Bayesianism, Infinite Decisions, and Binding[0]

Frank Arntzenius, Adam Elga, John Hawthorne

October 31, 2003

Penultimate draft. Revised version forthcoming in *Mind*

0. Introduction

When decision situations involve infinities, vexing puzzles arise. We describe six such puzzles below. (None of the puzzles has a universally accepted solution, and we are aware of no suggested solutions that apply to all of the puzzles.) We will use the puzzles to motivate two theses concerning infinite decisions. In addition to providing a unified resolution of the puzzles, the theses have important consequences for decision theory wherever infinities arise. By showing that Dutch book arguments have no force in infinite cases, the theses are evidence that reasonable utility functions may be unbounded, and that reasonable credence functions need not be either countably additive or conglomerable (a term to be explained in section 3). The theses show that when infinitely many decisions are involved, the difference between making the decisions simultaneously and making them sequentially can be the difference between riches and ruin. And they reveal a new way in which the ability to make binding commitments can save perfectly rational agents from sure losses.

Trumped[1]

Donald Trump has just arrived in Purgatory. God visits him and offers him the following deal. If he spends tomorrow in Hell, Donald will be allowed to spend the next two days in

[0] Thanks to Jamie Dreier, Robert Stalnaker, and an anonymous referee.

[1] "Trumped" is a modified version of a puzzle first put forward in Arntzenius and McCarthy (1997). See also the “EverBetter wine” puzzle of Pollock (1983, 417) and the “Devil's offer” puzzle of Gracely (1988). For a realistic variant of such puzzles, see Landesman (1995).

Heaven, before returning to Purgatory forever. Otherwise he will spend forever in Purgatory. Since Heaven is as pleasant as Hell is unpleasant, Donald accepts the deal. The next evening, as he runs out of Hell, God offers Donald another deal: if Donald spends another day in Hell, he'll earn an additional two days in Heaven, for a total of four days in Heaven (the two days he already owed him, plus two new ones) before his return to Purgatory. Donald accepts for the same reason as before. In fact, every time he drags himself out of Hell, God offers him the same deal, and he accepts. (See Figure 1.) Donald spends all of eternity writhing in agony in Hell. Where did he go wrong?

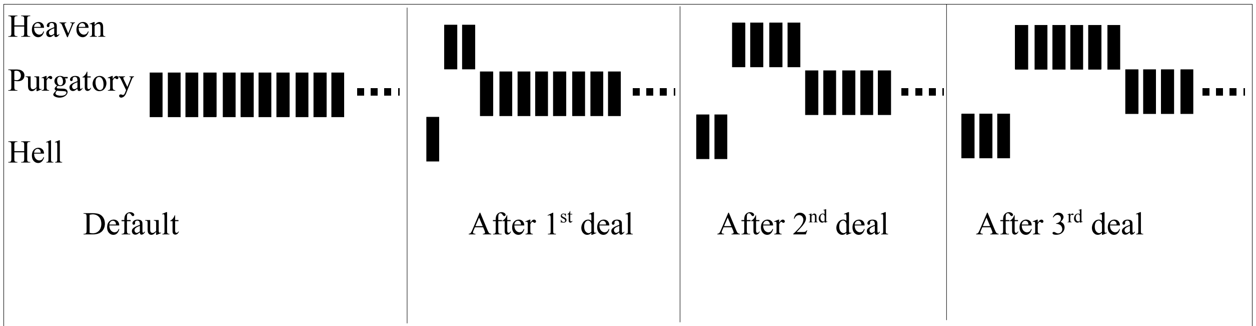

Figure 1: The afterlife that Donald has earned after each of God's deals. Donald's default afterlife is forever in Purgatory (each rectangle stands for one day). After the first deal, he has earned the right to two days in Heaven if he first stays one day in Hell. After the second deal, he has earned the right to two additional days in Heaven if he first spends one additional day in Hell.

Rouble trouble[2]

The Bank of Russia has offered Donald a sweet deal if he pays them a million dollars. If he pays for the deal, they will make him infinitely many offers of free money. In particular, at 12:00 they will offer him ten roubles, with serial numbers 1 through 10, on condition that he burns the

[2] "Rouble trouble" is a puzzle first put forward in Arntzenius and Barrett (1999). See also "Ross's paradox", described in Earman et. al. (1996, 239).

one with the lowest serial number, namely #1. If he declines the offer, nothing will happen at 12:00. At 12:30 they will offer him another ten roubles (the ten lowest-numbered roubles in their possession), on the condition that after he receives them he burns the rouble in his possession with the lowest serial number. At 12:45 they will offer him the next ten roubles, again on the condition that after he receives them he burns the lowest-numbered rouble in his possession. And so on. They will make him a countable infinity of such offers at times approaching 1:00.

Donald thinks a bit, and decides to accept the deal. He figures that even with inflation, there is a finite number of roubles that is worth more than a million dollars. By accepting the deal he expects to recoup his million dollars and then some.

The following events ensue. Every time the Bank makes Donald an offer, he accepts it and thereby increases the size of his pile by nine roubles. But to his utter amazement, at 1:00 he finds that he has no roubles at all. "Where did they go?!", he screams. "Well", explains a Russian official, "Where did #1 go?". "I burned that one at 12:00" responds Donald, still indignant. "How about #2?" asks the official. "I burned that at 12:30", responds Donald, who is becoming increasingly depressed as he sees where this reasoning is going. Still, he sputters "but I only burned one out of every ten I received". "Well", says the Russian "It is true that each time you received ten roubles you burned only one rouble, but that implies nothing about what you end up with at 1:00. What does matter is that every rouble that you received was labeled by a natural number, and that for every natural number there is a time prior to 1:00 at which the rouble labeled by that number was burned. It follows that all of the roubles we gave you were burned prior to 1:00". Donald, despondent, has one last go "But as the time approached 1:00, the total dollar amount in my possession kept increasing". "Yes Donald, but some functions are not continuous. And unfortunately the function that gives the total amount of dollars that you have as a function of time is provably (left-)discontinuous at 1:00".

That's all very well, but what should poor Donald have done?

Trouble in St. Petersburg[3]

[3] The puzzle that Bill faces when he goes to St. Petersburg is the "Airtight Dutch book" of McGee (1999).

Wandering along the Nevsky Prospect, Bill Gates encounters Ivan, a shady character who is casually tossing a coin. "Hey Biell, you vant some acsion?” Ivan asks. “Here's the deel. I offer some bets. You take or leaf. That's iet." Ivan explains that he will toss the coin until it lands tails, and offers Bill the following bets:

Bet 1: You lose $1 if it never lands tails, you gain $3 if it first lands tails on toss 1. If it first lands tails on some other toss, the bet is void.

Bet 2: You lose $4 if it first lands tails on toss 1, you gain $9 if it first lands tails on toss 2. If it first lands tails on some other toss, the bet is void.

Bet 3: You lose $10 if it first lands tails on toss 2, you gain $21 if it first lands tails on toss 3. If it first lands tails on some other toss, the bet is void.

And so on, i.e. loss(Bet n+1)=2xloss(Bet n)+$2, and gain(Bet n+1)=2xgain(Bet n)+$3. (See Table 1.)

	The coin never lands tails	***The coin takes 1 toss to land tails***	***The coin takes 2 tosses to land tails***	***The coin takes 3 tosses to land tails***	***The coin takes 4 tosses to land tails***
Chance	0	½	¼	1/8	1/16
Bet 1	Lose $1	Gain $3			
Bet 2		Lose $4	Gain $9		
Bet 3			Lose $10	Gain $21	
Bet 4				Lose $22	Gain $45

Table 1: A partial listing of the bets that Ivan offers Bill. Reading across each row, one can see that Bill counts each bet as favorable. Reading down each column, one can see that taking all of the bets guarantees a loss of $1. (Note that if the coin takes 4 tosses to land tails, then the $45 gain on bet 4 will be more than offset by a $46 loss on bet 5, which is not listed.)

Bill reasons as follows that each of these bets has positive expected dollar value. The chance that the coin will never lands tails is 0, so Bet 1 has a positive expected value. And for each of the other bets, although winning is only half as likely as losing, the reward for winning is more than twice as great. So each of the other bets has a positive expectation as well.

Since each bet has a positive expectation, Bill takes them all. Ivan says "Gud. No neet to sro ze coin. Just gif me vun dollar. Sanks." Ivan is right. If the coin first lands tails on toss n, Bill will lose one more dollar on bet n+1 than he gains on bet n and all other bets will be void. If it never lands tails Bill loses $1 on Bet 1 and all other bets are void. But where did Bill go wrong?

