11-2023

# How to Save Pascal (and Ourselves) From the Mugger

Avram Hiller
*Portland State University*, ahiller@pdx.edu

Ali Hasan
*University of Iowa*

## Citation Details

Hiller, A., & Hasan, A. (2023). How to Save Pascal (and Ourselves) From the Mugger. Dialogue: Canadian Philosophical Review/Revue canadienne de philosophie, 1-17.

*Dialogue* (2023), 1–17
doi:10.1017/S0012217323000392

CAMBRIDGE
UNIVERSITY PRESS

ARTICLE

# How to Save Pascal (and Ourselves) From the Mugger

Avram Hiller[1] and Ali Hasan[2]

[1]Department of Philosophy, Portland State University, Portland, OR, USA and [2]Department of Philosophy, University of Iowa, Iowa City, IA, USA
Corresponding author: Avram Hiller; Email: ahiller@pdx.edu

**Abstract**

In this article, we re-examine *Pascal's Mugging*, and argue that it is a deeper problem than the St. Petersburg paradox. We offer a way out that is consistent with classical decision theory. Specifically, we propose a "many muggers" response analogous to the "many gods" objection to Pascal's Wager. When a very tiny probability of a great reward becomes a salient outcome of a choice, such as in the offer of the mugger, it can be *discounted* on the condition that there are many other symmetric, non-salient rewards that one may receive if one chooses otherwise.

**Résumé**

Dans cet article, nous réexaminons *Le braquage de Pascal* et soutenons qu'il s'agit d'un problème plus profond que le paradoxe de Saint-Pétersbourg. Nous proposons une issue cohérente avec les théories de la décision classiques. Plus précisément, nous formulons une solution « braqueurs multiples » analogue à l'objection « dieux multiples » quant au pari de Pascal. Lorsque le résultat notable d'un choix est une probabilité infime d'obtenir une grande récompense, comme dans l'offre du braqueur, on peut l'écarter pourvu que de nombreuses récompenses symétriques mais non notables soient disponibles si l'on fait un choix différent.

**Keywords:** Pascal's mugging; decision theory; St. Petersburg paradox; probability discounting; Nick Bostrom

We pick up the scene from Nick Bostrom (2009): Pascal is approached in a dark alley by a "Mugger" who insists that he hand over his wallet (and the 10 livres in it). Though the Mugger has no gun or other means of threat, he promises to return the next day with double the amount in Pascal's wallet. Pascal declines the offer, at which point the Mugger offers 10 times the amount. When Pascal declines again, and says that he thinks the chances that the Mugger delivers on the promise is one in 1,000, the Mugger offers to return with 2,000 times the value, and Pascal remains sceptical. This goes on for a while, with the Mugger ultimately claiming that he is from the Seventh Dimension and has magical powers to deliver on *any* promise he might make.

## Part I. In Which Pascal Does Not Agree to Answer the Mugger's Ultimatum Question

**Mugger**: "[W]hat's your probability that I not only have magic powers but that I will also use them to deliver on any promise — however extravagantly generous it may seem — that I might make to you tonight?"[1]

**Pascal**: *Any* promise, however extravagant!?! After all, you might make a promise that is *impossible* to deliver. Maybe you'll promise to bake me a round square chocolate cake!

**Mugger**: OK, I will assure you that I will only promise you something that is metaphysically possible. If it isn't, the deal is off. Now that that's clear, let me ask again: what probability do you assign to my delivering on any promise that I might make?

**Pascal**: I'm still not sure I agree, but let's suppose that you have the magical power to deliver on any metaphysically possible promise. I might give it a very, very low probability that you have that power. But I worry that after I assign some probability, you will propose some incredible reward that I hadn't thought of, which will make me think that the probability is lower than I'd previously taken it to be.

**Mugger**: Well, there's nothing wrong with that. You will have received new information about the content of the promise, and it might then make sense for you to assign a lower probability once you do. But I'm asking you to assign a probability *now* to the claim — which you are granting is possible — that I have magical powers to deliver on any promise I might make. That's compatible with assigning a lower probability once you hear my promise.

**Pascal**: That seems right.[2] However, I feel that I can propose some probability in response to a particular reward, but not in response to the general claim that you will deliver on any promise you might make.

**Mugger**: That's fine. I'm even OK with you giving me a range.

**Pascal**: Alright. I'm still concerned about giving you *any* non-zero value, but let me say, well, um, … let's say that the probability is no greater than one in 10 quadrillion. And the low end of the probability range goes down to infinitesimally above zero.

