How to Avoid Maximizing Expected Utility

Bradley Monton
Wuhan University

Abstract

The lesson to be learned from the paradoxical St. Petersburg game and Pascal's Mugging is that there are situations where expected utility maximizers will needlessly end up (with high probability) poor and on death's door, and hence we should not be expected utility maximizers. Instead, when it comes to decision-making, for possibilities that have very small probabilities of occurring, we should discount those probabilities down to zero, regardless of the utilities associated with those possibilities.

1. Introduction

According to orthodox decision theory, one should always maximize expected utility. For example, here's a standard statement of the view: orthodox decision theory

> specifies that decision makers should maximize expected utility – period. It says that it is wrong always, everywhere, and for anyone, to do otherwise. (Hájek 2014, 555-6)

But this procedure of maximizing expected utility (period) is sometimes an unreasonable – perhaps even irrational – decision procedure to follow. This (I maintain) is something that we (collectively) have known (or should have known) for about 300 years (going back to the St. Petersburg game, which I'll discuss shortly). But what is not generally known – at least, what I haven't known, until recently – is how best to rationally avoid maximizing expected utility (period).

In this paper, I'll present and defend the solution: when it comes to decision-making, for possibilities that have very small probabilities of occurring, we should discount those probabilities down to zero, regardless of the utilities associated with those possibilities. Only after doing that should we maximize expected utility. (That's the core proposal of this paper that is the least deviating from orthodox decision theory. More aggressive deviations will be hinted at later.)

It turns out that this discounting method has been on the table for about 300 years as well – but despite that, discussions of why this method is

the right one have been limited. I'll critically evaluate six different justifications of this method.

In addition to the question of why this discounting method is the right one to follow, there are open questions about the actual content of the method, and 300 years hasn't been enough time to resolve them. I'll go on to give answers to three such open questions. Then I'll talk about dominance reasoning, and respond to Pascal's Wager. And stick around for the conclusion, with a big-picture thought about decision theory.

2. The St. Petersburg Game

In 1713, Nicolaus (I) Bernoulli (who I'll call "Nicolaus" for short) wrote a letter presenting the St. Petersburg game; in 1738 his cousin Daniel Bernoulli first published the game.[1] The game is played by flipping a fair coin until it lands heads, and then the game stops. The prize is $\$2^n$, where n is the number of flips.[2] The game is considered paradoxical because the expected value of the game is infinite, but intuitively people aren't willing to pay that much to play – in fact, it is intuitively irrational to give the majority of one's wealth to play a game that only has a tiny chance of yielding significant reward.

Some people are tempted to reject that intuition. Harris Nover and Alan Hájek (2004, 309) write:

> in response to the St. Petersburg paradox, there is something to be said for the bullet-biting response: 'the game *should* be valued infinitely, and any intuition to the contrary should be dismissed as an artifact of our finite minds not fully appreciating the true nature of the game; we should learn to live with decision theory's verdict'.

I maintain that this bullet-biting response is mistaken. Let's suppose that our finite minds cannot appreciate the true nature of infinite games. Admitting that doesn't resolve the paradox, because there is a finite version of the St. Petersburg game that is also paradoxical. The paradox arises in the finite

---

[1] Robert Martin (2017) says that Nicolaus and Daniel are brothers, but this is (as far as I can tell) mistaken; Dutka (1988, 18) says that they are cousins, and explains the family relationships carefully. Spieß (1975, 563) also says that they are cousins.

[2] In Daniel Bernoulli's (1738) version of the game, the price is $2^n/2$ ducats, but the formulation I give is the standard contemporary one. Daniel Bernoulli says that he would pay 20 ducats to play the St. Petersburg game; presumably he would pay $40 to play the standard contemporary version.

case because the expected value of the game is high, while rational people are not willing to pay more than a low amount to play.

Suppose that we truncate the St. Petersburg game at 999 tosses – if one gets tails 999 times in a row, then the game ends and the prize is what one would get if heads landed on the 1000th toss ($\$2^{1000}$). The expected value of this game is $1000. But rational people aren't willing to pay more than about $100 to play the original St. Petersburg game, and this finite version is worth even less, so again there is a disconnect between what rational people are willing to pay and what the strategy of maximizing expected utility recommends.

Daniel Bernoulli's (1738) now-standard (and correct) analysis of the original St. Petersburg game is that money has diminishing marginal utility, so while the expected monetary *value* of the finite game may be $1000, the expected *utility* of the game is much less. But we can just present new versions of the St. Petersburg game, replacing monetary values with utilities.

My favorite version is a days-of-life version, because I do not treat life-days as having diminishing marginal utility. I value two days of living twice as much as I value one day of living, I value 10,000 days of living 10 times as much as I value 1,000 days of living, and so on.[3] Suppose that you discover that, if you do nothing, you will die on The Critical Day, which is currently 1000 days from now. But you are offered a pill that with probability 1/2 will extend your day of death past The Critical Day by exactly 2 days, with probability 1/4 will extend your day of death past The Critical Day by exactly 4 days, and so on. (For a finite version of this game we can cap it at 2^{1000} days.) This pill has a cost though – it brings The Critical Day closer to the present by X number of days. How large would X have to be before you are unwilling to take the pill?

Since I do not treat life-days as having diminishing marginal utility, according to orthodox decision theory my expected utility of taking the pill (ignoring the cost) is an infinite number of life days (for the uncapped version) or 1000 life-days (for the capped version). But for either version, I am not willing to take the pill if the cost is moving The Critical Day 999 days closer to the present – I wouldn't take the pill if X were more than 50.

---

[3] Some interlocutors have objected that as a practical matter I couldn't have such insight into the details of my own utility function. I think that I do, but instead of arguing that point, we can just set me aside: imagine a hypothetical agent who does treat life-days in the way I'm describing, and consider how you would advise such a hypothetical agent who is presented with the days-of-life version of the St. Petersburg game that I will describe.

Why not? If I take the pill when X=999, there's a 7/8 chance that I'll live for less than 10 more days, and that would be sad. If I don't take the pill I'll live for 1000 more days – much better! But orthodox decision theory says that I should take the X=999 pill, because the expected utility of taking the pill is greater than the expected utility of not taking it. I am not willing to maximize expected utility here, and I maintain that maximizing expected utility here is an unreasonable thing to do.[4]

Now is a good time to make explicit that the scope of my paper is not limited to non-ideal boundedly rational agents like you and me; I intend my reasoning to also apply to imaginary ideally rational agents. My rejection of the orthodox decision-theoretic answer to the finite days-of-life version of the St. Petersburg game has nothing to do with the fact that I am bounded in my rationality – it's not the case that, if only I were more rational or could do some more sophisticated math, I could see that taking the X=999 pill is the right choice. For an agent who shares my utilities about days of life, taking the X=999 pill is so unreasonable that I take taking that pill as a datum in determining what counts as irrational behavior.

Note that there's not an unreasonable amount of idealization in the setup of the finite days-of-life version of the St. Petersburg game. Hájek points out that for the original infinite version of the St. Petersburg game, we are "up to our necks in idealization" (Hájek 2014, 565), and he gives that as a reason to accept the prima facie surprising orthodox-decision-theoretic claim that the game should be valued infinitely. But the finite days-of-life St. Petersburg game isn't overly idealized like that – while we may not currently have the medical resources to make such pills, we are still operating within the general constraints of the actual world (we are not dealing with infinite bankrolls or infinite sequences of coin flips, for example).[5] So when orthodox decision theory gives an unreasonable

---

[4] One might object that there is a disutility associated with the feeling of sadness, and that needs to be taken into account in deciding whether to take the X=999 pill. I agree, but for me at least, the disutility associated with such a feeling of sadness is very small compared to the full amount of utility associated with each day of life, so such feelings do not play a significant role in my expected utility calculation. Those who follow the adage "don't cry over spilt milk" should be on board with not assigning much disutility to these feelings of sadness or regret at the game not going the way one wanted.

[5] Anti-aging and life-extension technologies are actively under development by reputable scientists, but immunity from accidental death is (alas) not. So to make the game more realistic, we could add utility that compensates for the chance that (via accidental death) one might not live that long. We could do that by, for example, offering to give the

[continued on next page]

prescription for the finite days-of-life version of the St. Petersburg game (the prescription that one should take the X=999 pill), we shouldn't defend orthodox decision theory by saying that the prescription is permissible because we're talking about a highly idealized impossible situation. We're talking about a situation within the realm of future possible development, and orthodox decision theory gives the wrong answer, so orthodox decision theory has to go.

3. Pascal's Mugging

These St. Petersburg-style games are effective in illustrating a problem with orthodox decision theory, but the games are more complicated than they need to be to illustrate the problem. The relevant feature of the St. Petersburg-style games is that there is a very low probability of getting a very high reward. For a simpler game with that feature, consider Pascal's Mugging (Bostrom 2009).

Pascal's Mugging is a game where an agent named 'Pascal' has an opportunity to pay some money for a very small probability of getting a very large (but finite) amount of utility from a seeming mugger – utility both for Pascal and for lots of sweet little orphans. (In Bostrom's scenario the probability is one in 10 quadrillion.) Since the utility associated with the mugger's offer is high enough, Pascal calculates that the expected utility of giving the money is positive. No matter how low of a non-zero probability Pascal assigns to the hypothesis that he and the orphans will get the large amount of utility, there is a corresponding utility that he and the orphans could be offered such that Pascal deems the expected utility of the game to be positive, and hence gives the mugger the money.

Those who maximize expected utility will most likely end up with undesirable results when presented with the situations above. They will spend a lot of their wealth to play the St. Petersburg game, or give it to the grandiosely-promising mugger – and they can foresee that, based on their own probability judgements, it's very likely that they will have nothing to show for it. Expected utility maximizers, when faced with 1000 more days of life, will take a pill that will cause them to most likely end up with just a

---

person more and more money the longer they live.

Some interlocutors have been skeptical of the possibility of a pill accomplishing all that I describe. But the pill is just an artifice for simplicity – a more realistic account involves a game-playing, coin-flipping league of assassins and life-extension scientists.

few more days of life. In short, in these situations, expected utility maximizers are needlessly (with high probability) poor and on death's door, and I do not want to be one.