Two tickets to paradise[4]

Dutch state lottery officials have decided to hold a special lottery, which works as follows. There is one lottery ticket, A, which will be sold to the highest bidder. The day after it is sold, the dollar value of A will be determined by a chance process as follows:

$1 with chance 1/8
$2 with chance 7/32
$4 with chance 21/128
$8 with chance 63/512,

and so on. (The the remaining possible dollar values are all and only the remaining powers of 2. For all values n greater or equal to 2, if the chance of n is x/y, the chance of 2n is 3x/4y).

The expected dollar value of A thus is
1/8 ($1) + 7/32 ($2) + 21/128 ($4) + ...

The first term in this sum is $1/8. The next term $7/16, which is bigger. And each subsequent term is bigger than each previous term, since the probabilities get multiplied by 3/4 each time, and the dollar amounts double each time. So all terms are bigger than $1/4. So the

[4] The puzzle that Bill Gates faces in "Two tickets to paradise" is the "Two-envelope paradox" of Broome (1995), with an added Dutch book argument.

sum is infinite.[5]

Now, according to Bill Gates, dollars equal utilities. So he is willing to pay any amount for A. He does his research and wins the auction for the ticket. However, before the chance process which determines the worth of A has occurred, the lottery officials offer him a new deal:

"If you give us ticket A back, and pay us $0.01, we will give you a new lottery ticket B. The value of B will be determined as follows. After we have determined the value of A, we will roll a fair seven-headed die. If the die comes up 1,2 or 3, then B will be worth double what A is worth. Otherwise B will be worth half what A is worth. The only exception is that if A turns out to be worth $1, then, for sure, B will be worth $2. Do you want this deal?"

Bill thinks a little while and then accepts the deal. His reasoning is as follows. Given any amount $\$2^n$ that A could turn out to be worth, the expected value of this deal is positive. For if A is worth $1, then the deal, for sure, is worth $2-$1.01=$0.99. In any other case the expected value of the deal is

$$3/7x\$2^{n+1}+4/7x\$2^{n-1}-\$2^n-\$0.01=2/7x\$2^{n-1}-\$0.01,$$

which is always positive. So Bill takes the deal.

The Dutch officials then offer Bill another deal. They say: "We have with us a very special ticket. Its value will also be determined by a chance process. Given any possible value of B, the chance that this special ticket will be worth twice as much as B is 3/7, and given any value of B the chance that this special ticket will be worth half what B is worth is 4/7. The only exception occurs is if B is worth $1; in that case this ticket is worth $2. Do you want to give us ticket B and $0.01 in exchange for this very special ticket?" Bill reasons in exactly the same way as before that he should accept this deal. So he accepts the deal, and gives them B and $0.01.

To his astonishment the Dutch officials then give him ticket A back, and say that this concludes the deal. Despite Bill's astonishment, the Dutch officials are quite right: given any possible value of B (other than $1) the chance that A is worth half what B is worth is 4/7. And

[5] One might prefer to say that the expected amount is ill-defined rather than infinite, since the sum in question does not converge to a finite number. Whatever one wishes to call the expected amount in this case, the question that we discuss in this paper remains the same: is it coherent for people to prefer such an envelope to any envelope with a well-defined finite expectation?

given any possible value of B (other than $1) the chance that A is worth double what B is worth is 3/7.[6] But Bill is awfully puzzled. He started with ticket A. He then made two good deals, paid $0.02, and ended up ticket A again. Whatever did Bill do wrong?

Can God pick an integer at random?

God has created a countably infinite collection of planets. He tells Satan and the Archangel Gabriel that he plans to meddle with one of them, sparking the following conversation. Satan asks Gabriel, “which planet do you think He is going to pick?"

"I don't know,” Gabriel replies, “But I do know that He is completely fair. So I think each planet is equally likely."

"That is not possible."

"Of course it is! He is fair, and so each planet is equally likely."

"So how likely do you think it is that Earth will be chosen?"

"Same as every other planet, namely infinitesimal."

Satan then labels all the planets with natural numbers and offers Gabriel the following bets.

Bet 1: Gabriel pays Satan $2 if God chooses planet 1, and receives $1/2 from Satan otherwise.

Bet 2: Gabriel pays Satan $2 if God chooses planet 2, and receives $1/4 from Satan otherwise.

Bet 3: Gabriel pays Satan $2 if God chooses planet 3, and receives $1/8 from Satan otherwise.

.
.
.

Bet n: Gabriel pays Satan $2 if God chooses planet n, and receives $1/2n from Satan otherwise.

[6] If you want to check this quickly: the probabilities in this story are generated by

$\Pr(A=\$2^n \ \& \ B=\$2^{n+1})=1/8x(3/4)^n=\Pr(A=\$2^{n+1} \ \& \ B=\$2^n)$ for all nonnegative integers n. A very brief computation shows that this entails that $\Pr(A=\$2^n/B=\$2^{n+1})=4/7=(B=\$2^n/A=\$2^{n+1})$.

.
.
.

Gabriel counts each of these bets as favorable, since each offers a greater-than-infinitesimal chance of a finite gain, and threatens merely an infinitesimal chance of a finite loss. Satan then takes great delight in telling Gabriel that he has just lost more than $1. For Gabriel must lose one of his bets, at a cost of $2. And the amounts that he will win must add up to less than $1, since $1/2 + $1/4 + $1/8 + ... = 1. What did Gabriel do wrong?[7]

The magic dartboard[8]

Hansel and Gretel are lost in a dark German forest. They encounter a drunken mathematician who is throwing darts randomly at a black and white dartboard in the shape of a square. "My name is Sir Pinski," the mathematician slurs, "and I challenge you to the following betting game."

"After you are both blindfolded, one of you will throw a dart at the dartboard. Each point on the dartboard sits on exactly one vertical line and exactly one horizontal line. After your throw has landed, I will tell you, Hansel, what proportion of the vertical line that it landed on is white. And I will tell you, Gretel, what proportion of the horizontal line that it landed on is white. After I have given you that information you must not talk to each other. I will then offer each of you a bet on whether the dart landed on a white or black point. (You will be allowed to take whichever side of the bet you prefer.) If you together make money on that pair of bets, I will tell you the way out of the forest. We will play this game repeatedly until you make money on a pair of bets."

Hansel and Gretel can't believe their luck, and immediately accept. But to their dismay, every time Sir Pinski throws the dart, the following events ensue.

First Sir Pinski tells Hansel that the vertical line that the dart landed on is almost entirely

[7]This puzzle is an instance of the fact that failures of countable additivity entail vulnerability to countable Dutch books. See for example Seidenfeld (1983).
[8] This puzzle is based on a construction in Elga (1997).

black (i.e., measure 0 of it is white—see figure 2). Sir Pinski then offers Hansel either side of a 1:2 bet on black. In other words, he asks Hansel whether he wants to gain $1 if the point it landed on is black and lose $2 if it is white, or lose $1 if its black and gain $2 if it is white. Hansel always chooses to gain $1 if it is black and lose $2 if it is white, since he figures he is almost certain to win that bet.

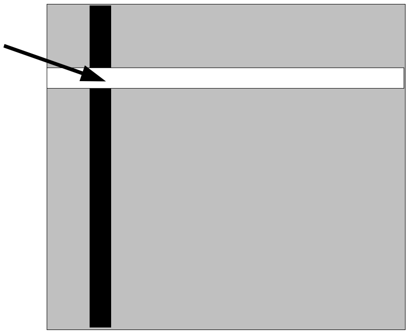

Figure 2. *Sir Pinski's dartboard. Notice that the dart (depicted as an arrow) sits on a horizontal line (thickened for illustration) that is almost all white, and a vertical line that is almost all black. Every point on the board has that same property.*

Then Sir Pinski tells Gretel that the horizontal line that the dart landed on is almost entirely white (i.e., measure 0 of it is black). He then offers Gretel either side of a 1:2 bet on white. In other words, he asks Gretel whether she wants to gain $1 if the point it landed on is white and lose $2 if it is black, or lose $1 if it is white and gain $2 if it is black. Gretel always chooses to gain $1 if it is white and lose $2 if it is black, since she figures she is almost certain to win this bet.

Sometimes the dart lands on white points, and sometimes on black ones. But either way, one of Hansel and Gretel wins $1 and the other loses $2.