**Mugger**: What exactly does "infinitesimally above zero" mean? What I'm proposing is metaphysically possible, and shouldn't all metaphysical possibilities have non-zero probabilities? I also told you[3] that my powers are strictly finite, and that the offer does not involve infinite values. So, please still go ahead and give me at least a lower bound that is non-zero.

[1] This line from the Mugger is quoted from Bostrom (2009, p. 444). The rest is our own alternate continuation of the dialogue.

[2] See Armendt (2019) on the appropriateness of assigning probabilities even while knowing that one may later update.

[3] Bostrom (2009, p. 445).

**Pascal**: The problem is that, given that I have already agreed that there is no temporal discounting of possible rewards,[4] there is a possible *infinity* of rewards that you might promise me. If your powers are really strictly finite, as you claim, would you mind telling me where they give out? What is the largest reward that you can deliver?

**Mugger**: In claiming that my powers are strictly finite, I meant that I can only deliver a finite amount, though I can deliver any finite amount you might want, without limit.[5]

**Pascal**: But, for whatever odds I give you as a proposed lower bound, you might promise me a reward that is so extravagant that it would be more unlikely than those odds. That isn't a problem with my predictive capacities or an irrationality of mine — that's just how a series of decreasing probabilities works. It's like playing a game of who can say a higher number — the person who goes second can always win! The core of your trick is just that whatever low odds I give, you can promise to match it with a correspondingly higher value to make it seem to be worth giving you my wallet. This isn't a question of "infinite values" (whatever those are!) — it's just a question of the limitless nature of our mathematical imaginations. So I'm not going to tell you any odds in advance of hearing the specific reward to which the odds directly apply. That seems only fair.

**Mugger**: Fine. I can understand your hesitancy to tell me the odds before I tell you the reward.

**Pascal**: Good! So, I'm keeping my wallet.

**Mugger**: Not so fast.

* * *

## Part II. In Which Pascal Agrees to, and Gives, a Decreasing Probability Function, and Believes That He Is Now Off the Hook

**Mugger**: How about this instead: you seem to be willing to tell me odds *after* I've told you what the reward would be. So, how about giving me, in advance, just a probability distribution of the chances of my fulfilling different possible rewards that I might promise you?

**Pascal**: Well, that seems incredibly difficult! I have no clue what you might promise!

**Mugger**: I'll make it easier on you. We in our Dimension are trained to give people many days of bliss. So, how about you give me a function, for each $x$, going from an input of $x$ days of bliss to an output of a probability of my being able to deliver that many days of bliss to you?

**Pascal**: [Thinks for a moment.] Well, I think that the chance of you giving me even one day of bliss is about one in a billion, and it's going to get lower from there. How about a reciprocal function? Say[6]:

[4] Bostrom (2009, p. 444).

[5] Bostrom (2009, p. 445).

[6] Assume that the Mugger might only promise a number of days of bliss greater than zero.

https://doi.org/10.1017/S0012217323000392 Published online by Cambridge University Press

$$\frac{1}{10^6 \ \times \ (x + 1{,}000)}$$

It’s just a simple function, and maybe you will allow me to add some minor tweaks later, but that should do, since it decreases just as fast as the days of bliss increase. And so, no matter how many days $x$ of bliss you promise me, the odds of you fulfilling that promise are less than one in $x$ million. And, at those odds, it would never be worth handing over my wallet, given that I have some livres in there to lose.

**Mugger**: That’s what you think.

* * *

### Part III. In Which the Mugger Changes the Terms, With a Distressing/Delightful Result in the Utility Function

**Pascal**: What do you mean?

**Mugger**: Well, first try doing a utility calculation with your reciprocal function.

**Pascal**: [Thinks for a moment.] Now that I’m thinking about it, it’s hard to figure out the utility function for my agreeing to your offer. The odds I just gave you are the odds that, for any given promised number of days $x$, *if* you promised me $x$ days of bliss, then you would fulfil that promise. But I can’t calculate the utility function for me giving you my wallet unless I know the odds, for each number of days, that you might promise me that many days of bliss.

**Mugger**: You’ll have to guess, then.

**Pascal**: Any help?

**Mugger**: Well, it’ll have to be a function that sums to one as it heads to infinity, because I *am* going to promise you something.