In the next section, I'll discuss how not to be such an expected utility maximizer. But first: my arguments in this paper aren't just a rejection of expected utility maximization; they're also a rejection of consequentialism. Consequentialists hold that "What we ought subjectively to do is the act whose outcome has the greatest expected goodness" (Parfit 1985, 25). That formulation mirrors the orthodox-decision-theoretic prescription to choose that act that maximizes expected utility, and as result everything I say about the perils of orthodox decision theory has a consequentialist analogue. For example, a consequentialist version of Pascal's Mugging could go as follows: a strange person hands you a baby and asks you to torture it, telling you that the torture will prevent significant unjust suffering of a large number of sentient creatures in some distant galaxy. While the probability is very low that torturing the baby will prevent the suffering, as long as the strange person makes the number of claimed distant-galaxy creatures high enough, according to consequentialism you should torture the baby. That's not good. But henceforth, I'll focus on decision theory; I'll leave the anti-consequentialist analogues of my arguments as an exercise for the interested ethicist reader.

Before moving on to the next section, where I discuss how not to be an expected utility maximizer, let's consider an objection to my St. Petersburg- and Pascal's Mugging-based rejection of orthodox decision theory. Some might admit that in those situations, expected utility maximizers would most likely end up poor and on death's door, but maintain that it doesn't follow that they are irrational. The objectors might say that the situation is analogous to what the two-boxers face in Newcomb's problem. In Newcomb's problem, a predictor puts $1 million in a box for you to take if and only if the predictor predicts that you will not also take a box with $1000. According to two-boxers like David Lewis (1981), the rational thing to do is to take both boxes, even though you will most likely not get the $1 million as a result: "The reason why we [two-boxers] are not rich is that the riches were reserved for the irrational" (Lewis 1981, 377). Could the expected utility maximizer make a similar move?

My answer is "no"; the situations are importantly disanalogous. In Newcomb's problem, the reason the two-boxers aren't rich is that there is an agent, the predictor, who is rewarding irrationality. As two-boxers Gibbard and Harper (1978, 153) put it, "If someone is very good at predicting behavior and rewards predicted irrationality richly, then irrationality will be

richly rewarded."[6] Even if one replaced the predictor in Newcomb's problem with an inanimate predicting machine, it is still the case that the person playing the game is being analyzed and (from the two-boxer's perspective, at least) the fact that the two-boxer has a rational intention-forming mechanism leads to the two-boxer being penalized. For people presented with a St. Petersburg-style game or Pascal's Mugging, in contrast, there is no evaluation of their mental state by another agent or a machine. As a result, there is nothing external that penalizes the people for the decisions they make – when expected utility maximizers end up poor and on death's door, they've done it to themselves.

4. Discounting Very Small Probabilities

So how can we avoid the fate of the expected utility maximizers? The simplicity of Pascal's Mugging makes clear that there are only two factors that are put in to yield the problematic result: the high-utility possibility and the small probability associated with that possibility. So to avoid the fate of the expected utility maximizers, we need to either limit that high utility number or discount that small probability number.[7]

The limiting utility move is not viable. A utility cap would be needed, but such a cap is ethically problematic. Here's why a cap would be needed: if one instead endorsed a principle that slowed the growth of utility without imposing a cap, then I could simply present a new version of the St. Petersburg game or Pascal's Mugging that offered utilities high enough to overcome the slowing-growth principle. But a cap is ethically problematic because one can always add more agents into the utility calculation – one can always consider the possibility of more sweet little orphans (either in existence now or in the future) who can benefit from one's decisions. I maintain, and so should you, that the utilities of those added sweet little orphans matter, regardless of how many sweet little orphans are already in existence.[8]

---

[6] The original quote has a semi-colon in place of the comma; I read that as a typo.

[7] At least, those are the two options under the assumption that preferences of an agent should be modeled via a utility function and a probability function; I'll say more about that in the conclusion.

[8] Moreover, I maintain, and so should you, that the utility of an added sweet little orphan matters just as much as that of the already-considered sweet little orphans – it's ethically problematic to say that adding more and more sweet little orphans contributes smaller and smaller amounts of utility, even though that would achieve the mathematical result of [continued on next page]

So limiting the high utility number isn't viable. This leaves the only other option: when it comes to decision-making, one should discount small probabilities, down to zero.[9]

Here's why going to zero is needed: if one instead had a principle that lowered the probability numbers without reducing them to zero, I could simply present a new game that offered utilities high enough to compensate for those reduced probabilities.[10]

Here's why the "when it comes to decision-making" qualifier is in there. When it comes to one's doxastic state, and updating one's doxastic state in response to new evidence, one should not discount small probabilities down to zero. The reason is that, in the Bayesian framework at least, a proposition that receives probability zero will always receive probability zero; an agent who assigns zero probability to a hypothesis is insensitive to new evidence regarding that hypothesis. But one should be sensitive to new evidence regarding propositions to which one initially assigns a small but non-zero probability, so when it comes to updating, one should not discount such probabilities down to zero.

My proposal is that for decision-making, small probabilities should be discounted down to zero before maximizing expected utility … but I know you're itching to ask: "how small?" My answer for now is "very small". Don't hate me! I'll have much more to say about this below.

This discounting idea is not original with me – in fact, it originated with the person who came up with the St. Petersburg paradox, Nicolaus (I)

---

having the utility asymptotically approach a cap.

What if there are already an infinite number of sweet little orphans in existence? Even then I maintain that adding more ones matters, as I argue in Monton 2010, Section 6.

[9] There are also mixed options: instead of just limiting the utilities, or just discounting the probabilities, one could combine the moves – one could hold that high-utility and high-disutililty possibilities need discounts on their probabilities, and high-probability possibilities need limits on their (positive or negative) utilities. I'm not a fan of mixed options; there are situations where they give unreasonable prescriptions. Imagine, for example, that you seem to experience death and then show up in a wonderful afterlife, and hence have you good reason to assign a high probability to the high-utility hypothesis that heaven exists and you are in it. It strikes me as unreasonable to discount the probability of that hypothesis just because that hypothesis has high utility – when all the evidence points to heaven existing and you being in it, it's unreasonable to say "this is too good to be true". (Of course, there are many reasons to doubt one's perceptions, but high utility by itself is not a good one.)

[10] For a mathematical example of this, see Buchak 2014, 73.

Bernoulli himself.[11] In 1714, Nicolaus, in a letter to Pierre Rémond de Montmort, endorses the following solution to his St. Petersburg paradox: "cases that have a very small probability must be neglected and assumed to be zero"[12], and as a result he says that the expected utility assigned to the St. Petersburg game by orthodox reasoning is too high. Nicolaus, unfortunately, doesn't have much more to say about his solution, but he does say "This is a remark which deserves to be thoroughly studied."[13] That thorough study has not yet happened; this paper is my belated attempt to rectify that.

I'll henceforth neologistically call that discounting-to-zero procedure "Nicolausian discounting". I endorse the doctrine that one should engage in Nicolausian discounting. But why, when it comes to decision-making, should very small probabilities be discounted down to zero? There's a sense in which I've already given an answer: in the games presented in the previous section, we rational agents don't want to just maximize expected utility, and capping utilities is a bad option, so discounting probabilities must be the good option. But I recognize that more would need to be said – and indeed, here and there in the literature, some more has been said. Building on those limited discussions in the literature, I'll now present and critically evaluate six reasons for engaging in Nicolausian discounting, in order from my least favorite to my most.

5. Six Reasons for Discounting

5.1 D'Alembert and "Certainement"

Jean Le Rond d'Alembert (1761) was the first to propose in print discounting very small probabilities down to zero.[14] D'Alembert endorses

---

[11] Interestingly, none of the people I discuss in this paper who endorse something akin to Nicolaus's discounting move (d'Alembert, Daniel Bernoulli, Buffon, Smith, Schwitzgebel, Buchak, Borel, Jordan, and Condorcet) cites Nicolaus as being the originator of this idea. In fact, the three contemporary philosophers I focus on in the next section (Smith, Schwitzgebel, and Buchak) do not cite anyone as being a precursor; they each introduce the idea *de novo*. (I don't mean this as a critique – it's just interesting.)

[12] This is my translation; the original is "les cas qui ont une tres petite probabilité doivent être negligés et censés pour nuls". Nicolaus's letter is published in Spieß 1975, 558-9.

[13] This is my translation; the original is "C'est une remarque qui merite d'étre bien approfondie."

[14] The historian of probability Isaac Todhunter (1865, section 644, page 344) incorrectly writes: "The doctrine that a very small chance is practically zero is due to D'Alembert". But we can give d'Alembert credit for being the first to propose the idea in print.

the following rule: "when the probability is very small, it must, in the ordinary course of life, be regarded as zero, & treated as such."[15] His initial rationale for this looks promising. He considers the standard rule of valuing a game by calculating the expected value (multiplying the probabilities and the utilities, and summing), and he says that the St. Petersburg game shows that this rule is flawed.[16] If instead one follows the rule of discounting very small probabilities down to zero and then valuing the St. Petersburg game via its expected value, one gets a more reasonable valuation for the game. (At least, one does if one picks a suitable threshold for what probabilities count as very small – more on this later.) So d'Alembert says that the fact that his discounting rule enables one to assign a reasonable value to the St. Petersburg game is evidence in favor of his rule.

So far, I agree. But d'Alembert recognizes that he needs to give more justification for his rule than that it gives a reasonable value to the St. Petersburg game, and it is in this further justification that he makes some problematic moves. D'Alembert considers a simplified version of the St. Petersburg game – essentially a Pascal's Mugging scenario. Pierre has the opportunity to pay 1 écu to play a game where if a coin lands heads for the first time on the hundredth toss, Pierre will get 2^{100} écus, and otherwise he'll get nothing. By the orthodox way of calculating the value of that game in terms of the expected value, the game is fair, but d'Albembert says that Pierre should not play. Why not?

There's an interesting translation issue here, so let's start with the French:

> Pierre ne doit pas donner cet écu; parce qu'il le perdra *certainement*, & que *croix* arrivera *certainement* avant le

---

[15] D'Alembert 1761, Memoir 10, Section 13, page 11, my translation. The original is: "On peut donc, ce me semble, poser pour régle, que quand la probabilité est fort petite, on doit dans l'usage ordinaire de la vie, la regarder comme zéro, & la traiter comme telle."

