility, \$100 is in box B. Moreover, conditional on her taking box B, Sue is certain that the predictor predicted that she take B. Hence, $p_G(InB|TakeB) = 1$. Note also that there are three equiprobable possibilities in which the green switch is on and the predictor predicted that Sue would take box A; and, in two of them, \$100 is in box A. Hence, $p_R(InA|TakeA) = 2/3$. Hence, after learning G, Sue's EEU of taking box B is 100 while her EEU of taking box A is 100(2/3). Hence, EDT requires that Sue take box B.

Suppose, on the other hand, that Sue learns $\neg G$. Then her new probability function is $p_{\neg G}$. Note that there are three equiprobable possibilities in which the green switch is off and the predictor predicted that Sue would take box B; and, in one of them, \$100 is in B. Hence, $p_{\neg G}(InB|TakeB) = 1/3$. Note also that there is only one possibility in which the green switch is off and the predictor predicted that Sue would take box A; and, in that possibility, \$100 is in B. Hence, $p_{\neg G}(InA|TakeA) = 0$. Hence, after learning $\neg G$, Sue's EEU of taking box B is 100(1/3) while her EEU of taking box A is 0. Hence, EDT requires that Sue take box B.

Hence, if Sue looks at the green switch, then, no matter what she learns, EDT will require that she take box B. **Claim 2** follows by the transparency assumption.

To prove **Claim 3**, first consider the option of looking at the red switch. The associated evidence partition is $\{R, \neg R\}$. We saw above that the EEU of the option that maximizes EEU relative to p_R (namely, $TakeA$) is 100, and the EEU of the option that maximizes EEU relative to $p_{\neg R}$ (namely, $TakeA$) is 100(1/3). Moreover, **Claim 1**, together with the transparency assumption, entails that $p(TakeA|LookR) = 1$. Since $p(PredictA|TakeA) = 1$ and $p(R|PredictA) = 1/4$, it follows that $p(R|LookR) = 1/4$. Hence, by (4),

$$EEU(LookR) = 100(1/4) + 100(1/3)(3/4) = 50.$$

Next consider the option of looking at the green switch. The associated evidence partition is $\{G, \neg G\}$. We saw above that the EEU of the option that maximizes EEU relative to p_G (namely, $TakeB$) is 100, and the EEU of the option that maximizes EEU relative to $p_{\neg G}$ (namely, $TakeB$) is 100(1/3). Moreover, **Claim 2**, together with the transparency assumption, entails that $p(TakeB|LookG) = 1$. Since $p(PredictB|TakeB) = 1$ and $p(G|PredictB) = 1/4$, it follows that $p(G|LookG) = 1/4$. Hence, the EEU of looking at the green switch is also 50.

It is simple to confirm that $EEU(TakeA) = EEU(TakeB) = 50$ as well. After all, $p(InA|TakeA) = p(InB|TakeB) = 1/2$. Hence, before looking at a

switch, the EEUs of each of Sue's four options are equal. Hence, EDT initially permits Sue to take any option, including either evidence gathering option. **Claim 3** follows. Hence, EDT permits rationalization in *The Switch Problem*.

CDT handles *The Switch Problem* much more sanely. Like EDT, CDT initially permits each of Sue's four options. The question is whether the demands of CDT change after Sue looks at a switch. Whether they do depends on Sue's beliefs about which box she will ultimately decide to take. We have not yet said anything about those beliefs, so, from a causalist perspective, our problem is not yet fully specified. Let us fully specify it. Let us stipulate that Sue is maximally unsure about what she will do, so that $p(TakeA) = p(TakeB) = 1/2$. We can now show that the CEU of taking box A will always equal the CEU of taking box B, no matter what Sue learns after looking at a switch.

Suppose that Sue looks at the red switch and learns R. Then her updated causal expected utilities are as follows:

$$CEU_R(TakeA) = p_R(InA)(100).$$
$$CEU_R(TakeB) = p_R(InB)(100).$$

Since Sue's degrees of belief are split evenly over her options, they are also split evenly over the two possible predictions. Hence, the four possibilities in which the red switch is on are equiprobable. Half of those possibilities are ones in which box A contains \$100, and the other half are ones in which box B contains \$100. Hence, $p_R(InA) = p_R(InB) = 1/2$. Hence,

$$CEU_R(TakeA) = 50 = CEU_R(TakeB).$$

By the symmetry of the problem, parallel reasoning shows that

$$CEU_{\neg R}(TakeA) = 50 = CEU_{\neg R}(TakeB),$$

as well. Indeed, the same holds true for the causal expected utilities of Sue's options after looking at the green switch.