This sequence of events is no coincidence. The dartboard is colored in such a way that measure 0 of *every* vertical line is white and measure 0 of *every* horizontal line is black. That such a dartboard is possible, assuming the axiom of choice and the continuum hypothesis, was first established by the Polish mathematician Sierpinski.[9]

Sir Pinsky's dartboard is colored in a bizarre way that involves nonmeasurable regions. But similar problems arise for boards colored in more mundane ways. For example, suppose that Gretel's beliefs meet the following condition:

> For any point of the board, and for any vertical line, one's degree of belief that the dart lands on the point, given that it either lands on the point or the line, is less than .999. In other words, for any point b and vertical line L:[10]
>
> $P(b \mid b \text{ or } L) < .999$.

That condition is very mild. To violate it, there must be a point and a vertical line such that, given that the dart landed either on the point or the line, one is almost certain that the dart landed on the point. If one's probability distribution over landing places for the dart is anything near smooth, there will be no such point and line. Nevertheless, if Gretel's beliefs meet this mild condition, a disastrous result follows.

It follows that there is a partition **T** of the board (a set of disjoint regions whose union is

[9] Proof: Identify the dart board with the unit square S = [0,1] x [0,1]. By the axiom of choice and the continuum hypothesis, there is a bijection f from [0,1] onto the set of countable ordinals. Let the set of B of black points be those points <x,y> for which f(x) < f(y). B intersects every horizontal line only countably many times. Therefore, by the countable additivity of Lebesgue measure, the horizontal line is almost all white. Similarly, the set W of white points intersects every vertical line only countably many times. Therefore, every vertical line is almost all black. Notice that B and W must fail to be Lebesgue measurable (else we would have a violation of Fubini's theorem, which allows one to reverse the order of integration when calculating the double integral of a measurable function). The authors learned of this type of construction from Pitowski (1983).

[10] Here we are letting “b” do double duty as both a name for a point and a name for the proposition that the dart lands at that point. We will continue with similar abuses.

the entire board), a rectangular region R, and a number x, so that

(1) Gretel's credence in R is less than x.

(2) For any T in **T**, Gretel's credence in R given T is greater than x.[11]

It further follows that Gretel can be Dutch booked. In other words, it follows that there is a set of bets, each of which Gretel finds favorable, which together lead to a sure loss. For example, suppose that P(R) = 1/3, and that for all T in **T**, P(R|T) = 2/3. Then Gretel will find this bet favorable:

> If the dart lands in R, pay \$3. If the dart lands outside of R, receive \$2.

And for each T in **T**, she will find this conditional bet favorable:

> If the dart lands in T, the following arrangement is activated: if the dart lands in R, receive \$2; if the dart lands outside of R, pay \$3.

Suppose that Gretel accepts all of the bets. Then no matter where the dart lands, exactly one of the conditional bets will be activated. So if the dart lands in R, Gretel will lose \$3 on the unconditional bet and gain \$2 on a conditional bet. On the other hand, if the dart lands outside R, she will win \$2 on the unconditional bet but lose \$3 on a conditional bet. In other words, if she accepts all of the bets she is sure to lose \$1. Where did Gretel go wrong?

1. Decisions sets and Binding

Our protagonists have all run into the same sort of trouble: they deemed each of a

[11] Proof: Again we identify the dart board with the unit square S = IxI = [0,1]x[0,1]. Let r = .999. (Any number less than 1 would work equally well.) Choose an open interval (m,n) such that P((m,n) x I) < 1-r. (To see that such an interval exists, let n be greater than 1/(1-r) and divide I into disjoint nonempty intervals I_1, I_2, ..., I_n. At least one of the regions I_k x I (1 <= k <= n) must get probability less than 1-r; otherwise their union would get probability greater than 1, which is impossible.) Let R = (m,n) x I. Choose a bijection f: (m,n) -> S - R. Use f to construct a partition **T** = {({x}x I) union {f(x)}: x in (m,n)}, each of whose elements is the union of a vertical line of R and a single point outside of R. For all T in **T**, P(R|T) = 1-P(S-R|T). Now, P(S-R|T) = P({b}| L union {b}) for some b in S-R and some vertical slice L of R. So by assumption, P(S-R|T) < r. Therefore for all T in **T**, P(R|T) > 1-r. Since R was chosen so that P(R) < 1-r, we are done.

countable collection of acts to be beneficial, but performing all of the acts led to a sure loss. Two unsatisfactory responses suggest themselves. The first response is to restrict decision theory so that it does not apply to the situations described. For example, one might ban decision situations with infinitely many options, require rational agents to have upper bounds on their utility functions, and ignore cases in which agents divide their credence among infinitely many alternatives. We recognize this craven line of retreat as a potential last resort, but as nothing more. For we are loath to constrain the scope of decision theory with such seemingly ad hoc bans. And we would be unsatisfied with a resolution of the puzzles that didn't reflect their common character.

Another response is to augment decision theory with new norms that yield special global judgments of rationality—judgments that don't attach to particular decisions, but rather to sequences of decisions. For example, one might grant that in “Trumped”, *each* of Donald's decisions to stay one more day in hell was perfectly rational, but insist that he has exhibited some sort of *global* irrationality in the pattern of his decisions. This response is unsatisfactory as well. Even if there were such irreducibly global norms of rationality, such norms would have no motivational force. Decision-makers face individual decisions at particular times. If each of an agent's choices is perfectly rational, then she is beyond rational reproach, as far as decision theory is concerned.

In contrast, we will argue that ordinary (causal) decision theory, carefully applied, is capable of handling all of the puzzles. We begin with "Trumped" and "Rouble Trouble", in which the fate of the protagonist involves no chancy lotteries. In section 2 we turn to “Two tickets to paradise” and “Trouble in St. Petersburg”, both of which involve unbounded utilities. Section 3 addresses “The magic dartboard” and “Can God pick an integer at random?”, both of which involve non-conglomerable probabilities. Section 4 exhibits some bizarre violations of van Fraassen's (1984) "Reflection Principle" that arise from instances of non-conglomerability.

*

In “Trumped” and “Rouble Trouble”, Donald accepts an infinite sequence of deals or offers. Each offer (receive ten roubles and burn one; spend one day in Hell to earn two days in Heaven) seems favorable, and yet accepting them all is disastrous. In accepting them all, where has Donald's reasoning gone wrong?

First notice that when deciding whether to accept one offer, Donald fails to consider what future offers he will accept. That's a mistake. For Donald's fortune depends on the complete pattern of offers he accepts and rejects. And in both of these cases, an offer that would be favorable on its own need not be favorable in combination with other offers. For example, consider the Bank of Russia's first offer: receive roubles #1-#10 on condition that he burn #1. Considered on its own, that offer is favorable, since it results in a net gain of nine roubles. But suppose that Donald will in fact accept infinitely many of the Bank's subsequent offers. In combination with those later acceptances (which guarantee that he burns every ruble he gets!), the Bank's first offer is not favorable. Given that he accepts infinitely many of the later offers, Donald ends up with only burnt rubles in his pocket whether or not he accepts the first offer.

A similar point holds for "Trumped": when deciding whether to accept one of God's deals, Donald should consider what future deals he will accept. But in the statement of the puzzle, we did not supply enough information to determine how he should do so. For example, we did not say what happens if Donald ever declines to stay in Hell one more day. Does God come back after Donald has spent some time in Heaven and/or Purgatory and make him another offer? Does Donald know in advance what offers will be made? We'd need to answer these questions in order to determine which of God's offers Donald should accept.

The bottom line: in “Trumped” and “Rouble Trouble”, Donald reasons incorrectly by failing to consider what future offers he will accept. That explains what is wrong with Donald's reasoning, and so resolves the puzzles, as stated. But that resolution is unsatisfying, for there is a more challenging version of the puzzles. Like "Trumped" and "Rouble Trouble", the more challenging puzzle features an agent who makes a number of seemingly favorable decisions, only to find that the decisions together lead to disaster. But unlike the previous puzzles, the agent does not make the mistake that Donald does (the mistake of failing to consider future offers, when evaluating a present one). The analysis of the more challenging puzzle will lead to two lessons, lessons which we will put to use to resolve all of the remaining puzzles.

Satan's Apple

Satan has cut a delicious apple into infinitely many pieces, labeled by the natural numbers. Eve may take whichever pieces she chooses. If she takes merely finitely many of the pieces, then she suffers no penalty. But if she takes infinitely many of the pieces, then she is expelled from the Garden for her greed. Either way, she gets to eat whatever pieces she has taken.

Eve's first priority is to stay in the Garden. Her second priority is to eat as much apple as possible. She is deciding what to do when Satan speaks. "Eve, you should make your decision one piece at a time. Consider just piece #1. No matter what other pieces you end up taking and rejecting, you do better to take piece #1 than to reject it. For if you take only finitely many other pieces, then taking piece #1 will get you more apple without incurring the greed penalty. On the other hand, if you take infinitely many other pieces, then you will be ejected from the Garden whether or not you take piece #1. So in that case you might as well take it, so that you can console yourself with some additional apple as you are escorted out."