[16] D'Alembert (1761), Memoir 10, Section 2, page 2. He imagines the St. Petersburg game offered to Pierre, and says the standard way of valuing games has the consequence that Pierre must give an infinite sum for the game to be fair. He then writes: "independent of the fact that an *infinite sum* is a chimera, there is no person who would give it in order to play this game, I say no infinite sum, nor even a rather modest sum. The rule [of valuing a game via its expected value] appears therefore to be flawed". This is my translation; the original is: "indépendamment de ce qu'une *somme infinie* est une chimere, il n'y a personne qui voulût donner pour jouer à ce jeu, je ne dis pas une somme infinie, mais même une somme assez modique. La régle paroît donc être en défaut".

centiéme coup, bien qu'il ne doive pas arriver *nécessairement*.
(d'Alembert 1761, Memoir 10, Section 10, pages 8-9)

You might think that this should be translated as something like:

Pierre must not give this écu, because he will *certainly* lose it, & heads will *certainly* arrive before the hundredth toss, even though it will not *necessarily* arrive.

And indeed, this is the sort of way it is standardly translated.[17]

The problem with this translation is that it has the consequence that what d'Alembert is saying doesn't make sense – at least, in this context, I don't see how to understand this distinction between the claim "X will certainly happen" and the claim "X will necessarily happen". We should wonder whether that translation is uncharitable – and indeed, there is another way to translate "certainement" besides as "certainly". It can also mean "most probably" or even "probably".[18] On that more charitable interpretation of d'Alembert, we have him saying:

Pierre must not give this écu, because he will *most probably* lose it, and heads will *most probably* arrive before the hundredth toss, even though it will not *necessarily* arrive.

It's true that most probably, Pierre won't get 100 tails in a row. The problem now is that we lose the motivation for d'Alembert to say that Pierre should not give the écu. It's true that he most probably will lose it, but he might get a huge reward. What's so bad about that?

The *certainement*/*nécessairement* distinction isn't the only one d'Alembert uses to try to elucidate his point. He also says that getting 100 tails in a row is not "possible *physiquement*", even though it is "*métaphysiquement* possible" (d'Alembert 1761, Memoir 10, Section 12,

---

[17] See Todhunter (1865, section 472, page 262), Samuelson (1977, 43), and Pulskamp (2009, 4).

[18] For "most probably" see *Collins French Dictionary*, https://www.collinsdictionary.com/dictionary/french-english/certainement, archived on 2017-6-22 at http://www.webcitation.org/6rPE8Vq9m; and http://www.french-linguistics.co.uk/dictionary/certainement.html, archived on 2017-6-22 at http://www.webcitation.org/6rPFZ2Upb. For "probably" see *Cambridge Dictionary*, http://dictionary.cambridge.org/dictionary/french-english/certainement, archived on 2017-6-22 at http://www.webcitation.org/6rPEaLAsm.

D'Alembert's contemporary Buffon also used the term "certainement" to describe a measure of probability that comes in degrees – see Hey, Neugebauer, and Pasca (2010, footnote 2): "we should warn readers at the outset that Buffon uses the word certainty not in the absolute sense that we currently use. … He refers to certainties of different orders and hence uses the word certainty as a synonym for probability or likelihood."

page 10). The distinction between physical possibility and metaphysical possibility makes sense in this context; the problem is that the claim that it's physically impossible to get 100 tails in a row seems false. On the contemporary standard way of understanding physical possibility, at least, getting 100 tails in a row is physically possible – in fact, that string of results has the same probability as any other string of results one gets from flipping a fair coin 100 times.

But let's again be charitable to d'Alembert – he's working with a different understanding of physical possibility. He clarifies this when he considers rolling double-sixes on two dice 100 times in a row, and says that such a sequence of events is physically impossible "because this has never happened, and will never happen."[19]

This claim is interestingly non-modal. D'Alembert is not saying that such a sequence of events *could* not happen; he's just saying that it has not happened in the past, and it will not happen in the future. His claim may be correct.[20] But suppose it turns out that just moments ago, Lily flipped a fair coin and got 100 tails in a row – d'Alembert's reasoning about the St. Petersburg game would no longer go through. This is unreasonable – we shouldn't have to change our philosophical analysis of the St. Petersburg game as a result of what Lily just did with her coin.

Let's sum up. D'Alembert infamously got lots of things wrong when it comes to probability theory (he endorsed the gambler's fallacy, for example). And I'm not particularly impressed by the reasons he gives for this rule of discounting very small probabilities down to zero. But I think he's right to endorse the rule, even if he hasn't hit on the right reasons. Let's see if his contemporary, Georges-Louis Leclerc de Buffon, can do any better.

5.2 Buffon and the Probability of Dying in a Day

Buffon (1777, Section 16) considers the St. Petersburg game, and says that there are two reasons the game should not be valued highly. One reason is the diminishing marginal utility of money; the other is: "the probability must

---

[19] D'Alembert 1761, Memoir 10, Section 12, page 10; the original is "parce que cela n'est jamais arrivé, & n'arrivera jamais."

[20] I intend this to, in part, be an explanation of d'Alembert's reasoning in response to Paul Samuelson (1977, 43), who disdainfully quotes d'Alembert with parenthetical interrogation marks, as follows: "it is *physically* impossible because it has never [?] happened and never will [??] happen."

be considered as zero as soon as it is very small, that is, below 1/10,000".[21] Setting aside the issue of diminishing marginal utility, Buffon's 1/10,000 threshold would lead one to value the St. Petersburg game with prize $\$2^n$ at \$13, since the probability of getting heads on the 13th toss is 1/8,192, while the probability of getting heads on the 14th toss is 1/16,384. So Buffon's discounting strategy gives a reasonable result to the St. Petersburg game, and that provides some evidence in its favor, but what further reason does Buffon give for his discounting strategy? And why the 1/10,000 threshold?

Buffon's reasoning starts by considering what sorts of possibilities people ignore, and what the associated probabilities are for those possibilities. For an possibility one doesn't care about that much, one might be willing to ignore it even if the probability is not very low. So Buffon considers an important possibility that one cares a lot about: death. But he is not interested in just any person's fear of death – he's interested in the fear of death in someone "when reason has attained its full maturity and the experience all its force", and he considers a 56-year-old man to be such a person (Buffon 1777, Section 8). Consulting mortality tables, he sees that the probability of a 56-year-old man dying in the course of 24 hours is 1/10,189. But such a man who is doing well "has no fear of death in the 24 hours". Buffon then says:

> from this I conclude that any equal or smaller probability must be regarded as zero, since any fear or any hope below ten thousand must not affect us or even occupy for a single moment the heart or the mind. (Buffon 1777, Section 8)

Since such a man is willing to treat his death in a day as a zero-probability possibility, we should treat any probability below 1/10,000 as zero. (Buffon calls the sort of certainty such a man has that he will not die in a day "moral certainty", and distinguishes it from the stronger notion of "physical certainty".)

Daniel Bernoulli, intriguingly, endorses Buffon's discounting strategy, though Bernoulli takes issue with the choice of threshold. Buffon quotes from a 1762 letter that Bernoulli sent him, where Bernoulli writes:

> I strongly approve, Sir, your way to estimate the limits of moral probabilities; you consult the nature of man by his actions … you conclude that one ten-thousandth of probability should not

[21] All quotations in this subsection on Buffon are from the excellent translation by Hey, Neugebauer, and Pasca (2010).

> make any impression in the mind of man, and consequently this one ten-thousandth has to be regarded as an absolute nothing. (quoted in Buffon 1777, Section 8, footnote 3)

Bernoulli then says that some 56-year-old men are ill, and such men would rightly fear dying in the next day. Bernoulli points out that only healthy men would regard the probability of their death in the next day as an absolute nothing. Since the mortality table for 56-year-old men includes both ill men and healthy men, the probability of death of a healthy man is much less than 1/10,000. Bernoulli concludes: "I do not fight your principle, but it seems rather to lead to 1/100,000 rather than to 1/10,000." (quoted in Buffon 1777, Section 8, footnote 3)

After quoting this letter, Buffon makes a slight concession: "there is maybe a little more than ten thousand to bet against one, that a man, healthy and strong will not die in the twenty-four hours; but it is hardly necessary to increase this probability to one hundred thousand." Buffon then says that "this difference [between a 1/10,000 threshold and a 1/100,000 threshold], although very large, changes nothing of the main implications that I draw from my principle." And indeed, I agree. With the 1/100,000 threshold, the value the St. Petersburg game with prize $\$2^n$ would be \$16; so that's also a reasonable result. But what about the reasoning that leads to either of these thresholds?

Well, there is certainly a lot to criticize in the details. First, people other than 56-year-old men have the full maturity of reason and the full force of experience – though perhaps a 21st-century Buffon would argue that, whatever a healthy person's mortality rate is in a day, that determines what that person's probability-discounting threshold is.

Second, many even healthy people do rationally fear dying in the next 24 hours – for many healthy people, there are moments where the fear of immediate death occupies their hearts and their minds. Perhaps Buffon would maintain that that only happens in frightening situations, where the probability of death in a day is greater than 1/10,000. But even in everyday life situations, one can arguably worry about dying soon without thereby being irrational.

Third, why the restriction to a day? Why not two days, or 17 seconds? Buffon doesn't give a reason for that arbitrary time frame.

Fourth, Buffon picks a person's own death because he says that "of all the possible moral probabilities, the one that most affects man in general is the fear of death", but that is arguably false. One could care more about the death of one's family than one's own death. Pandemic researchers try to reduce the probability of human civilization ending, and while that has a

lower probability than the death of a 56-year-old man in a day, it's a correspondingly more significant event.

But despite my criticisms of the details, there's something to be said for the general idea that we should look to rational-seeming people's behavior to adjudicate between decision theories. There are very-small-probability possibilities that rational-seeming people ignore, and we want our decision theory to capture that. One could object that that the *ideally* rational agent should take all such possibilities into account, in the standard-decision-theoretic way – but this would lead to the ideally rational agent giving up many days of life to play the days-of-life version of the St. Petersburg game, and handing over the money in a Pascal's mugging scenario. Those are not reasonable results.

So let's continue to look for reasons to follow the "Nicolausian discounting" rule I'm endorsing: when it comes to decision-making, for possibilities that have very small probabilities of occurring, we should discount those probabilities down to zero, regardless of the utilities associated with those possibilities.

5.3 The Virtues (and Vices) of Tolerance

Nicolausian discounting enables one to assign a reasonable value to the St. Petersburg game. It also enables one to assign a reasonable value to the Pasadena game, a game invented by Nover and Hájek (2004), that is like the St. Petersburg game, but with the distinctive feature that the possible outcomes of the game switch in value between negative and positive, depending on whether the game ends with an odd or even toss.[22] Nover and Hájek show that the Pasadena game has the problematic feature that it has no unique expected utility, because there is no unique value for the sum of the probability times the utility for each of the infinite possible outcomes of the game.