What happens if we relax the assumption that Sue is maximally unsure of what she will do? Then CDT's recommendations may change. For instance, if Sue is antecedently *certain* that she will take box A, and she sees that the red switch is on, then CDT requires that she take box A.[11] But if, on the other hand, Sue sees that the red switch is off, CDT requires that she take box B.[12] So,

[11] *Proof:* Suppose that Sue learns R. By hypothesis, $p(TakeA) = 1$. Hence, in this case, the causal expected utility of taking box A equals its evidential expected utility, which, we have seen, is 100. The causal expected utility of taking box B is $p_R(InB)(100)$, or equivalently $p_R(InB|TakeA)$. We have seen that $p_R(InA|TakeA) = 1$. Hence, $p_R(InB|TakeA) = 0$. Hence, in this case, after learning R, Sue's causal expected utility of taking box A exceeds that of taking box B by 100.

[12] *Proof:* Suppose that Sue learns $\neg R$. By hypothesis, $p(TakeA) = 1$. Hence, in this case, the causal expected utility of taking box A equals its evidential expected utility, which, we have seen, is 100/3. The causal expected utility of taking box B is $p_{\neg R}(InB)(100)$, or equivalently $p_{\neg R}(InB|TakeA)(100)$. We have seen that $p_{\neg R}(InA|TakeA) = 1/3$. Hence, $p_{\neg R}(InB|TakeA) = 2/3$. Hence, in this case, after learning $\neg R$, Sue's causal expected utility of taking box B exceeds that of taking box A, 200/3 to 100/3.

although CDT sometimes changes its requirements, Sue cannot be antecedently certain about the way in which the theory will change its recommendations, since she cannot be antecedently certain that—for example—the red switch is on. As we saw in §1, there is nothing odd about rationalizations that aren't anticipated with certainty. So there is nothing odd about CDT's treatment of *The Switch Problem*.

5

The Switch Problem is similar to *Viewcomb* in that the agent can predict with certainty that if she makes a particular observation, the demands of evidential rationality will change in a particular way. In *Viewcomb*, the agent can predict with certainty that if she looks in the opaque box, EDT will no longer permit one-boxing. In *The Switch Problem*, the agent can predict with certainty that if she looks at the red switch, EDT will no longer permit taking box B (and that if she looks at the green switch, EDT will no longer permit taking box A).

However, in *Viewcomb*, EDT does *not* permit gathering evidence in such a way as to predictably manipulate the demands of rationality.[13] This fact may be seen as a sort of saving face for the theory. Although it is physically possible for an agent to manipulate EDT's demands in *Viewcomb* (by looking in the box), it is rationally impermissible, according to the theory. If we personify EDT as an advisor, it is as if the advisor is saying: "You should one-box. Of course, if you look in the box, then, no matter what you see, I will tell you to two-box. But I don't want to tell you that, since I think you should one-box. So you shouldn't look in the box." This advice may seem odd but it is at least self-reinforcing. The advisor acknowledges that her advice will change from what it is at the outset, but she does not support that change.

The Switch Problem is different. There, EDT *permits* gathering evidence in such a way as to predictably manipulate the demands of rationality. It is as if the EDT-advisor is saying: "You are permitted to take box B. Of course, if you look at the red switch, then, no matter what you see, I will tell you that you are *not* permitted to take box B—that you should rather take box A. But I have no problem telling you that. So go ahead, look at the red switch. Or don't. I'll tell you whatever you want to hear." Not only is this advice odd, it rings of self-doubt. Not only does the advisor acknowledge that her advice will change, she apparently endorses that change at the outset.