Eve finds this reasonable, and decides provisionally to take piece #1. Satan continues, "By the way, the same reasoning holds for piece #2." Eve agrees. "And piece #3, and piece #4 and..." Eve takes every piece, and is ejected from the Garden. Where did she go wrong?

There are two versions of the above puzzle. In the *synchronic* version, Eve must decide all at once on a complete profile of which pieces to take and which to reject. In the *diachronic* version, Eve must first decide whether to take piece #1, then decide whether to take piece #2, etc. These versions call for different analyses, and so we consider them in turn.

Start with the synchronic version. Eve must make a single, gigantic decision: she must decide on a complete profile of which pieces to take and which to reject. Satan has an argument that she is rationally required to take all of the pieces. Where does the argument go wrong?

Satan's argument works one piece at a time. He notes that for piece #1, taking dominates rejecting. That means: regardless of what other pieces Eve takes or rejects, she does better to take piece #1 than she does to reject it. Satan infers from this that Eve is rationally required to choose a complete profile that involves taking piece #1.

Satan's inference is incorrect.

From the fact that for piece #1, taking dominates rejecting, it does not follow that Eve is rationally required to choose a profile that involves taking piece #1. Why, then, does the inference seem so plausible? And what is wrong with it?

The inference seems plausible because in finite cases—the sorts of cases on which our intuitions are based—similar inferences are correct. For example, suppose that you must decide all at once on a complete outfit to wear to a party. You have a big choice of what complete outfit to wear. And this big choice can be thought of as a product of a finite number of sub-choices: what shoes to wear, what socks, what pants, and so on. Suppose that for each category of clothing, you have finitely many options. Finally, suppose that, when it comes to pants, your super-cool green pants are the dominant option. That means: regardless of what choices you make for your other clothing categories, your green pants combine better with those choices than any of your other pants do.

From the fact that your green pants are the dominant choice of pants it *does* follow that you are rationally required to choose an outfit that includes them. Proof: You have a choice of finitely many outfits. Therefore, at least one of these has maximal utility, and so you are rationally required to choose an outfit that has maximal utility. But only outfits that include your green pants have maximal utility. That's because outfits that don't include the green pants can be improved by replacing their pants with the green pants. (Notice that this reasoning depends on your having merely finitely many outfits to choose from. For if you had infinitely many outfits to choose from, there would be no guarantee that any of them had maximal utility. So in that case you needn't be rationally required to choose an outfit with maximal utility.)

Moral: if you are choosing among finitely many outfits, and your green pants are your dominant choice of pants, then you are rationally required to choose an outfit that includes them. And the corresponding "dominance" inferences in other finite cases are right, as well.

But Eve's case is not a finite case. And in infinite cases, such inferences can go wrong. Proof: In Eve's case, for any particular piece of apple, taking that piece is the dominant option with respect to that piece. So if rationality always required one to choose dominant options, it would require Eve to take every piece. But it is absurd that rationality requires Eve to take every piece, since she has plenty of better overall choices available (for example, taking no pieces). So rationality doesn't always require one to choose one's dominant options. Even though taking piece #1 is the dominant thing to do with respect to piece #1, it doesn't follow that rationality requires Eve to take piece #1. In slogan form:

> In infinite cases, rationality does not require one to choose one's dominant options.

That is the first lesson of “Satan's apple”, and the first of the two theses mentioned in our introduction. It is the key to resolving the synchronic versions of all of the puzzles.

*

In the synchronic version of "Satan's apple", Eve must all at once decide on a complete profile of which pieces to accept and which to reject. Satan argues that she should take every piece, and we have seen where that argument goes wrong. But the question remains: what profile *should* Eve choose?

Notice that for any profile Eve might choose, there is a superior one. (Taking finitely many pieces is better than taking infinitely many. And for any choice of finitely many pieces, there is a superior choice that involves taking those pieces, and one additional piece.) So whatever she does, she'll later be in a position to see that she could have done better.

This problem is not particular to the above type of puzzle. Whenever one has no best option, there is no univocal answer to the question, "What should I do?" Suppose, for instance, that God offers to let you live any finite time of your choosing. Assuming that your utility is an increasing bounded function of the length of your life, there is no answer to the question: what life-span should you choose? Where there is a lowest upper bound on your utility, one could perhaps give useful vague guideline: pick a large number. By picking a large number, you can come very close to the lowest upper bound on your utility. So you should pick a very large number. (In Eve's case, we might say that she should take some very large finite number of pieces.) But if your utility function is linear with length of life (and so has no upper bound) not even such guidelines are available (except relative to some arbitrarily stipulated threshold). In such a case, we should set aside the question of what is rational simpliciter, and instead make do with a ranking of the rationality of various choices. According to such a ranking, someone who asks for one year of life acts better than someone who asks for zero years. Someone who asks for two years of life acts better than someone who asks for one. And so on.

In addition to the normative issue, there is something of a motivational puzzle here. What exactly would cause you to ask for one lifespan rather than another? But this puzzle is nothing new. We are already used to the idea that, pace Buridan, an ass confronted with equally attractive bales of hay will go to one of them rather than die of indecision.

*

So much for the synchronic version of "Satan's apple". What about the diachronic version? In that version, Eve does not make a single big decision about what profile of pieces to take an reject. Instead, she must first decide whether to take piece #1, then decide whether to take piece #2, and so on. (We may suppose that the decisions are made faster and faster, so that after an hour, they have all been made.)

What Eve should do depends on two factors. It depends on whether she can bind herself to future courses of action. And it depends on the manner in which she expects that her decision about one piece will influence her decisions about future pieces. Neither factor was specified in the statement of the puzzle, and so the puzzle is incomplete as it stands. But let us fill in these missing details in a few different ways, to see how they affect what Eve should do.

First, suppose that Eve is capable of irrevocably binding herself to a plan of action. For example, she might at the start commit herself to taking the first million pieces, and no others. Once she has set herself on that course, she will be unable to reconsider, and so will have no future decisions. Given this ability, Eve has more than two options when she is faced with piece #1. For in addition to deciding whether to take piece #1, she also has the opportunity to bind herself to whatever sequence of future choices she wishes.

When Eve has the power to bind herself in this way, her options at the start resemble her options in the synchronic case. Just as in the synchronic case, there was no best total *profile* for her to choose, in the diachronic case there is no best *plan* for her to choose. The result: another situation in which an agent has an infinite sequence of increasingly good options. And another situation in which there is no univocal answer to the question, "Which of these increasingly good options should the agent choose?"

Now suppose that Eve cannot irrevocably commit herself to a plan: at each decision point, she will be able to reject or revise whatever provisional plans she may have made up until then. Consider her decision whether to take piece #1. Whether she should take piece #1 depends on what she believes about how taking piece #1 will influence her future choices. For example, suppose that Eve is confident that taking piece #1 would cause her to take all subsequent pieces, and refusing piece #1 would cause her to refuse all subsequent pieces. Given those beliefs, she ought to refuse piece #1, since she prefers taking no pieces to taking all of them.[12]

The above example shows that there was a missing premise in Satan's argument, applied to the sequential version of the puzzle. Satan noted that—holding her subsequent choices fixed—Eve does better to take piece #1 than to refuse it. That much is correct. But he concluded that she should take piece #1. That conclusion does not follow without an additional assumption. For as we just saw, it may be that Eve thinks that her present choice can influence her subsequent choices. If so, Eve should *not* hold fixed her subsequent choices in deciding what to do. Instead she should take into account how her present choice will influence them.

Satan's conclusion does follow with an additional assumption: that Eve is sure that she

[12] As Jamie Dreier has pointed out to us, this requires that Eve be confident that there is an extremely robust causal connection between her present choice and her future choices. For suppose that Eve is confident that her future choices are subject to independent chancy influences (however tiny). For example, Eve might be confident that if she refuses the current piece, then for each subsequent piece, the chance that she will accept that piece is 0.0000001. And she might be confident that these chances are independent. Then she will be confident that with chance 1, she will end up accepting infinitely many pieces of apple. For the same reason, the ability to bind herself only helps Eve if it is an extremely robust sort of binding, not subject to independent chancy influences.

cannot by her present choice influence her future choices. Given that belief, Eve *is* entitled to treat her future choices as fixed. So if Eve has such a belief when deciding whether to take piece #1, she is rationally required to take piece #1. And if she always believes that her present choice has no influence over her future choices—call this the “no influence” case—then she is rationally required to take every piece.[13]

This can seem an unwelcome conclusion. It is the conclusion that in the no influence case, Eve is rationally required to eject herself from the Garden!