---

[22] My claim that Nicolausian discounting allows one to assign a reasonable value to the Pasadena game is controversial. Kenny Easwaran (2008) argues that the correct value of the Pasadena game is $\ln 2$, and Hájek (2014, page 551, footnote 12) proves that a discounting strategy like mine cannot assign the Pasadena game that irrational value. But I don't know of anyone who believes that the case for valuing the Pasadena game at $\ln 2$ is conclusive – for example, Easwaran (2008, 634) suggests that his reasoning has more of a "pragmatic" than a "rational" basis. Moreover, J. McKenzie Alexander (2011) gives a promising criticism of Easwaran's valuation, that is in the spirit of my reasoning in this paper.

Nicholas J. J. Smith (2014) wants to solve the problems for orthodox decision theory posed by the Pasadena game, and to do so he endorses the following principle:

> For any lottery featuring in any decision problem faced by any agent, there is an $e > 0$ such that the agent need not consider outcomes of that lottery of probability less than e in coming to a fully rational decision. (Smith 2014, 472)

Smith calls this the "Rationally Negligible Probabilities principle" (RNP).

D'Alembert, Buffon, Nicolaus and Daniel Bernoulli, and I all endorse RNP. In fact, we endorse a claim that's stronger than RNP – we endorse the claim that one *should* not consider very-small-probability outcomes in coming to a fully rational decision, while Smith just holds that it is *rationally permissible* not to consider such very-small-probability outcomes.

Why does Smith endorse RNP? One reason is that it can yield a reasonable result for the Pasadena game. But Smith recognizes that more would need to be said – so he gives the following argument based on *tolerance* reasoning. His argument in a nutshell is: decision theory specifies that rational agents should ignore outcomes with zero probability; given that decision-making is a practical activity, infinite precision cannot be required; so instead "there must be some finite tolerance – some positive threshold such that ignoring all outcomes whose probabilities lie below this threshold counts as satisfying the norm" of rationality (Smith 2014, 472).

I have concerns with decision-theoretic reasoning based on tolerance. Obviously I am happy with RNP, but my worry is that if one blindly applies tolerance-based reasoning one will get unacceptable results. Here is an example of what I have in mind.[23]

Suppose you are a scientist who assigns probability b to hypothesis B, and probability c to hypothesis C. It's taken you much research to arrive at these probability assignments, and you've done so with the goal of answering the important scientific question of whether c is higher than b. You have discovered that c is slightly higher. But by tolerance reasoning: decision-making is a practical activity, and infinite precision cannot be required, so it's rationally permissible to treat c as slightly lower than b. So you report the results of your research as showing that hypothesis C is less likely than hypothesis B, despite the fact that $c > b$.

---

[23] And for many more pages of trenchant criticism of Smith's "tolerance" move, see Hájek (2014, 551-560) – my criticism here is complementary to his.

I maintain this is unreasonable – and Smith would presumably agree. What this shows is that Smith does not endorse tolerance reasoning in general in decision theory – tolerance reasoning can be used to support RNP, but it can also be used to support unreasonable results. Perhaps there is a more limited form of tolerance reasoning that supports RNP but elegantly evades the unreasonable results – I'd be happy to see that reasoning spelled out, and in principle I'd be happy to endorse it. But I don't see that reasoning spelled out by Smith, and I don't have any insights to add. For now, my view is that there are better, non-tolerance-based, reasons to support RNP – I'll present three such reasons in the next three subsections. (But for the first two reasons, I still have some qualms, as you'll see.)

5.4 Loss Aversion

I think that it is sometimes rationally permissible to desire to maximize expected utility. But there are other desires that are rationally permissible too, that conflict with the desire to maximize expected utility. One such desire is the desire to *avoid loss*. Eric Schwitzgebel (2015, 283), for example, has this desire: he says that he will sometimes make a decision that increases the chances of avoiding loss, even if that decision is not the one that maximizes expected utility. He goes on to say:

> I might even reasonably decide that at some level of improbability – $1/10^{30}$? – no finite positive or negative outcome could lead me to take a substantial almost-certain loss. And if the time and cognitive effort of sweating over decisions of this sort itself counts as a sufficient loss, then I can simply disregard any possibility where my credence is below that threshold. (Schwitzgebel 2015, 283)

By endorsing Nicolausian discounting, in a Pascal's Mugging scenario Schwitzgebel can avoid the almost-certain loss of giving the mugger his money, regardless of what positive finite utility the mugger is offering. And for the St. Petersburg game, the $1/10^{30}$ threshold would lead Schwitzgebel to value the game at \$99 – in the ballpark of reasonableness.[24]

Schwitzgebel's rationale that time and cognitive effort themselves count as a loss is an promising but limited one. It is promising in that it

[24] The probability of getting heads for the first time on the 99th toss is $1/2^{99}$, which is about $1.6\text{x}10^{-30}$; the probability of getting heads for the first time on the 100th toss is $1/2^{100}$, which is about $8\text{x}10^{-31}$, less than the 10^{-30} threshold.

offers good pragmatic justification in some contexts for ignoring small-probability options. Our time and cognitive effort in life is painfully finite, so it's not worth worrying about possibilities that are very unlikely to occur. But Schwitzgebel's rationale is limited because it only apples to non-ideal agents who need to spend time and cognitive resources to think. An ideal agent would be able to think about arbitrarily complex topics infinitely quickly, and so wouldn't face that limitation. But I maintain that even an ideal agent should not give up a lot of utility to play a St. Petersburg game, and should not hand over money to Pascal's mugger.

Let's set aside the idea that time and cognitive effort themselves count as a loss, and look directly at the connection between loss aversion and Nicolausian discounting. While I share Schwitzgebel's desire to avoid loss, and I agree with Schwitzgebel that we should do Nicolausian discounting, I want to raise two worries regarding the connection between that desire and that strategy.

Here is a story to illustrate my first worry. Imagine that Schwitzgebel hires Bernie as his money manager. Schwitzgebel, worried that Bernie might invest in a Pascal's Mugging scenario, tells Bernie to disregard any possibility where the probability is below the threshold of $1/10^{30}$, regardless of the finite outcome associated with that possibility. Bernie then transfers Schwitzgebel's money from Bank A to Bank B, for the sole reason that the transfer fee is a $1/10^{31}$ chance of Bernie getting all of Schwitzgebel's money.

Bernie can truthfully tell Schwitzgebel that things will (most probably) be fine. But Bernie is exposing Schwitzgebel to a potential loss that Schwitzgebel didn't need to be exposed to – Bernie's action is not an action that a loss-averse person would endorse.

Here is the lesson I draw from this. Nicolausian discounting is a rational thing to do, and having the desire to avoid loss is a rational desire. For Pascal's Mugging, that strategy and that desire cohere, and both give the result that one should refrain from giving the mugger money. What the money transfer scenario shows is that there are other situations where the relationship between that strategy and that desire is more complicated than Schwitzgebel suggests.[25]

---

[25] What's my own view about Bernie's behavior? From the perspective of the Nicolausian-discounting-based decision theory I'm promulgating, Bernie's behavior is rationally permissible (setting aside dominance considerations, which I'll discuss in Section 7). But Bernie's behavior is not *ethically* permissible. This raises a key question: what is the relationship between decision theory and ethics?
[continued on next page]

Here is my second worry regarding the connection between avoiding loss and Nicolausian discounting – but unlike with the first worry, I think that this second worry is fully resolvable. The worry is that engaging in Nicolausian discounting makes one a potential money pump, where one continually faces potential loss with no countervailing gain. This is not the sort of situation a loss-averse agent would want to find herself in. Continuing the story above, imagine that Bernie excitedly does the transfer, only to find that he did not get the money. He then decides to set up an automated system that transfers Schwitzgebel's money back and forth until Bernie hits jackpot with the transfer fee.

The lesson to be learned here is that an agent engaging in Nicolausian discounting will have to evaluate potential gambles collectively – the choices the agent makes may be different than when evaluating gambles individually. But this sort of thing is true for orthodox decision theory too – consider for example the immortal person with the ever-better bottle of wine, where each day she is better off not drinking the wine, but those choices taken together will lead to her never drinking the wine. She will have to

---

Consider the formulation of orthodox decision theory we discussed at the beginning of this paper: "it is wrong always, everywhere, and for anyone" to do anything other than maximize expected utility. I ask the orthodox decision theorists – "wrong" in what sense? If they answer "in the sense of being irrational" then their view has the problematic consequence that rationality and ethics are in conflict. Sometimes the action that maximizes my expected utility is an unethical action (such as when I have the opportunity to steal from someone I hate and I'm guaranteed I won't get caught).

Perhaps the orthodox decision theorists would reply that the view of utility I'm utilizing is overly egocentric – perhaps they would say that I shouldn't focus on maximizing my own utility, but on everyone's utility. But on that view, orthodox decision theory surprisingly entails utilitarianism. This is problematic, because utilitarianism is false – it, for example, licenses sexual assault. (Imagine that Sam and Pat are forever alone on a desert island, and Sam wants to kiss Pat, but Pat does not want to be kissed by Sam. If Sam would get more utility from the kiss than Pat would get disutility, the moral action according to utilitarianism is Sam kissing Pat.)

One way to resolve the worry I'm raising is to endorse a lexical view of the relationship between decision theory and ethics: ethical principles conceptually come first; they are used to determine the permissible actions that are then chosen between by decision theory. I'm unhappy with that lexical view, because it doesn't capture the richness of human experience – we make ethical decisions under uncertainty; ideally decision theory would help with those decisions too.

(For a useful introduction to the (not sufficiently developed) decision theory/ethics literature, see Lumer 2010.)

evaluate her choices collectively to avoid ending up in the clearly inferior situation where she never drinks the wine.[26]

5.5 The Unbearable Amorphousness of Very Small Probabilities

Let's turn to another reason to engage in Nicolausian discounting. In the Pascal's Mugging game, there is one salient very-small-probability possibility – that paying the mugger yields a vast increase in utility. But there are other very-small-probability possibilities that aren't made salient but are also prima facie relevant to deciding what one should do. For example, there is the possibility that another mugger is waiting a five-minute walk down the road, and this mugger will offer Pascal vastly more utility for the money Pascal would have given the first mugger. There is the possibility that Pascal is wrong about his moral judgements, and increasing the utility of sweet little orphans is actually a morally horrendous thing to do. The list

---

[26] For many more orthodox-decision-theoretic examples, see Arntzenius, Elga, and Hawthorne 2004.