**Pascal**: OK. How about the odds of you promising me $x$ days of bliss are $\frac{1}{2^x}$?

**Mugger**: Isn’t that rather pessimistic about what I’m going to promise you? I’m going to promise you a *lot* of days of bliss! But you know what? Fair enough. Try putting that into your utility function.

**Pascal**: It’s:

$$\sum_{n=1}^{\infty} \frac{n}{2^n \times (10^6 \ \times (n + 1{,}000))}$$

That sum is equal to about $\approx 1 \times 10^{-9}$. This is obviously much lower than the expected value of keeping the 10 livres! So, it is definitely not worth me giving you my wallet.

**Mugger**: [Motions with hand, another person appears in the alleyway.]

**Pascal**: Who’s that? She is looking at me … menacingly.

**Mugger**: Oh, that's my Boss, also from the Seventh Dimension. You see, we have a system set up.

**Pascal**: A system? This is not sounding good.

**Mugger**: No, no, you will like this. It is good news for you! For problem cases like you — and actually, we've never had one quite like you before! — we are prepared to offer you a special deal. My Boss has powers *much* greater than mine. If I were going to offer you $x$ days of bliss, she has assured me [Mugger and Boss nod to each other] that there will instead be $2^x$ others just like me who are now silently promising to give you that same number of days of bliss. [Several others now walk into the alleyway.] Since presumably the probability of any of them coming through on the promise isn't any less than the probability of *me* coming through, you'll need to change your utility function to the following:

$$\sum_{n=1}^{\infty} \frac{2^n \times n}{2^n \times (10^6 \ \times (n + 1{,}000))}$$

[**Mugger continues**] … and notice that this utility function does not converge.

**Pascal**: Well, first of all, your story about your fellow people from the Seventh Dimension seems crazy. Given what we had discussed earlier, about you giving a high reward after I'd told you my odds, I'm not sure that this is fair.

**Mugger**: Well, all I asked for is odds that *I* would come through on any promise *I* might give you. Those are *other* people who might make you a promise.

**Pascal**: Well, that still sounds crazy, but I'll go along with you. The odds that you are serious about those other people from your dimension seems crazy, like one in 1,000 quadrillion.

**Mugger**: I can understand your hesitancy about that. Why don't you go ahead and divide the sum of the utility function by 1,000 quadrillion. Tell me now — what does that come to?

**Pascal**: [Thinks for a moment.] Whelp. That utility function still doesn't converge.

**Mugger**: That is correct.

* * *

## Part IV. In Which Pascal Thinks He Has Seen This Trick Before, and Mugger Begs to Differ

**Pascal**: OK, so you all seem to have me in a tricky spot. But, if I may say so, now you're just not being very *original* — and I would've expected more originality from someone from another dimension. I was just in St. Petersburg last week.[7] They had a coin flip game there. For however many times a coin comes up heads before the first tails — call that $n$ — one gets $2^n$ rubles as a prize.[8] Someone there convinced me, using expected value theory, that I could expect to get a super-high reward for playing. They convinced me that even though the odds of getting one of the better outcomes gets smaller and smaller, the outcomes themselves get exponentially better and better, and so my expected value from playing is super high. *Unlimited*, in fact. I put in a million rubles to get in the game, but I only came away with four. What is going on here now sounds the same as their game, and I don't want to get caught in that kind of thing again. Tiny odds for each grand possibility, but an infinity of them.

**Mugger**: I know that game, and it sounds like you made the right choice to get into it, though you got unlucky. Formally, it may look like what we are offering you is the same, and that is indeed also why you should also accept *our* offer (despite your bad luck last week). And I understand that you might think that now what we are doing isn't anything original. But if I may say so, what we are doing is a bit harder on you. Or, I should really say, *better* for you! Anyway, the St. Petersburg "house" took on more immediate risk to themselves than we are taking on — they had to provide a coin, and they had to show you that they were good for some money. Did you check that they had a lot of money before you agreed to it? And I'm assuming that they were honest, and would've given you the rubles if you had actually won more than you put in.

**Pascal**: Yes, actually, I did check, and they showed me some stacks of cash.

**Mugger**: But how could you be really sure that they could give you any amount of rubles with no limit, if your coin flips came out right? What we have is a different, ummm … a different opportunity for you.