One might try to avoid that conclusion by suggesting that Eve should make a (defeasible) plan at the beginning of the experiment, and then stick to it. Eve might declare at the beginning: "I will take the first million pieces, and that's all!". But we have supposed that Eve lacks the ability to irrevocably bind herself to any plans. So despite her declaration, after she has taken the first million pieces, she will have the option of reneging on her plan. Furthermore, she will have compelling reason to do so. The reason: taking the million-and-first piece has a higher expected utility than refusing it. Therefore making defeasible plans does not help Eve at all.

[13] Another case to consider: Fred must say "0" or "1" on an infinite series of occasions. If he says "1" on infinitely many of those occasions, he wins a big prize. If not, not. He doesn't have the ability to irrevocably bind himself to courses of action. Furthermore, on each occasion he is certain that his choice on that occasion will not influence what he chooses on future occasions. At each step, Fred reasons as follows: “No matter what I say in the future, my payoff will be the same whether I say '0' or '1' now. So I am indifferent between saying '0' or '1'.” In fact, Fred says "0" on every occasion, and so does not win the prize. Has Fred exhibited any failure of rationality?

We say that the case has been underdescribed. Even though Fred does not have the ability to make an irrevocable plan, he may have a more modest ability: the ability to—at the start of the game—form an *intention* to say “1” on every occasion. We assume that forming such an intention would make it more likely that Fred ends up saying “1” on infinitely many occasions. If Fred has such an ability, and nevertheless does not exercise it, then he has been irrational in failing to do so. On the other hand, if Fred lacks that ability, or if he forms the intention to always say “1” but gets unlucky and ends up saying “0” on every occasion, then Fred has made no irrational decisions.

The above analysis is guided by the principle that there are no irreducibly global norms of rationality, a principle that we defended briefly at the end of section 1. Those who accept irreducibly global norms of rationality may say that when Fred chooses all “0”s, his sequence of choices may be irrational even though no individual one of them is irrational. For an analysis of this kind (of a game-theoretic example involving imperfect recall), see section 6 of Stalnaker (1999). We are indebted to Robert Stalnaker for helpful correspondence on this point.

So we stand by the conclusion that in the no influence case, Eve is rationally required to eject herself from the Garden. To make the conclusion more palatable, consider a multi-person version of the case, in which Eve is replaced by an infinite sequence of advisors. The first advisor is in charge of deciding about piece #1, the second about piece #2, and so on. The advisors are forced to make their decisions simultaneously, in separate cubicles. Each advisor has only Eve's interests at heart. What should they do?

Consider advisor #37. He is certain that his choice won't influence any of the other advisors. Given that certainty, he should advise in favor of taking piece #37. (To make this more vivid, we may suppose that all of the other advisors have made their choices, and that it only remains for #37 to choose.) And the same goes for the other advisors. The incentive structure of the situation makes it impossible for the advisors to rationally coordinate their choices in a way that keeps Eve in the Garden.

Note that if advisor #37 believes that he is similar to the other advisors, then by refusing his piece, he can make himself *more confident* that they refuse their pieces as well—even though he cannot causally influence their decisions. An evidential decision theorist would count this as a reason for advisor #37 to refuse his piece. Being causal decision theorists, we cannot condone this reasoning.[14]

The lesson is that under certain circumstances, the following ability can be incredibly helpful: the ability to have one's present choices causally influence one's future choices. In its most extreme form, this amounts to the ability to irrevocably bind oneself to future courses of action. It's old news that such an ability helps those who—like Ulysses sailing toward the Sirens—expect their values to be distorted in a way they'd like to resist.[15] What we have seen is that such an ability can also be indispensable in situations in which one's values remain unchanged, and in which one remains perfectly rational at all times.

[14] On causal decision theory, see Lewis (1981).
[15] See Gauthier (1997).

The lack of such ability is not, we say, a deficiency in rationality. (We do not wish to indulge in endless quibbles concerning the term 'rationality' here. But it does seem ill-advised methodologically to respond to the puzzles by proposing sundry ad hoc constraints on the true meaning of rationality, rather than staying with the broad outlines of the Bayesian conception that is both intuitively compelling and reasonably well understood.) But where a capacity to bind is available (say in the form of an ability to form a plan and cause oneself to be someone who will unreflectively follow it), that capacity can be useful exploited. In brief, rational individuals should welcome the capacity to bind themselves. Meanwhile:

> Rational individuals who lack the capacity to bind themselves are liable to be punished, not for their irrationality, but for their inability to self bind.

Two theses have sprung to light in this section. First: suppose that an agent is faced with an infinite number of offers, and that she must decide all at once on a profile of which offers to accept and which to reject. Suppose that for each offer, “taking” dominates “rejecting”: each offer is such that—no matter what other offers she takes or rejects—she does better to take it than to reject it. We saw that it does *not* follow that she should accept all of the offers. Instead she should select a combination of offers that produces an outcome that she desires.

Second: sometimes a single individual is faced with the prospect of a infinite sequence of offers, delivered over time. If that individual is able to bind his future self, this case can be treated on the model of the previous one. If not, and if the agent believes that his present choices do not influence his future ones, then the agent may be unable to rationally coordinate his choices in order to avoid a disastrous outcome. Some agents who are led to ruin in this way are perfectly rational. It's just that certain situations exploit rational agents who are unable to self-bind.

2. Unbounded Utilities

In "Two tickets to paradise" and “Trouble in St. Petersburg”, Bill Gates gets into trouble because of his unbounded utilities. (In other words, for any natural number *n*—however large—

there is a state of the world to which Bill assigns utility greater than *n*.) In each puzzle, he is offered a countable sequence of (conditional) bets. Each bet, considered on its own, has a positive expected value. But accepting all of the bets together guarantees a net loss.[16]

The standard way to avoid this kind of trouble is to require rational agents to have bounded utility scales. We would prefer not to impose this restriction. Some seemingly perfectly coherent systems of preferences can only be represented in terms of degrees of belief and utilities if these utilities are unbounded. For instance, suppose that Methuselah has beliefs which accord with the Principal Principle (roughly, the principle that one's degrees of belief ought to accord with one's estimate of the objective chances; see Lewis (1980)). Suppose that Methuselah were to truly report, “For any real number *p* greater than zero, there exists a natural number *n* such that I prefer chance *p* of extending my life by *n* years, to receiving $1 with certainty.” Methuselah's beliefs and preferences seem to be perfectly coherent. But his utility scale must be unbounded.[17,18] It would be nice if decision theory could accommodate such people.

What stands in the way from it doing so are puzzles such as “Trouble in St. Petersburg” and “Two Tickets to Paradise”. But we can resolve such puzzles using the two main themes from the previous section, as follows.

*

Each puzzle comes in two versions: synchronic and diachronic. We will focus on the

[16] The reason for this surprising fact is that when one has an infinite collection of bets with unbounded payoffs, the sum of the expectations need not be equal to the expectation of the sum. Indeed the expectations can all be positive, while the expectation of the sum is negative. See Norton (1998).

[17] Proof: For simplicity, assume that Methuselah increases his utility by U(n) units when he extends his lifespan by n years, and by U(dollar) units when he receives a dollar. Then for any p>0, there is an n such that pU(n) > U(dollar), which entails that U(n) > (1/p)U(dollar). But (1/p) approaches infinity as p approaches zero, so for any finite bound B, there must be an n such that U(n) > B.

[18] In this example the representation in terms of utilities and credences is not only constrained by his preferences, it is also constrained by the fact that his credences have to equal the objective chances. One can also give examples in which just the preferences force a representation in terms of unbounded utilities. See, for instance, Jeffrey (1983, chapter 10).

synchronic version of “Trouble in St. Petersburg” and some diachronic variants of “Two tickets to paradise”.

In the synchronic version of “Trouble in St. Petersburg”, Bill must all at once choose a complete profile of which of Ivan's bets to accept and which to reject. Ivan might argue as follows: “Bill, think about this one bet at a time. Start with bet #1. *Premise*: No matter what other bets you take or reject, you do no worse (in expected gain) to take bet #1 than to reject it. *Conclusion*: you should choose a profile that includes bet #1. A similar argument shows that you should take each of the remaining bets as well.” Ivan's premise is correct. But his conclusion does not follow.