Evaluating gambles collectively allows one to avoid the objection to Smith's RNP (and, by extension, Nicolausian discounting) by John Matthewson, as presented in Hájek 2014, pages 561-2. The objection presents a gamble with an infinite set of possibilities that, taken together, give one probability 1/2 of getting $1, but holds that Smith could unreasonably value that gamble at less than $0.50, because some of the possibilities in the gamble have very low probability. I maintain that a proponent of Nicolausian discounting should evaluate the possibilities collectively, and hence value the gamble correctly at $0.50.

Such collective evaluation could also arguably be used to avoid the following sort of objection to Nicolausian discounting. The objection runs as follows: if the threshold is, say, 1/10,000, then consider a scenario where there are 10,001 mutually exclusive and exhaustive equiprobable possibilities – by the lights of Nicolausian discounting, each of those possibilities should have its probability discounted to 0, but then we are left with an absurd decision-making context where there is a probability of 0 that anything will occur. My reply to this objection is that, by engaging in collective evaluation, one can apply Nicolausian discounting to a particular possibility, while still recognizing that one of the other 10,000 possibilities will occur.

If this paper were a full-fledged defense of Nicolausian discounting, then I would have to present a precise procedure for such collective evaluation. I don't have such a procedure, but (as those who aren't reading this footnote will find out when they get to the conclusion) the ultimate point of this paper isn't to defend Nicolausian discounting; it's to use this discussion of Nicolausian discounting to motivate folks to come up with a decision theory that's better than the flawed orthodox decision theory.

goes on and on, and we shouldn't expect Pascal to even be aware of the space of all such remote possibilities, let alone have precise probability assignments for each of the possibilities. Pascal should recognize that his probability assignments for such possibilities are very small, but beyond that we should expect his probability assignments to be *amorphous*.

As a practical matter, we aren't cognitively capable of making well-thought-out precise probability assignments for that immense space of remote possibilities. Hence, it would be irrational to focus on one such possibility simply because the mugger has made it salient, when there are other such possibilities that would also need to be taken into account to make a considered decision. Because we aren't cognitively capable of taking them all into account, and because the probabilities associated with the possibilities are all very small, the best thing to do is to ignore all such possibilities, by discounting their probabilities to zero.[27]

Let's compare my amorphousness argument to the symmetry argument given by Ord, Hillerbrand, and Sandberg (2010, 203). Imagine that you find yourself worrying that dropping a pencil will destroy the world. They write: "for events like the dropping of a pencil which have no plausible mechanism for destroying the world, it seems just as likely that the world would be destroyed by *not* dropping the pencil. The expected losses would thus balance out." A similar argument is given by Schwitzgebel (2015, 283): he says that his "credences about such extremely remote possibilities appear to be approximately symmetrical and canceling", and hence the remote possibilities should be discounted.

---

[27] For those familiar with Yoaav Isaacs' (2016) objection to Smith (2014): my discussion can be construed as a reply to Isaacs. Isaacs writes that, according to Smith's RNP principle:

> lotteries with a very high probability of producing a penny and a very low probability of producing a calamitous loss may be rationally evaluated as though they were sure to produce the penny … But it seems deeply wrong to ignore the obviously salient possibility of calamitous loss. (Isaacs 2016, 761)

My reply to Isaacs is that, when it comes to decision-making, ignoring that very-low-probability possibility is indeed a reasonable thing to do (unless one can take it into account in a cost-free way, as I'll discuss in Section 7). Isaacs' focus on the salient possibility is parochial – there are always many very-low-but-not-zero-probability possibilities that involve calamitous loss, and we rationally implicitly set these aside all the time.

The symmetry argument is different from my amorphousness argument above. I'm not saying that for every remote possibility there is a symmetric countervailing possibility with equal probability. To say that would require too much cognitive sophistication – it would require an unreasonable amount of knowledge of the immense space of remote possibilities, and also an unreasonable level of precision in making probability assignments to this space of possibilities. So I reject the symmetry argument.

I'm a fan of my amorphousness argument, but I recognize that it has limitations. I'll point out two.

One limitation is that my amorphousness argument only applies to rationally limited agents – it does not apply to a rationally perfect being who is aware of the whole space of possibilities, and can assign probabilities and utilities to each of those possibilities. But I maintain that even such ideally rational agents should not give a high value to the St. Petersburg game, and should not give the money to the mugger.

Another limitation of my amorphousness argument is that, even for rationally limited agents like us, it does not apply in all situations. Consider a group of effective altruists who have $10 billion to give away, so they spend years developing the cognitive sophistication to assign precise probabilities to all the remote possibilities they are aware of. And suppose it turns out that, while their credences about these remote possibilities are *approximately* symmetrical and canceling, they are not *exactly* symmetrical and canceling. Should the altruists then simply do the math and spend the $10 billion in whatever way maximizes expected utility, even if it is spent on a possibility associated with a very small probability?

I want to say "no", but I don't think my amorphousness argument provides a strong enough reason to say "no". We can get a better argument by coupling that amorphousness argument with a judgement-stability line of reasoning: we should prefer to make life-changing decisions that we would not call into question upon making tiny adjustments in our probability assignments. The effective altruists should recognize that there are competing possibilities with respect to how a group of rationally-minded people would make such judgements. If a different cognitively sophisticated group would get a vastly different judgement about the most effective action by making a tiny change in the probability assignments, this should lead the initial group to recognize that their judgements are not very *stable*. Moreover, one should expect such instability, given that one is dealing with very small probabilities.

But even when we add the judgement-stability line of reasoning, my amorphousness argument is limited. The effective altruists could recognize that their judgement is imperfect, but at some point one has to act. Maybe the $10 billion will disappear if they don't give it away, or maybe they simply recognize that, if they don't give it away, future decision-makers will also feel uncertain in their judgement, and will have the same reason to not give it away. One never reaches a stage where one can say with certainty 'all the judgements have been made; all the information is in' – Neurath's ship keeps sailing. So at some point one has to act – but I maintain that there are situations where that act should not be the one that maximizes expected utility.

5.6 YOLO

It's time for my favorite reason to engage in Nicolausian discounting. The reason is that it is rational to value how your life actually goes. In the end, looking back on your life, it doesn't matter for your well-being to what extent you maximized expected utility. What matters is the utility you actually achieved – how your life actually went. The orthodox decision theorists' focus on maximizing expected utility is a mistake, because (to quote Goethe's *Clavigo*) "one lives but once in the world". Or to quote millenials c. 2012: "YOLO" – you only live once. The prescription to maximize expected utility does not take seriously the importance of how one's life actually goes.

The orthodox decision theorist might reply: the whole point of decision-making is that we are doing it under uncertainty; we don't know how things will actually go. Of course the goal is to maximize actual utility, but given the epistemic uncertainty, the best we can do is to maximize expected utility. What more do you want?

There is something fundamentally right about this reply – maximizing expected utility does in many cases capture the best way to make a decision under uncertainty. But there is something fundamentally wrong about the orthodox decision theorist's reply too, and this wrongness is evident in situations involving low-probability high-utility possibilities. Maximizing expected utility can put one in a situation where one has a high probability of having one's life actually go badly. Because one only lives once, one has good reason to avoid choosing an action where one has a high probability of having one's life go badly, regardless of whether or not the action maximizes expected utility.

I have used the days-of-life St. Petersburg game and Pascal's Mugging to present situations where the choice that maximizes expected utility also gives one a high probability of having one's life go badly. Because how your one-off life actually goes is of crucial importance to you, maximizing expected utility in those situations is not a reasonable thing to do. Nicolausian discounting allows one to avoid these unreasonable consequences of orthodox decision theory.[28]

Moreover: orthodox decision theory specifies that the value of an action is the same, regardless of whether it is one-off or part of an iterated series. I maintain that, because of the importance of how one's life actually goes, how rational agents value an action that they have the opportunity to repeat over and over might be different than how they value an action that they only have the opportunity to do once. Pascal's Mugging provides a good example of this. Suppose I think that there is a 10^{-40} chance that the mugger will give me the prize, but I have the opportunity (and the financial resources) to play the game 10^{50} times (and each trial is independent, and playing the game 10^{50} times can happen quickly, and so on). In this situation I would want to play the game – there is a high probability that I will come out ahead. But if I only have the opportunity to play once, I'd pass – there is a high probability that I will come out behind. Orthodox decision theory specifies that I should play the game regardless of whether it is iterated or not – orthodox decision theory specifies that both the one-off version and the iterated version have positive expected utility. But what orthodox decision theory fails to adequately take into account is my probability of *actually* winning money.

Nicolausian discounting allows one to avoid playing the one-off version of the game because the probability is too low that one will win; that probability should be discounted to zero for the purposes of decision-making. For the iterated version of the game, the probability is high that one will win,

---

[28] Note that my YOLO argument is in some sense stronger than my Nicolausian discounting conclusion. I'd use the same sort of YOLO argument to support Buchakian risk-weighting (which I'll talk about in the next subsection). And the same sort of YOLO argument could potentially be used to support even larger deviations from orthodox decision theory. Indeed, I take Nicolausian discounting (and Buchakian risk-weighting) to be needed first steps away from orthodox decision theory and toward rationality (or at least reasonableness). But I don't have a perfectly rational, reasonable decision theory to hand you just yet (sorry). I'll have a bit more to say about this when we get to this paper's rousing conclusion.

so – when the iterations of the game are evaluated collectively, as they should be – Nicolausian discounting does not apply.[29]

---

[29] Jeff Jordan (1994, 215-6) endorses a principle closely related to Nicolausian discounting, and argues for it using YOLO-style reasoning. But his argument ends up being confused. For details, keep reading.

> Jordan calls the principle he endorses the "Sure Loss Principle" (SLP):
> if two acts A and –A are such that the EU [expected utility] of A is greater than the EU of –A, but the probability of the favorable consequence of A occurring is such that the performance of A will probably result in a significant net loss for the agent, then the agent ought either to (i) perform –A, if there is no risk of great loss; or (ii) perform neither A nor –A. (Jordan 1994, 215)

Applied to Pascal's Mugging, where A is paying the mugger a significant amount of money and –A is refraining, SLP has the consequence that the agent should perform –A, because paying the mugger will probably result in significant loss. Note that in this situation SLP dictates that the agent should perform –A regardless of how high the expected utility of A is: "According to the SLP one should decline *any* act in which a resultant net loss is practically certain" (Jordan 1994, 215, my emphasis).