To make things vivid, imagine that Sue, who only speaks English, has a

[13]There is a famous theorem due to I. J. Good (1967) according to which it is always rational to gather more evidence before making a decision, provided that the cost of so doing is negligible. *Viewcomb* is a counterexample to the version of Good's theorem wherein 'rational' is given an evidential interpretation. EDT's violation of Good's theorem is often used as an argument against the theory, but Maher (1990) has shown that CDT also violates Good's theorem on occasion. Hence, violations of Good's theorem do not, on their own, cut any ice in the debate between EDT and CDT. Nevertheless, this paper suggests that there is a problem for EDT surrounding its treatment of evidence gathering that arises not when the theory *prohibits* the collection of cost-free evidence, but rather when it *permits* such collection.

Chinese-speaking duplicate, Lu. Suppose that Sue and Lu face *The Switch Problem* together, and they have shared interests: they will split whatever earnings they make from whichever box they collectively decide to take. Moreover, they are both convinced by evidential reasoning when it comes to making choices. At the outset Sue and Lu are indifferent as to which box to take, and they are also indifferent as to which switch to look at. As it happens, Sue looks at the red switch and Lu looks at the green switch. After looking at their respective switches Sue and Lu reconvene to decide which box to take. Sue now knows the status of the red switch (say, that it's on) and Lu knows the status of the green switch (say, that it's off), but neither knows the statuses of both switches, and they are unable to communicate with one another. Sue reaches for box A but just as she is about to take it, Lu grabs her arm and gestures for them to take box B. Sue shakes her head and points at box A. They start fighting.

But should they fight? Sue and Lu both want the same thing. And they both agree on how their beliefs and desires should combine to guide their decision. Of course, they have different beliefs: Sue believes that the red switch is on but has no idea whether the green switch is on, whereas Lu believes that the green switch is off but has no idea whether the red switch is on. There is nothing odd about two people with different beliefs rationally disagreeing about what to do, even when their interests align. What is odd is that Sue knew, in advance, that after looking at the red switch she would prefer that they take box A, and that after Lu looked at the green switch, Lu would prefer that they take box B. And Lu knew this as well. So they knew that they would fight, and they knew which sides they would be taking in the fight. In that case, why not start fighting right then and there, before looking? Either way, fighting over which box to take at any stage in the problem seems perverse. The location of the money was determined by the toss of a fair coin, so, intuitively, taking either box is permissible—regardless of what is known about the switches.

David Lewis (1981, p. 5) once criticized evidential decision theory for commending "an irrational policy of managing the news." To this we should add that the theory commends an irrational policy of managing its own requirements.

6

We began in §1 with a case in which one option—slacking off—was antecedently irrational, and we asked whether it was possible to 'rationalize' that option. We found that it was not. It is worth noting that the case with which we ended takes a slightly different form. In *The Switch Problem*, both box-taking options are antecedently permissible, yet it is possible to render one or the other option uniquely rational.

Two questions arise. First, is the type of rationalization permitted under EDT in *The Switch Problem* any less toxic than the type of rationalization identified in §1? In other words, does the fact that the initial expected utilities of the options in *The Switch Problem* are equal make EDT's treatment of the

case any less problematic? I see no reason to think that it does, though I think that this question is worth pursuing further.

Second, is it possible to design a case in which EDT permits the rationalization of an initially impermissible option? If it was, then the type of rationalization in such a case would seem exactly analogous to the studying problem with which we began.

To this end, consider a variation on *The Switch Problem* in which there is a small prize—say, one cent—for looking at a switch. The addition of the prize tips the initial expected utilities in such a way that taking a box straightaway is now impermissible. So we have designed a case in which EDT permits the rationalization of an initially impermissible option. Still, the case is such that the initial expected utilities of the two box-taking options are equal—a feature not shared by the studying case. I leave it as an open question whether the two cases can be made exactly the same.

I want to end by addressing two objections.[14] The first objection centers on the 'transparency' assumption of §4. For convenience let us specify a point in time before the agent facing *The Switch Problem* observes a switch and call it 'stage 1'; and let us call the point after the agent observes a switch 'stage 2'. In deriving the results of §4, I assumed that (a) the agent who follows the advice of EDT believes at stage 1 that she will follow the advice of EDT at stage 2, and (b) the agent who follows the advice of CDT believes at stage 1 that she will follow the advice of CDT at stage 2. But that means that the two agents under comparison have different beliefs and are thus associated with different probability functions. If we identify a "decision problem" in part with a probability function representing the beliefs of the agent facing the problem, then it appears that there are really two decision problems here: the one faced by someone who believes they are an evidentialist, and the one faced by someone who believes they are a causalist.