Given the lesson from the synchronic version of Satan's Apple, this is no surprise. In the Satan's Apple case we saw that just because taking piece #1 dominates rejecting it, Eve is not rationally required to accept a set of pieces that includes #1. Similarly, in St. Petersburg case, just because accepting bet #1 (weakly) dominates rejecting it, Bill is not rationally required to accept a set of bets that includes bet #1.

What set of Ivan's bets *should* Bill accept? As before, there is no best set. In fact, matters are slightly more complicated, since there are many sets of bets which have no well-defined expected value. For instance, in the St. Petersburg case, the set consisting of the odd-numbered bets has no well-defined expected value, since its expected value equals

$$\tfrac{1}{2}\ (\$3) + \tfrac{1}{4}\ (-\$10) + 1/8\ (\$21) + 1/16\ (-\$46) + 1/32\ (\$93) + \dots .$$

The probabilities in this sum get halved in each consecutive term, and the utilities switch sign and are more than doubled in each consecutive term, so this sum does not converge. Nonetheless there are many sets of bets with well-defined expected utilities. Among those, there are better and worse sets. Accepting all of Ivan's bets, for instance, is worse than accepting none of them, and accepting none of them is worse than accepting a finite number of them. As for the sets of bets whose expected value is undefined: there is in general no answer to how they compare with other sets. (That's a fine conclusion—it is no embarrassment to an analysis that it offer no verdict where there is, intuitively, none to be had.)

In the last section we contrasted a case in which a single individual controls each of an infinite collection of offers with a case in which an infinite collection of individuals each have control over a single offer from the same collection. That contrast is also instructive here. Consider a variant of the St. Petersburg situation. Suppose that there is a countable collection of people, and that person #n takes out bet #n with Ivan. Ivan is guaranteed to make $1. And each person is making a bet with positive expected utility. It is a good deal for everyone! Of course the countable collection of people is guaranteed to make a $1 loss. But nonetheless each one of them is taking a bet with positive expected utility. How can this be? Surely there are no free lunches! We say that there *are* free lunches. But oddities such as free lunches are unsurprising, given that infinities are involved.

For example, suppose that a countably infinite group of people can, at least temporarily, avail themselves of an unbounded good. They can make a pattern of non-probabilistic deals that benefit each and every one of them. Imagine friends # 0,1,2,3, ... all standing in a line. For each *n*, friend *n* hands $*n* to friend *n*-1. After this transaction each one of them will have gained $1. In such circumstances non-probabilistic free lunches exist. So probabilistic free lunches, when there are unbounded utilities, are to be expected.

In fact we can transform the St. Petersburg case into a probabilistic free lunch for each of a countable collection of people, each of whom has *bounded* utilities. Instead of having one person takes out bet #n with Ivan, divide bet #n up into as many bets as there are dollars to be lost in bet #n. Then have Ivan offer each of these bets to a different person. Each bet has a potential gain of more than $2 set against a potential loss of merely $1. As a result, each bet has positive expected utility, even for people with bounded utility functions. And the package consisting of all of the bets still has positive expected utility for Ivan. So in order to construct a free lunch setup, we merely need an unbounded supply of some good, and people who always prefer more of the good to less of it. This indicates that the puzzles have little to do with unbounded utilities, but have much to do with the fact that the sum of expectations need not equal the expectation of the sum.

*

A second theme from the last section is the advantage of binding: someone with Bayesian rationality and the ability to bind himself can expect certain advantages over a Bayesian with no such ability. The idea can be put to good work in connection with some variants of "Two Tickets to Paradise". First a diachronic puzzle: The lottery from “Two Tickets to Paradise” has been run in secret, and the monetary values of tickets A and B have been determined. You have not seen the lottery outcome or the resulting ticket values. You will play the following two-round game.

Round 1: You will be told the value of ticket A. Then you will be given the following choice: “Either **stick**, in which case nothing happens, or else **pay** $0.01 to have [value of B - value of A] added to your bank account.”

Round 2: Your memory of the value of ticket A will be erased (though not your memory of your choice on the previous round). You will then told the value of ticket B. Finally, you will be given the following choice: “Either **stick**, in which case nothing happens, or else **pay** $0.01 to have [value of A - value of B] added to your bank account.”[20]

At each stage, the option of paying offers a positive expected gain. Yet paying on both rounds results in a net loss of $0.02, while sticking on both rounds results in breaking even.

Suppose that you lack the ability to self-bind. You are told that ticket A is worth $4. Suppose that you form the plan "I will pay in the first round, but stick in the second round." You pay at the first round, have your memory erased, and are informed that ticket B is worth, say, $8. You must now make your choice for the second round. By hypothesis, you remember your earlier plan. But why stick to it? You now expect to gain by paying. Of course, you realize that by paying in both rounds you will end up losing $0.02, which is worse than what you would have gotten by sticking in both rounds. But having paid in the first round, the option of sticking on both rounds is no longer available. And from your current perspective, you reckon it rationally compulsory to pay.

[20] Note that unlike standard two envelope and St. Petersburg games, the relevant decisions in this and the preceding game are made at times when one's expected return from payment is *finite*.

The situation is not one where a failure of rationality is displayed. For how can the reasoning at either round be faulted? The situation is rather one where the rational person is punished. Lacking the ability to control his future self, the earlier self launches a sequence of actions that can forseeably be improved upon but is forseeably unavailable given the rational dispositions of his future self.

The situation is different, of course, for an agent who can irrevocably commit to a plan. Before playing the game, such an agent can and should commit to sticking in both rounds. Such an agent need never find herself in the situation where she helplessly pays to switch twice.

The problem becomes even more poignant—and the importance of the capacity to self-bind even more vivid—in a social version of the last puzzle. Suppose that neither you nor your friend know the values of tickets A or B. And suppose that the two of you agree to equally divide your net gains from playing the following game. Tomorrow the Dutch lottery officials will put you and your friend in separate rooms. They will tell you the value of ticket A, and offer you the following choice: "Either **stick**, in which case nothing happens, or else **pay** $0.01 to have [value of B - value of A] added to your bank account." They will tell your friend the value of ticket B, and offer him the following choice "Either **stick**, in which case nothing happens, or else **pay** $0.01 to have [value of A - value of B] added to your bank account."

You both know in advance that given the value of A, no matter what that value is, the expected utility of B is more than $0.01 higher than that of A. And you know in advance that given the value of B, no matter what that value is, the expected utility of A is more than $0.01 higher than that of A. So once you and your friend are told the values of your respective tickets, each of you will have incentive to pay. But if you both pay, the pair of you will lose $0.02. So unless you are able to commit to a plan beforehand, the two of you will forseeably lose $0.02. Furthermore, the problem is not that there is some conflict of interest between you and your friend. There is no conflict. What benefits you benefits your friend, and vice versa.

What if two players in the above game are allowed to communicate, once one player has been told the value of ticket A, and the other has been told the value of ticket B? (Assume that neither player is allowed to reveal the ticket value that he has been told.) "There is no point in us both paying," each will agree. The trouble is that the players will have incompatible preference rankings. Each player will prefer most his paying, second, neither paying, third, both paying, last just the other paying. Suppose that the two compromise and agree that both will stick.

But now consider what happens when it is time to act. If each acts according to causal decision theory, and attaches no particular utility to keeping agreements for their own sake, the fact of the agreement will not make any difference: each will pay. Repeated episodes of the game will make no difference. Rationality will be punished again and again.

*

Summing Up: We can resolve puzzle cases such as "Trouble in St. Petersburg" and "Two tickets to paradise" without requiring that all agents have bounded utility functions. We can do so by attending to our two main themes: (1) When analyzing the choice of an agent who makes infinitely many decisions at once, take the agent's options to be complete patterns of decisions. (2) Recognize that in various diachronic versions of the puzzles, rational agents who cannot causally bind their future selves will be hampered relative to those for whom such binding procedures are available.

3. Non-conglomerable probabilities

Suppose that conditional on its being cold tomorrow, you are confident that it will be sunny. Suppose further that conditional on its *not* being cold tomorrow, you are *also* confident that it will be sunny. It would be odd indeed if you in addition were confident that it *wasn't* going to be sunny tomorrow. The odd feature your probability function has in this case is *nonconglomerability*. Your beliefs are nonconglomerable when there is a proposition A, a partition of propositions (in the above case {*Sunny*, *Not sunny*}), and a number x such that:

- Your credence in A is less than x. (In other words, $P(A) < x$.)
- Conditional on each member T of the partition, your credence in A is greater than x (in other words, $P(A|T) > x$).