Jordan gives two arguments for the SLP. The first is unproblematic: he says that "something like the SLP is needed whenever infinite utilities are included". The second is where the trouble arises. Jordan writes:

> Secondly, a violation of the SLP will lead, almost certainly in the short term and often in the long, to the agent suffering a loss of the stake with little, if any, gain. Even if one has great resources, a net loss is practically certain, given the Law of Large Numbers. (Jordan 1994, 216)

If we look at a one-off version of the Pascal's Mugging game, then indeed a net loss is practically certain, and that's a reason to decline to play – that's YOLO-style reasoning. But Jordan is mistaken to say that the law of large numbers implies that a net loss is practically certain in a sufficiently iterated version of the Pascal's Mugging game. Since the expected utility of the game is positive, then what the law of large numbers implies is that in the long run the person playing the Pascal's Mugging game will come out ahead. (Think about it in a casino context: the house, with its positive expected utility and great enough bankroll, always wins in the long run.)

So what was Jordan thinking? He continues the discussion from the quote above by applying the law of large numbers to a single iteration of the St. Petersburg game. But his discussion of violating the SLP "in the long" term, and his mention of the person being offered the game having great resources, implies that at that point he was thinking of a situation where the game in question is iterated. My interpretation is that Jordan got confused here, but we can set aside the confusion and recognize that there's a YOLO-style insight at the core of his second argument. The insight is that a violation of the SLP can lead to an agent suffering a significant loss with near certainty, and that is a good reason to endorse something like the SLP.

To further motivate my YOLO reasoning, let's think about a one-off Pascal's-Mugging-style scenario from the perspective of effective altruists. If the effective altruists' only desire were to maximize expected utility, then (let's suppose) they would give all their donation money to a very-small-probability cause. But rational effective altruists could have other desires too, such as the desire to *actually do good*. In a situation where they have these multiple reasonable yet conflicting desires, it would be rational for them to reject the option that maximizes expected utility but has a very small probability of doing good, and instead donate in such a way that has lower expected utility but a higher probability of actually doing good.

At this point some interlocutors have objected: the effective altruists who lower the probability of something bad happening count as actually doing good, simply because they have lowered the probability. I maintain that this is mistaken – at least, this is not what I mean by "actually doing good". Here's an intuition-pumping example. Imagine that, 10 years from now, scenario A will occur with high probability, or scenario B will occur with low probability, and the scenarios are mutually exclusive and exhaustive. If scenario B occurs, there is a very high probability that everything will be fine, but there is a very low probability that something catastrophic will happen. Suppose that a group of people spend billions of dollars and years of their lives to reduce the chance of something catastrophic happening in scenario B. Then, scenario A occurs. It's true that those people lowered the probability of something bad happening, and this is often a prudent thing to do. But there's an important sense in which their work was in vain. And, if the initial probability of catastrophe is low enough, then at the time that they are deciding how to proceed, it is rational for them to spend their time and money in such a way that may have lower expected utility, but a higher probability of actually doing good.

5.7 What About Risk?

We're done with my discussion of the six reasons to engage in Nicolausian discounting. But you might wonder why I didn't talk about *risk*. Lara Buchak (2014) has a beautiful risk-weighted decision theory, which I'm on board with – I would want to incorporate Buchakian risk-weighting into my preferred decision theory. But does such risk-weighting help resolve the St. Petersburg paradox or Pascal's Mugging?

Buchak (2014, 73) recognizes that, for a normal risk-weighting function that does not drive probabilities to zero, the answer is 'no'. In either game, the utilities can be increased to compensate for the risk of loss, in

such a way that even a risk-averse agent would deem a suitable version of the St. Petersburg game to have infinite risk-weighted expected utility, and a suitable version of the Pascal's Mugging game to have positive risk-weighted expected utility.

Intriguingly, Buchak does give a proposal for how we can avoid the judgement that the St. Petersburg game has infinite expected utility: her proposal is that the agent utilize a risk-weighting function that has the mathematical effect of discounting very small probabilities to zero. But Buchak gives only a one-sentence justification for this proposal: "we might think that we ought not consider the chance of a good outcome in our calculations if that good outcome is extremely unlikely to happen" (Buchak 2014, 74). What I've tried to do in this section is give reasons for thinking that.

6. Three Open Questions

A key feature of Nicolausian discounting is the low-probability threshold below which, for the purposes of decision-making, one discounts probabilities to zero. This leads to three unresolved questions. Is this threshold subjective or objective? Does the value of the threshold depend on context, or is it invariant? How should one handle probabilities greater than but close to the threshold? I'll give my own answers to those three questions now – but of course someone could be a fan of Nicolausian discounting and yet give different answers to these questions.

6.1 Is the Threshold Subjective?

People who believe that the threshold is objective face the burden of telling us what the value of the threshold is – and to their credit, Buffon and Daniel Bernoulli are happy to provide numbers. They aren't the only ones: Le Marquis de Condorcet, for example, endorses a threshold of 1/144,768. Condorcet's argument for this threshold is that it is the difference between the probability that a 47-year-old man would die in the course of a day and the probability that a 37-year-old man would, and that is a difference that would keep no man awake at night.[30]

[30] Condorcet criticizes Buffon's choice of threshold, and then provides his own; he calls 1/144,768 "the greatest risk that can be regarded as null". This is my translation, the original is "le risque le plus considérable qu'il foit permis de regarder comme nul" (Condorcet 1785, page cxiii).

The fact that this is so amusing in its specificity is an indicator that it can't be objectively right. Contrast that experience with the experience of having some objectively true philosophical insight – such as reading Gettier's (1963) paper and saying "oh yes, now I see, knowledge is not justified true belief". I can't imagine having an experience like that with Condorcet's argument – "oh yes, now I see, the threshold is 1/144,768." I can't think of any argument for an objective threshold that fares better than Condorcet's in that amusingness regard, so I conclude that the threshold is *subjective*.[31]

At least, it is subjective within reason. If someone set the threshold to, for example, 1/2, I'd deem them irrational (and, with their resultant poor decision-making skills, they would probably not be long for this world). And if someone set the threshold to an astronomically small number, I'd deem them at least unreasonable – they would, for example, be willing to assign too high a value to playing the St. Petersburg game.

So what is your threshold? One way to get insight is to contemplate how much you would pay to play a version of the St. Petersburg game with no diminishing marginal utility. For me, the finite days-of-life version I presented in Section 2 is such a game, and recall that I wouldn't be willing to give up more than 50 days of life to play the game. So for me, the threshold is somewhere between $1/2^{50}$ and $1/2^{51}$. That means that, for the purposes of decision-making, I am treating the probability of getting tails 50 times in a row, and then heads on the 51st toss or later, as a probability 0 event. The number $1/2^{50}$ is about $9x10^{-16}$, while $1/2^{51}$ is about $4x10^{-16}$, so my threshold is about $5x10^{-16}$ – about 1 in 2 quadrillion. I'm happy with that as my threshold – in the course of decision-making, I'm happy to ignore possibilities with probabilities less than 1 in 2 quadrillion.

6.2 Is the Threshold Context-Dependent?

Suppose you use the above method of introspecting how many days of life you would give up in the finite days-of-life version of the St. Petersburg game to determine your threshold. Now we can ask the question: is this your

[31] My appeal to subjectivity defangs the critique of arbitrariness from Kenneth Arrow (1951, 414): "This principle [of engaging in Nicolausian discounting] seems extremely arbitrary in its specification of a particular critical probability [i.e. the threshold]". I say: a choice of threshold is unproblematically arbitrary because it is based on subjective preference, in the same way that a choice between eating vanilla or chocolate ice cream is unproblematically arbitrary.

threshold for all contexts, or just contexts where your life is at stake? Perhaps one's threshold should be larger for less important contexts, like deciding what to have for dinner, and one's threshold should be smaller for more important contexts, like when the fate of the whole world is at stake.

Émile Borel, who is a proponent of Nicolausian discounting, maintains that one's threshold should vary depending on the context. I'll present his position, and then reject it.

In Borel's 1943 book *Le Probabilités et la Vie* (translated into English as Borel 1962), Borel endorses this claim:

> *Events with a sufficiently small probability never occur;* or at least, we must act, in all circumstances, as if they were *impossible*. (Borel 1962, 3)

In this ensuing discussion, it's clear that he's not really acting *in all circumstances* as if they were impossible – he's happy to do math calculations involving very small probabilities where he gives the right answer, which is a non-zero answer. The charitable way to read Borel is that he is talking about *decision-making* circumstances. So I take Borel to be endorsing Nicolausian discounting.[32]

But what is Borel's threshold? Borel gives four different numerical thresholds, for probabilities which are negligible on the human scale, the terrestrial scale, the cosmic scale, and the supercosmic scale. His corresponding threshold values are 10^{-6}, 10^{-15}, 10^{-50}, and 10^{-500}. The human scale corresponds to possibilities regarding the life of a person; the terrestrial scale corresponds to possibilities regarding human civilization as a whole; the cosmic scale corresponds to possibilities regarding the observable universe, and the supercosmic scale corresponds to possibilities regarding violations of the laws of physics.

---

[32] But *why* does Borel endorse Nicolausian discounting? It's not clear. Borel says on the first page of his book that the calculus of probabilities permits prediction as a result of "the *single law of chance*", which is "phenomena with very small probabilities do not occur" (Borel 1962, 1). I take Borel to be endorsing Antoine-Augustin Cournot's (1843) claim this single law of chance is what gives empirical meaning to probability theory. While I am skeptical of this foundational claim, discussing it is beyond the scope of this paper. (See Shafer and Vovk (2006, 73) for some discussion.)

A nomenclature note: this foundational claim is sometimes called "Cournot's Principle", but sometimes "Cournot's Principle" is just used to refer to something more like the rule that one should engage in Nicolausian discounting. Let's give credit where credit is due; since Cournot was the first to state the foundational claim, let's call *that* claim "Cournot's Principle".

This initially has an air of reasonableness, that the threshold should vary depending on the types of possibilities under consideration. But I will argue that a reasonable person should have a single threshold that does not vary with context.