Now, if we focus on either one of these problems individually, EDT and CDT give the same advice at stage 1. But that's no surprise. After all, EDT and CDT give the same advice at stage 1 no matter what the agent believes about what she will later do. Both theories permit taking any of the four options at stage 1. The theories only differ in their recommendations at stage 2, once the agent acquires evidence about a switch. The agent's belief at stage 1 about what she will do at stage 2 is only relevant insofar as it puts her in a position to know that the demands of rationality will change in a particular way upon viewing a switch. Thus, the relevant question is whether, once we focus on a single doxastic state—the state of believing that you will follow EDT at stage 2—both the evidentialist and the causalist are able to rationalize an option. If they were, then *The Switch Problem* would not cut any ice in the debate between the two theories. That is the objection.

My response is that that the conflict between EDT and CDT persists even after holding fixed the doxastic state of the agent. As shown in §4, EDT permits an agent who believes that she will follow EDT at stage two to manipulate

[14]Thanks to an anonymous referee for bringing these worries to my attention.

rationality in foreseeable ways by choosing a switch to observe. The question is whether CDT similarly permits an agent who believes that she will follow EDT to manipulate rationality. The answer is no. To see why, we must imagine an agent who is subject to the demands of rationality as determined by CDT but who nevertheless believes that she will follow the demands of EDT. Of course, this deluded causalist that we are imagining believes, at stage 1, that, at stage 2, after viewing a switch, the demands of rationality will change (since the demands of EDT would change). For example, she believes that if she views the green switch then, no matter what she learns, rationality will demand that she take box B. But rationality will not *in fact* demand that she take box B, since the demands of rationality at stage 2, for the causalist, are causalist demands, following directly from the agent's causal expected values, and, as shown in §4, the causal expected values of the agent's options after seeing either switch remain unchanged: no matter what the agent learns after observing a switch, the agent is permitted by CDT to take either box, just as she was at stage 1. So, while the deluded causalist believes that the demands of rationality will change, she cannot know that they will, since they won't. The manipulation never happens, so it cannot be foreseen to happen. Thus, CDT does not permit rationalization, even for an agent who believes that she will follow EDT. We therefore have a single decision problem in which EDT permits rationalization but CDT does not.

Let us, then, grant my claim that EDT permits rationalization in *The Switch Problem*. Still, I hear a second objection. It is that I simply have not proven that rationalization is impossible. Without such a proof, a defender of EDT may bite the bullet and maintain that rationalization—though perhaps impermissible in most ordinary cases such as those discussed in §1—is permissible in certain contrived cases such as *The Switch Problem*. I concede that I have offered no general proof that rationalization is impossible. Nevertheless, I believe that the preceding discussion is significant, since it reveals a new consequence of EDT that is not had by CDT. Although I have offered some intuitive reasons to worry about this consequence, I am ultimately happy leaving it to the reader to decide whether the consequence is a feature of the theory or a bug.

References

Adams, E. and Rosenkrantz, R. 1980. "Applying the Jeffrey decision model to rational betting and information acquisition." *Theory and Decision* 12:1–20.

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

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

Gibbard, A. and Harper, W. 1978. "Counterfactuals and Two Kinds of Expected

Utility." In Leach J. Hooker, A. and E. McClennen (eds.), *Foundations and Applications of Decision Theory*, 125–162. Reidel.

Good, I. J. 1967. "On the Principle of Total Evidence." *British Journal for the Philosophy of Science* 17:319–321.

Hedden, B. 2015. *Reasons without Persons*. Cambridge University Press.

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

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

Maher, P. 1990. "Symptomatic Acts and the Value of Evidence in Causal Decision Theory." *Philosophy of Science* 57:479–98.

Meacham, C. 2010. "Binding and Its Consequences." *Philosophical Studies* 149:49–71.

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

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

Skyrms, B. 1990. "The Value of Knowledge." *Minnesota Studies in the Philosophy of Science* 14:245–266.

Spencer, J. and Wells, I. Forthcoming. "Why Take Both Boxes?" *Philosophy and Phenomenological Research* .

Wells, I. Forthcoming. "Equal Opportunity and Newcomb's Problem." *Mind* .