Consider the following example. One can assign equal probability to each possible outcome n of a countable lottery by assigning equal infinitesimal probability to each integer n.[21] In doing so, one adopts a probability function that violates countably additivity. (A probability function P is *countably additive* iff P(A1 v A2 v ...) = P(A1) + P(A2) + ... for any pairwise incompatible propositions A1, A2, ...) Consider now a set of outcomes that has probability ½—say the even numbers. Let us partition the set of all possible outcomes into the following subsets:

S_1={1, 2, 4}, S_2={3, 6, 8}, S_3={5, 10, 12}, ...

(Each subset contains one odd number and two even numbers.) Since each number gets equal infinitesimal probability, the probability of an odd number conditional upon any element S_n of this partition is 1/3. So this probability distribution violates conglomerability. As Gabriel discovers to his chagrin in "Can God pick an integer at random?", having such a probability function makes one vulnerable to countable Dutch books.

Gretel learns a similar lesson in "Sir Pinsky's dartboard". Recall that for any point and any vertical line, Gretel's credence that the dart will land on the point, given that it will land either on the point or the line, is less than 0.999. It follows that Gretel has non-conglomerable degrees of belief.[22] As a result, whether or not Hansel is in the picture, Gretel will be subject to an uncountable Dutch book.

What should Gabriel and Gretel do?

Start with Gabriel. Faced with a collection of bets that together guarantee a net loss, which combination of them should he accept? It's a familiar story: in the synchronic case, there is no best combination. Gabriel does well to accept any finite number of Satan's bets, and the more the better. In the diachronic case, what Gabriel should do depends on how he expects his present choices to influence his future choices, in the way outlined in the discussion of the diachronic version of "Satan's Apple".

Noncongomerable probabilities also make one vulnerable to diachronic Dutch books. For example, assume that Gretel's probability function is noncongomerable for the following reason: There is a region R of the board and a partition **T** of regions of the board such that P(R) = 1/3,

[21] For a highly readable exposition of the construction that enables one to do so, see Appendix 6 of Skyrms (1980).

[22] More generally, when probability functions are defined over spaces that admit of uncountable partitions, noncongomerability is extremely difficult to avoid. See Kadane et. al. (1986).

but for each T in **T**, P(R|T) = 2/3.

Suppose that Gretel is told that she will play the following two round game. In the first round, (after the dart has been thrown, but before she has seen where it landed) she will be offered the following bet:

> If the dart landed in R, pay $3. Otherwise receive $2.

Then she will be told which element of **T** the dart landed in. Finally, she will be offered the following bet:

> If the dart landed in R, receive $2. Otherwise pay $3.

She foresees in advance that she will accept both bets. She will accept the first bet because at the time she is offered it, her degree of belief that the dart landed in R will still be 1/3. She will accept the second bet because it the time she is offered it, her degree of belief that the dart landed in R will have increased to 2/3. Furthermore, she foresees that the bets together guarantee a loss of $1.

This scenario is familiar enough from previous discussion: if Gretel cannot bind herself, she will be in an unfortunate situation in which a rational person is helplessly exploited. If can can bind herself, at the start of the game she will commit to declining the second bet and so avoid the forced loss.

Hansel and Gretel face a socialized version of the above problem. Even if they are betting for their mutual benefit, their epistemic perspectives will enjoin bets that collectively will be to their mutual detriment. Assuming that they attach no particular utility to conformity, and that they cannot enter into binding agreements, they will be collectively exploitable by Sir Pinski. Where a capacity for social binding is available, and consultation possible, social contracts between Hansel and Gretel will be individually rational and mutually beneficial.

*

In the above discussion, we have assumed that decision theory applies to Gabriel and Gretel even though their credence functions are non-conglomerable. Some would disagree, and would require rational agents to have conglomerable credence functions. (This requirement is in the same spirit as the requirement that rational agents have bounded utility scales.) We think that the cases of Gabriel and Gretel do not call for such a requirement: as explained above, the cases can be handled using our two main themes. But are there are independent motivations for imposing such a requirement?

One proposed motivation derives from countable additivity. Countable additivity is a consequence of natural-looking continuity principles. And the mathematical development of probability theory goes more smoothly when countable additivity is assumed. Considerations of this kind have convinced some that countable additivity is a rational requirement (Jaynes 2003, Chapter 15). Others have thought that such considerations have little force, and have claimed that countable additivity is no requirement of rationality (de Finetti 1974). In any case, requiring countable additivity rules out some instances of non-conglomerability (those instances in which the relevant partition is countable). So requiring countable additivity goes part way towards requiring conglomerability.[23]

But only part way. For some probability functions—such as Gretel's—are non-conglomerable even though they are countably additive. How might one rule out these remaining instances of non-conglomerability? The most promising way would be to say that a rational agent need not have well-defined degrees of belief conditional on propositions in which the agent has degree of belief zero. This would automatically be the case if one adopted the standard definition of P(A/B) as the quotient P(A&B)/P(B).[24] Only if one takes conditional degrees of belief as primitive (and subject to certain axioms) can one allow for well-defined degrees of belief conditional upon propositions to which one attaches zero degree of belief.[25] If Gretel has no primitive conditional degrees of belief, she will have no well-defined degrees of belief conditional upon the dart landing on a line. And her degrees of belief will therefore be conglomerable.

Someone who rejects primitive conditional degrees of belief is obliged to say how Gretel

[23] We are indebted to an anonymous referee for bringing to our attention the theoretical considerations sometimes adduced for requiring countable additivity.

[24] Hájek (Forthcoming) argues convincingly against adopting that standard definition.

[25] On the axiomatization of primitive conditional probability, see Renyi (1955), Popper (1952), and Hájek (Forthcoming).

should update her beliefs when she learns a proposition to which she'd previously assigned probability zero. We do not think that this will be an easy task. Nor do we find it plausible that Gretel's updating in this case is subject to no constraints at all. (Nor are we eager to simply declare decision theory inapplicable when uncountable infinities are involved.) So this way of ruling out non-conglomerable degrees of belief is unattractive.

Bottom line: one can inoculate against Gabriel and Gretel's troubles by both requiring countable additivity and banning primitive conditional degrees of belief. But the motivation for doing the former is controversial, and the cost of doing the latter is prohibitive.

*

It is often thought that vulnerability to Dutch books is a symptom of irrationality. For example, a famous justification of the finite additivity requirement is that agents who violate it are vulnerable to Dutch books.[26] One might similarly argue that Bill Gates, Gabriel and Gretel are irrational, on the grounds that they are vulnerable to Dutch books. Such arguments—if they worked—would show that boundedness of utility functions, countable additivity and conglomerability are rational requirements. But such arguments do not work. Whatever force Dutch book arguments have in finite cases, they have no force whatsoever in infinite cases.

To see why, consider an agent who is vulnerable to a Dutch book. Here is the best argument for thinking that this vulnerability is a symptom of irrationality: In judging each bet in the book as favorable, the agent is committed to judging the whole book as favorable. But the agent can also deductively infer that the book leads to a sure loss. There is an unacceptable tension between those two judgments.[27]

This argument depends on the following principle:

Bet Agglomeration Principle: Suppose that one is offered a package of bets, and that one judges each bet in the package as favorable (when considered on its own[28]). Then one is

[26] On Dutch book arguments, see Ramsey (1931), de Finetti (1974), and Earman (1982).

[27] Here we follow Christensen (1996) in preferring a nonpragmatic version of the Dutch book argument , and adapt his statement of such a version.

[28] Each bet is B assumed to be favorable in the following sense: no matter what other bets in the

committed to judging the whole package as favorable.

This principle should look familiar—and suspicious! It is the betting analogue of the following principle:

Deal Agglomeration Principle: Suppose that one is offered a package of (deterministic) deals, and that one judges each deal in the package as favorable (when considered on its own[29]). Then one is committed to judging the whole package of deals as favorable.

We have already seen that the Deal Agglomeration Principle is false. It can fail when the relevant package of deals is infinite. Recall that for each n, Eve judges the deal "Take piece #n of the apple" as favorable when considered on its own. But she is *not* thereby committed to judging the package "Take every piece of apple" as favorable.

The failure of the Deal Agglomeration Principle robs the Bet Agglomeration Principle of any plausibility in cases involving infinite packages of bets. Start by accepting that infinitely many favorable deals may combine to form an unfavorable package. Given that, there is every reason to think that infinitely many favorable *bets* may combine to form an unfavorable package.

Bottom line: when it comes to infinite packages of bets, the Bet Agglomeration Principle is false. So the argument above—the best argument that vulnerability to a Dutch book is a symptom of irrationality—does not work for infinite Dutch books. Other arguments along the same lines do no better. There simply need not be any tension between judging each of an infinite package of bets as favorable, and judging the whole package as unfavorable. So one can be perfectly rational even if one is vulnerable to an infinite Dutch book.