To start my argument, let's look at what Borel says about carrying an umbrella. Suppose that you're in France, it's 10 am, the weather is nice, you haven't seen a weather forecast, and you're about to head out the door. Borel guesses that the probability is somewhat larger than 1/1000 that in such a situations, it rains in the afternoon. But unless you're frail, "we will not consider [you] careless" if you go out without an umbrella (Borel 1962, 30). So far I agree. But Borel is using this umbrella example to illustrate the following claim:

> much larger probabilities [than the human-scale threshold of 10^{-6}] must also be disregarded in the numerous cases where the event corresponding to such probabilities does not represent for us a grave misfortune, but merely a disagreeable incident. (Borel 1962, 30)

What this shows is that, by Borel's lights, there aren't just four thresholds; there are many, corresponding to the relative graveness of the possible misfortune under consideration. Thus, in Borel's decision theory, an agent's preferences would have to be modeled not only by a probability function and a utility function, but also a complicated threshold function. (Borel doesn't spell it out that way (he doesn't spell it out at all), but I think that's how the account would most reasonably go.)

I reject this proposal of a threshold function. Orthodox decision theory already has the resources to explain the decision not to carry an umbrella in the situation Borel describes. When you're about to leave your house at 10 am, you (at least implicitly) do an *expected utility calculation* to decide whether it's worth it to carry the umbrella. Consider the process of carrying the umbrella – just the act of carrying itself, setting aside the umbrella's usefulness if it rains. This process of carrying the umbrella has some negative utility for you, and the expected utility calculation determines whether carrying the umbrella is all-things-considered worth it, given the small chance of it doing any good for you. We don't need Borel's threshold function to account for umbrella-carrying behavior; orthodox decision theory does just fine.

But that just implies that using Borel's threshold function is unnecessary; I'll now show that it's a mistake. By the lights of orthodox decision theory, if the process of carrying the umbrella had zero utility for you, then you're irrational if you don't carry it, since there's a chance it will

do some all-things-considered good. But by Borel's lights, you should apply the threshold to treat the probability of rain in the afternoon as a zero-probability event, and hence you should be indifferent between carrying the umbrella or not. Orthodox decision theory is closer to the truth here – it's irrational (or at least careless) *not* to carry the umbrella, in a situation where the process of carrying it doesn't have any negative utility.

I'm not aware of any proposal to have the threshold vary with context that is better than Borel's proposal, and Borel's proposal fails. So that's a reason to conclude that a reasonable person should have a single threshold that does not vary with context. Here is another consideration in favor of that conclusion – or at least, here is my explanation for why I personally have a single threshold that does not vary with context.

Consider the finite days-of-life version of the St. Petersburg game, and recall that I'm willing to give up 50 days of life to play the game. What if I am making the choice, not for my life, but for the life of a chinchilla? For a chinchilla due to die in 1000 days, how many days of the chinchilla's life am I willing to give up to play the game to determine when the chinchilla's life will actually end? Or what if I am making the choice for all of human civilization, so that civilization as a whole is scheduled to end in 1000 days, and I am the one to decide how many days of civilization's continued existence to give up to play the game to determine when civilization as a whole will actually end?

Whether it's for a chinchilla or civilization, I'd still be willing to give up no more than 50 days to play the game, and hence my threshold in both these contexts is about 1 in 2 quadrillion. But the death of a chinchilla and the death of civilization are vastly different contexts; this provides evidence that *my* threshold, at least, does not depend on context.

6.3 How Should One Handle Probabilities Close to the Threshold?

Let's suppose that 1/1000 is our threshold. How should we treat probabilities greater than but close to this threshold, like 1/999 and 1/998 – should we discount them somewhat? D'Alembert poses this question, asking:

> If we do not look at these probabilities as smaller than they actually are, I ask how the probability 1/999 suddenly becomes = 0 in the case where it is 1/1000? Can the expression of

> probability thus shift sharply & without gradation, from a finite expression to a null value?[33]

D'Alembert does not give answers to these questions – he simply concludes that "the general rule for the calculus of probabilities is defective in certain respects."[34] With this I agree, but nevertheless, working within the context of Nicolausian discounting, we can answer his questions.

My answer is: yes, in a decision-making context, the expression of probability does shift sharply and without gradation to zero when one reaches the threshold. At least, this is the answer I give as an agent with limited cognitive resources. While I share d'Alembert's implicit worry that the difference between 1/999 and 0 is large, I do not feel the same about the difference between 1/2000000000000000 and 0 – with my limited cognitive resources, these numbers seem really close. Moreover, if I were to try to implement some discounting function that discounts probabilities close to the threshold to a lower but non-zero number, that needlessly adds to my already-limited cognitive load. Would an ideal agent implement a discounting function instead of a single threshold? Perhaps; my limited cognitive resources prevent me from giving a more definitive answer.[35]

Nicolaus himself worried both about the issue of what the threshold is, and the issue of how handle probabilities near the threshold. In a 1728 letter to Gabriel Cramer, Nicolaus writes:

> it is necessary … to determine how far the quantity of a probability must diminish, in order for it to be considered as zero; but this is impossible to determine, whatever supposition one may make, one always finds difficulties; the limits of these small probabilities are not precise, but they have a certain latitude which cannot be fixed easily…[36]

---

[33] This is d'Alembert 1761, Memoir 10, Section 13, page 11, my translation. The original is: "S'il ne faut pas regarder ces probabilités comme plus petites qu'elles ne sont en effet, je demande comment la probabilité 1/999 devient tout d'un coup = 0 dan le cas où elle est 1/1000? L'expression de la probabilité peut-elle passer ainsi brusquement & sans gradation, d'une expression finie à une valeur nulle?"

[34] D'Alembert 1761, Memoir 10, Section 14, page 12. This is my translation; the original is "la régle générale du calcul des probabilités est défectueuse à certain égards."

[35] Presumably Buchak, whose cognitive resources are less limited than mine, would utilize a continuous risk function that drops smoothly to zero.

[36] This is my translation; the original is "Il faut donc … determiner jusqu' où la quantité d'une probabilité doit diminuer, afin qu'elle puisse étre censée nulle; mais voilà ce qui est impossible de determiner, quelque supposition que l'on fasse, on rencontre toujours de difficultés; le limites de ces petites probabilités ne sont pas precises, mais elles ont une [continued on next page]

This isn't the first time philosophers have faced a situation where it looks like a line needs to be drawn but any choice of where to draw it seems problematic. I'm happy to make an arbitrary, subjective choice, but that's only because I don't know of a better option.

7. Objection: Domination (And, Pascal's Wager)

There's no point in worrying about the open questions if Nicolausian discounting itself is a fundamentally flawed decision-theoretic procedure. In this section, I'll consider the one objection I find most pressing.

I endorse Nicolausian discounting with a threshold of $5x10^{-16}$. You might think that it follows that, when it comes to decision-making, I treat the following two possibilities indifferently, with respect to their expected utility:

> *Option A*: getting \$2 with probability $(1-10^{-50})$, and losing \$1 with probability 10^{-50}
>
> *Option B*: getting \$2 with probability $(1-10^{-50})$, and losing all one's money with probability 10^{-50}

But the objector would say: surely (surely!) when it comes to decision-making, one should not be indifferent between Option A and Option B, and hence Nicolausian discounting is a flawed decision procedure.

To orthodox decision theorists who present such an objection, I reply: *tu quoque*. As Hájek (2014, 556-7) points out, orthodox decision theory faces the same sort of problem of prescribing indifference between options that shouldn't be considered indifferently. According to orthodox decision theory, these two options both have expected utility of 0:

> *Option 1*: A fair coin is flipped an infinite number of times in a finite amount of time; you go to heaven if the coin lands heads on every toss, and get nothing otherwise.
>
> *Option 2*: Same as above, except you go to hell if the coin lands heads on every toss.

Surely, we should not be indifferent between those two options.

But "tu quoque" by itself isn't a solution to the problem posed by the objection; it just means that orthodox decision theorists and Nicolausian-discounting decision theorists are in the same boat when it comes to solving the problem.

---

certaine latitude que l'on ne peut pas fixer aisément…". This letter is published in Spieß 1975, page 562.

I (following Hájek) propose to the orthodox decision theorists that they solve the problem by specifying that indeed, when faced with a choice between the zero-probability heaven and hell options, they prescribe that a rational agent choose the dominating heaven option. They need to supplement the ranking of options based on expected utility with dominance reasoning. But the dominance reasoning should only be implemented when it is cost-free. It's not worth spending any time or money to get the heaven option over the hell option, since the two options have the same expected utility.

I propose that Nicolausian-discounting decision theorists make the same move. When faced with the choice between Option A and Option B as specified above, they should choose Option A, but only if it is cost-free. It's not worth spending any time or money to get Option A over Option B, since, post-Nicolausian-discounting, the two options have the same expected utility.

I understand if you feel uncomfortable with this thought of not placing any value on small-probability possibilities in decision-making contexts, especially if such possibilities lead to ruin (losing all one's money, for example). But I maintain that you, 21st-century intellectual that you are, already make such disregarding judgements in the course of decision-making – and moreover, you are correct to do so.

To see this, let's (finally) take up Pascal's Wager – Blaise Pascal's argument that it is pragmatically rational to believe in God, even if one initially assigns a small probability to the hypothesis that God exists, because of the possibility of obtaining the infinite reward offered in heaven. Let's focus on a hell-based version of Pascal's Wager – consider the hypothesis that the traditional Christian God exists, and all and only redeemed believers go to heaven; everyone else goes to hell.[37] 21st-century intellectual that I am, my probability for that hypothesis is very low. But it's not zero – I recognize that there is a tiny chance that that hypothesis is correct. Moreover, when I consider other hypotheses that involve the possibility of ending up in hell (the hypothesis that God sends all and only philosophers to hell, for example) those hypotheses strike me as less probable than the traditional Christian hypothesis, in part because of the number of believers involved. (It is striking that millions of Christians believe that I am going to hell, and I assign a very low but non-zero probability that they are correct.) So out of all the hypotheses that involve

[37] Whether Pascal himself endorsed a hell-based version of Pascal's Wager is controversial – see Section 4 of Hájek 2017.

people going to hell, I consider the traditional Christian one the most probable.[38] Specifically, the probability I assign to that hypothesis is 10^{-50}. But now let's apply orthodox-decision-theoretic reasoning: the disutility associated with hell is, if not infinite, at least vastly more than 10^{50} times larger than the utility I get from my godless earthly life, so orthodox decision theory specifies that the action that maximizes expected utility is to try to become a redeemed believer.

Nicolausian discounting can be utilized to provide a response to Pascal's Wager. By my lights, any hypothesis that involves people going to hell has probability below my threshold, and hence, for the purposes of decision-making, I discount the probabilities of such hypotheses to zero. The process of trying to become a redeemed believer has negative utility for me; hence I rationally decide to refrain from trying to become a redeemed believer.

Perhaps you were uncomfortable with engaging in decision-making while disregarding the small-probability possibility in Option B above, where there was a small chance of losing all your money. Are you, 21st-century intellectual that you are, similarly uncomfortable with engaging in decision-making while disregarding the small-probability possibility that you will spend eternity in hell? I'm guessing – and hoping – that you are in fact comfortable disregarding such silly eschatological hypotheses; it's not healthy to make life decisions with the worry that one might end up in hell.[39] But, from a doxastic standpoint, it's unreasonable to assign all such hypotheses zero probability; there is a chance that such a hypothesis is right, unlike, say, there being a chance that 2+2=5. So I maintain that, when faced with Pascal's Wager, you are already implicitly engaging in Nicolausian

---

[38] For the standard argument that comparing the probabilities of the various hypotheses matters, even when all hypotheses involve infinite utilities, see Schlesinger 1994. For a sophisticated decision theory that allows one to make such comparative probability judgements in an infinite-utility context, see Bartha 2007.

[39] Another silly eschatological hypothesis that I'm hoping you're comfortable ignoring is that of Roko's Basilisk – the hypothesis that an evil future artificial intelligence will eternally punish those of us who did not help it come into existence. (For more on Roko's Basilisk, see for example Auerbach 2014.)

In case you are actually worried about Roko's Basilisk: I invite you to also consider the hypothesis that a future artificial intelligence will be morally good except when it comes to vengeance, and will eternally punish those people who tried to help bring Roko's Basilisk into existence (with a higher cardinality of punishment than whatever punishment you were imagining Roko's Basilisk would dole out). And then I invite you to recall my amorphousness argument of Section 5.5.

discounting. What I have tried to do in this paper is to make such reasoning explicit, and defend its reasonableness.

8. Conclusion

The St. Petersburg game and Pascal's Mugging show that orthodox decision theory is flawed, and I have put forth Nicolausian discounting as a means of avoiding some of the problematic prescriptions of orthodox decision theory. My guess is that open-minded decision theorists are intrigued but unsatiated. Perhaps they want a calculus of decision into which I can input my probability and utility functions to get a cardinal ranking of actions that captures all the decision-theoretic preferences I've expressed above. Or maybe they want something like a representation theorem that shows how all my decision-theoretic preferences can be represented as resulting from some utility function and probability function. I admit, I don't have that calculus, and I don't have that theorem. In fact, I'm skeptical of the view that it's correct to express decision-theoretic preferences solely in terms of a utility function and a probability function. (And, as much as I admire Buchak's work, I'm also skeptical of the view that it's correct to express decision-theoretic preferences solely in terms of a utility function, a probability function, and a risk function.) We need to go back to the drawing board; my ultimate goal of this paper is to provide an impetus to do that.

I admire orthodox decision theory, but orthodox decision theory is predicated on a mistake. A key feature of orthodox decision theory is that maximizing expected utility is always the rational thing to do, but (as you've gathered by now) I think that's problematic. I admire orthodox decision theory because it gets the right answer in a lot of cases, and because it helps people see that often their intuitive judgements about probability are the wrong ones. But it turns out that some of the decisions prescribed by orthodox decision theory are the wrong ones too.

The formal system that orthodox decision theorists have developed is impressive, but the existence of an impressive formal system does not entail that the theory is giving the right prescriptions for decisions. We shouldn't let the existence or non-existence of a formal system blind us to the fact that maximizing expected utility is sometimes an unreasonable – perhaps even irrational – decision procedure to follow.[40]

---

[40] Thanks to my interlocutors: Nick Bostrom, Lara Buchak, Eddy Chen, Kenny Easwaran, Dien Ho, Yitu Hu, Mike Huemer, Brian Kierland, Seahwa Kim, Matt Lutz, Chad Mohler, [continued on next page]

References

Alexander, J. McKenzie (2011), "Expectations and Choiceworthiness", *Mind* 120: 803-17.

Arntzenius, Frank, Adam Elga, and John Hawthorne (2004), "Bayesianism, Infinite Decisions, and Binding", *Mind* 113: 251–83.

Arrow, Kenneth (1951), "Alternative Approaches to the Theory of Choice in Risk-Taking Situations", *Econometrica* 19: 404-37.

Auerbach, David (2014), "The Most Terrifying Thought Experiment of All Time", *Slate*, available at http://www.slate.com/articles/technology/bitwise/2014/07/roko_s_basilisk_the_most_terrifying_thought_experiment_of_all_time.html. Archived on 2018-9-21 at http://www.webcitation.org/72bQ5w5xU.

Bartha, Paul (2007), "Taking Stock of Infinite Value: Pascal's Wager and Relative Utilities," *Synthese* 154: 5–52.

Borel, Émile (1962), *Probabilities and Life*, New York: Dover Publications.

Bostrom, Nick (2009), "Pascal's Mugging", *Analysis* 69: 443-5.

Buchak, Lara (2014), *Risk and Rationality*, Oxford: Oxford University Press.

Bernoulli, Daniel (1738 [English translation 1954]), "Exposition of a New Theory on the Measurement of Risk", *Econometrica* 22: 23–36.

Toby Ord, Ric Otte, Eric Schwitzgebel, Kyle Scott, Philip Trammell, Bas van Fraassen, Paul Weirich, the usual anonymous referees, and my lovely wife Lily. Thanks also to audiences at Lingnan University, Nanyang Technological University, and Wuhan University. I was inspired to write this paper by Luke Muehlhauser's March 2017 Open Philanthropy Project blog post "Technical and Philosophical Questions That Might Affect Our Grantmaking", where he mentions Pascal's Mugging.

Condorcet, Le Marquis de (1785), *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*, Paris: L'Imprimerie Royale.

Cournot, Antoine-Augustin (1843), *Exposition de la théorie des chances et des probabilités,* Paris: Hachette.

d'Alembert, Jean le Rond (1761), *Opuscules Mathématiques ou Mémoires sur différens sujets de Géométrie, de Méchanique, d'Optique, d'Astronomie &c., Tome Second*, Paris: David.

Dutka, Jacques (1988), "On the St. Petersburg Paradox", *Archive for History of Exact Sciences* 39: 13–39.

Easwaran, Kenny (2008), "Strong and Weak Expectations", *Mind* 117: 633-641.

Gettier, Edmund (1963), "Is Justified True Belief Knowledge?", *Analysis* 23: 121-3.

Gibbard, Allan and William Harper (1978), "Counterfactuals and Two Kinds of Expected Utility", in C. A. Hooker, J. J. Leach, and E. F. McClennen (eds.), *Foundations and Applications of Decision Theory, Volume 1*, Dordrecht, Holland: D. Reidel, pages 125-62.

Hájek, Alan (2014), "Unexpected Expectations", *Mind* 123: 533-567.

Hájek, Alan (2017), "Pascal's Wager", *The Stanford Encyclopedia of* Philosophy, Edward N. Zalta (ed.), https://plato.stanford.edu/archives/fall2017/entries/pascal-wager.

Hey, John, Tibor Neugebauer, and Carmen Pasca (2010), "Georges-Louis Leclerc de Buffon's 'Essays on Moral Arithmetic'", *Luxembourg School of Finance Research Working Paper Series*, available at https://eprints.luiss.it/769/1/10-06_hey_2010.pdf. Archived on 2018-9-21 at http://www.webcitation.org/72bQvQRge.

Isaacs, Yoaav (2016), "Probabilities Cannot Be Rationally Neglected", *Mind* 125: 759–62.

Jordan, Jeff (1994), “The St. Petersburg Paradox and Pascal's Wager”, *Philosophia* 23: 207-22.

Lewis, David (1981), “Why Ain’cha Rich?”, *Noûs* 15: 377-80.

Lumer, Christoph (2010), “Introduction: The Relevance of Rational Decision Theory for Ethics”, *Ethical Theory and Moral Practice* 13: 485–496.

Martin, Robert (2017), “The St. Petersburg Paradox”, *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), https://plato.stanford.edu/archives/fall2017/entries/paradox-stpetersburg.

Monton, Bradley (2010), “Against Multiverse Theodicies”, *Philo* 13: 113-35.

Nover, Harris, and Alan Hájek (2004), “Vexing Expectations”, *Mind* 113: 305-17.

Ord, Toby, Rafaela Hillerbrand, and Anders Sandberg (2010), “Probing the Improbable: Methodological Challenges for Risks with Low Probabilities and High Stakes”, *Journal of Risk Research* 13: 191-205.

Parfit, Derek (1985), *Reasons and Persons*, reprinted with corrections, Oxford: Oxford University Press.

Pulskamp, Richard (2009), “Jean Le Rond d'Alembert on Probability and Statistics”, available at http://cerebro.xu.edu/math/Sources/Dalembert. Memoir 10 archived on 2017-06-21 at http://www.webcitation.org/6rOw6h63E.

Samuelson, Paul (1977), “St. Petersburg Paradoxes: Defanged, Dissected, and Historically Described”, *Journal of Economic Literature* 15: 24-55.

Schlesinger, George (1994), “A Central Theistic Argument”, in Jordan, Jeff (ed.), *Gambling on God: Essays on Pascal’s Wager*, Lanham, MD: Rowman & Littlefield, pp. 83–99.

Schwitzgebel, Eric (2015), “1% Skepticism”, *Noûs* 51: 271-290.

Shafer, Glenn and Vladimir Vovk (2006), "The Sources of Kolmogorov's *Grundbegriffe*", *Statistical Science* 21: 70-98.

Smith, Nicholas J. J. (2014), "Is Evaluative Compositionality A Requirement of Rationality?", *Mind* 123: 457-502.

Spieß, Von O. (1975), "Zur Vorgeschichte des Petersburger Problems", in BL van der Waerden (ed.), *Die Werke von Jakob Bernoulli*, *Bd. 3: Wahrscheinlichkeitsrechnung*, Basel: Birkhäuser, pp. 557-67.

Todhunter, I. (1865), *A History of the Mathematical Theory of Probability From the Time of Pascal to That of Laplace,* Cambridge and London: Macmillan and Co.