4. Reflection

We have suggested that decision theory can allow for agents who have unbounded utilities

package one accepts, the expected gain from accepting those bets along with B is higher than the expected gain from accepting the other bets alone (unless both expected gains are undefined).

[29] Each deal D assumed to be favorable in the following sense: no matter what other deals in the package one accepts, the gain from accepting those deals along with D is greater than the gain from accepting the other deals alone.

and non-conglomerable credences. That suggestion has bizarre consequences for the manner in which one's current beliefs and desires should mesh with one's estimates of one's future beliefs and desires.

Updating by conditionalization in cases in which conglomerability is violated entails a violation of van Fraassen's (1984) "Reflection Principle": the principle that one's current degree of belief in a proposition A should equal one's expectation of one's future degree of belief in A. Suppose that you have non-conglomerable degrees of belief in A relative to a particular partition. And suppose that you are certain that tomorrow you will be told which element of the partition is true. Then you should be certain that your degree of belief in A is about to decrease. That is a counterexample to the Reflection Principle.

Nonconglomerability of one's credence function can lead to other oddities. In certain cases, it allows one to change one's degree of belief in a proposition to whatever value one wishes. For example, Gabriel can force himself have degree of belief 1/5 that God will choose an odd numbered planet. To do so, Gabriel would start by partitioning the natural numbers into sets of five in such a way that each set contains exactly one odd number. Then he would ask God which element of that partition contains the number of the chosen planet. Upon hearing the answer, Gabriel will come to have degree of belief 1/5 that God picked an odd-numbered planet. And for any degree of belief r (where r is a rational number) there exists a corresponding partition. So for any rational-valued degree of belief, Gabriel can make sure that he acquires that degree of belief that God will choose an odd-numbered planet.

In the case of unbounded utilities, one cannot choose what to *believe*. But one can choose what to options to prefer. For instance, in the case of the two tickets, suppose that one can choose whether one will be told the value of the ticket that one possesses or the value of the other ticket. By so doing, one can control whether one will prefer to stick with one's ticket, or to switch tickets. It's a strange world. Here we have a failure of the "Preference Reflection Principle": that the value that one currently attaches to an option should match one's expectation of one's future value of that option.

Some of us have already been convinced that cases of known future irrationality or information loss make trouble for Reflection principles.[31] The cases described above are much more shocking: they arise even when one is sure that one's future beliefs will be gotten from one's present beliefs by ordinary conditionalization. It is counterintuitive that rational agents can be subject to these more shocking Reflection violations. Our view—that reasonable agents may have non-conglomerable probabilities—entails this counterintuitive consequence. That is a cost of our view.[32] A benefit of our view is that it allows decision theory to apply to agents whose uncertainty ranges over a continuous space of outcomes. As the case of Gretel shows, competing views will be hard-pressed to do the same.

5. Conclusion

Problems raised by unbounded utilities, countable additivity violations, and probabilities conditional upon probability-zero propositions have a common character. The common character is that a collection of deals can be individually attractive but jointly harmful. When an agent can decide all at once on which deals to accept, her options consist not of individual deals, but rather of complete patterns of which deals to accept and which to reject. The agent's choice in such cases should be guided by her preferences over these various patterns.

When the agent is offered the deals sequentially, and cannot bind herself to a plan of action, the above resolution does not apply. In that case the agent should accept or reject each deal according to the (causal) expected utility of doing so. It may well be that the sequence of decisions endorsed by that rule lead to the agent's ruin. That is, it may be that the agent's rationality (combined with her inability to commit to a plan) prevent her from coordinating her actions to prevent her ruin. Similar situations arise when agents need to coordinate not with their future selves, but with each other.

We have suggested that to handle these puzzles we need not ban unbounded utilities and nonconglomerable probabilities, nor need we institute countable additivity as a constraint upon credences. Instead we recommend following the argument where it leads. Certain rational agents

[31] For trouble cases for Reflection arising from cognitive mishaps, see Skyrms (1987) and Christensen (1991); for cases involving information loss see Talbott (1991), Elga (2000), and Arntzenius (2003).

[32] We are indebted to an anonymous referee for pointing this out.

will, owing to the protocol of the case, be led to ruin without being irrational thereby. In such cases, we have self-consciously deployed that language used by Gibbard and Harper (1978) in analyzing Newcomb's problem. In these cases too, rational agents are frequently punished, and irrational ones rewarded. Which rational agents will be punished? Those who do not value conformity to their earlier selves and each other as ends in themselves, and who cannot causally bind their future selves to their past selves or themselves to each other. The situation is very different for agents who can do the relevant kinds of binding. In such cases, binding can forseeably bring dividends. We thus all have reason to hope that we are agents who can self-bind, and who can bind ourselves to each other.

References

Arntzenius, F. and D. McCarthy.1997. The two envelope paradox and infinite expectations *Analysis* 57(1):28-34.

Arntzenius, F. and J.Barrett. 1999. An infinite decision puzzle. *Theory and Decision* 46:101-103.

Arntzenius, F. 2003. Some problems for conditionalization and reflection. *Journal of Philosophy* 100(7): 356-371.

Broome, J. 1995. The two envelope paradox. *Analysis* 55(1): 6-11.

Christensen, D. 1996. Dutch Books Depragmatized: Epistemic Consistency for Partial Believers. *Journal of Philosophy* 93:450-479.

de Finetti, B. 1974. *The Theory of Probability*. New York: Wiley.

Earman, J. 1992. *Bayes or Bust?* Cambridge: MIT Press.

Earman, J., and J. D. Norton. 1996. Infinite pains: the trouble with supertasks. In *Benacderraf and his Critics.* Eds. A. Morton and S. Stich, 231-61, Cambridge, MA: Blackwell.

Elga, Adam. 1997. Failures of Dominance. Manuscript.

Gauthier, D. 1997. Resolute Choice and Rational Deliberation: A Critique and a Defense. *Nous* 31:1-25.

Gibbard, A., and W. Harper. 1978. Counterfactuals and two kinds of expected utility. In *Foundations and Applications of Decision Theory, vol. 1*. Ed. Hooker, Leach, and McClennen, 125-162. Dordrecht: Reidel.

Gracely, E. 1988. Playing Games with Eternity: The Devil's Offer. *Analysis* 48(3): 113.

Hájek, A. Forthcoming. What Conditional Probability Could Not Be. *Synthese.*

Jaynes, E. T. 2003. *Probability Theory: The Logic of Science*, ed. G. Larry Bretthorst. Cambridge: Cambridge University Press.

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

Lewis, D. 1980. A subjectivist's guide to objective chance. In *Studies in Inductive Logic and Probability*, Vol. II, ed. R. C. Jeffrey. University of California Press.

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

Kadane, J., M. Schervish and T. Seidenfeld. 1986. Statistical implications of finitely additive probability. In *Bayesian Inference and Decision Techniques, volume 6*. Ed. P. Goel and A. Zellner, 59-76. Amsterdam: Elsevier Science Publishers.

Landesman, Cliff. 1995. When to terminate a charitable trust? *Analysis* 55(1): 12-13.

McGee, V. 1999. An airtight Dutch book. *Analysis* 59(4):257-265.

Norton, J. 1998. When the sum of our expectations fails us: The exchange paradox. *Pacific Philosophical Quarterly* 79:34- 58.

Pitowski, I. 1983. Deterministic model of spin and statistics. *Physical Review D* 27(10):2316-2326.

Pollock, John L. 1983. How do you Maximize Expectation Value? *Nous* 17(3):409-421.

Popper, K. 1952. *The Logic of Scientific Discovery*, 2nd edition. New York: Basic Books.

Ramsey, F.P. 1931. *The Foundations of Mathematics and other Logical Essays*. London: Routledge.

Renyi, A. 1955. On a new axiomatic theory of probability. *Acta Math. Acad. Scient. Hungaricae* 6: 285-335.

Seidenfeld, T. and M. Schervish. 1983. A conflict between finite additivity and avoiding dutch book. *Philosophy of Science* 50:398-412.

Stalnaker, R. 1999. Extensive and strategic forms: Games and models for games. *Research in Economics* 53: 293-319.

Skyrms, B. 1980. *Causal Necessity*. New Haven: Yale University Press.

van Fraassen, B. 1995. Belief and the problem of Ulysses and the sirens. *Philosophical Studies* **77**: 7-37.

van Fraassen, B. 1984. Belief and the Will. *Journal of Philosophy* 81:235-256.