**Pascal**: Hmm ….

**Mugger**: We all know that *money* has marginally decreasing value, and you have a limited amount of time on the planet, things that might in practice have swayed intuitions in St. Petersburg to not give them even *more* money up front.[9] We from the Seventh Dimension don't have nearly the same limitations in our negotiations — we can offer you anything you'd like. From your perspective, the *only* risk is the chance that I might not come through for you. Of course, you think that that chance is high, but it's your one and only risk, and that risk is *already accounted for* in your utility function.[10] In St. Petersburg, you had to risk *both* getting an unlucky

[7] It should be noted that the Pascal in this dialogue does not reside in the 17th century, when Blaise Pascal lived.

[8] See Peterson (2020).

[9] See Hájek (2014), and Hájek & Nover (2006).

[10] Pascal is not risk averse (Bostrom, 2009, p. 444). One might think that there is a risk in losing one's wallet, but there is also a huge risk in missing out on quadrillions of days of happiness, so *risk* aversion should not explain an aversion to giving up a wallet. However, *loss* aversion (Kahneman & Tversky, 1979) may explain intuitions in Pascal's Mugging (see also Schwitzgebel 2017, p. 283), but loss aversion

roll — which you did — plus *another* risk that the house might not come through for you if you did hit it big.[11]

**Pascal**: That sounds right.

**Mugger**: And the other big difference is that — and I don't mean to sound like a stalker — our bargain is ever-present in a way that the St. Petersburg game is not.

**Pascal**: What do you mean?

**Mugger**: Honestly, I *could* even make it an ongoing offer. I could tell you that, at any point in time, you could just wire us 1,000 livres and we'd promise to give you some fantastic number of days of bliss. You don't even need to be in a place with a coin flip system set up.

**Pascal**: Hmm …. Hmm ….

**Mugger**: [Smiles.] In fact, [*breaks the fourth wall and addresses reader*] *anyone* can simply *imagine* that I am secretly the next person that they will talk to, and they should give away their wallet to that person in exchange for a possible, unspoken, offer from my Boss and me that they can imagine might come true.

* * *

## Part V. In Which Pascal Revises His Probability Function

**Pascal**: Well, you might think you have me, but there is actually an easy fix. Sure, your offer is tough because you folks from the Seventh Dimension can offer me fantastical rewards at any point in my life, unlike St. Petersburg. But, *I* can also re-imagine my *own* probability function, also unlike the coin flips in St. Petersburg. Your Boss adding all of those other people from your dimension threw me for a loop. At first, I suggested that I lower my probability by a fraction. But I can resolve this if I just change my whole probability function for you.

**Mugger**: OK, what does that look like?

**Pascal**: Something like this. If you want to play the game of exponentially increasing unlikely-to-occur rewards, I can imagine any kind of odds function, including one with an exponent in the reciprocal, like this. The odds of you coming through with $x$ days of bliss, in the context of you also telling me that there are lots ($2^x$) of others who are also going to be promising me the same, is this:

may not be rational, and for the purposes of this article, we can stipulate that Pascal, as an adherent of classical decision theory, is not loss averse.

[11] See Jeffrey (1983, §10.2), who writes that "it is essential that we avoid the St. Petersburg paradox." He continues, "our rebuttal … consists in the remark that anyone who offers to let the agent play the St. Petersburg game is a liar, for he is pretending to have an indefinitely large bank." At least from Pascal's perspective, the Mugger in the present dialogue does not have that restriction, since all that matters is the fact that there is a chance that he might come through. He does not need to give any other evidence up front. Thus, the Mugger paradox is in this way a deeper paradox than St. Petersburg.

https://doi.org/10.1017/S0012217323000392 Published online by Cambridge University Press

$$\frac{1}{10^6 \times (2^x + 1{,}000)}$$

And, if I do the utility calculation on that, even including all your friends (but dividing now by one in 10 quadrillion), the utility will come out as something very tiny, approximately $1 \times 10^{-25}$. So no, you are definitely *not* getting my wallet.

* * *

## Part VI. In Which the Mugger Points Out That This Is *Ad Hoc*, and Pascal Sums up the Quadrilemma He Is In

**Mugger**: So, you think you are so smart? You might think that *more* math will save you, and formally, I confess that I *think* you can do it. At least, rather than ever handing over your wallet, you perhaps can keep engaging me in conversation by constantly updating your probability functions as I keep telling you about more and more of my Boss's exponentially more impressive powers.

**Pascal**: Maybe I would do that, or more likely I would just walk away!

**Mugger**: Well, let me ask you this. *Why* are you trying so hard to resist us this way? If I may say so, it is rather *ad hoc* to now posit this new function when earlier you seemed content with a different function. All I wanted at the start was an answer to the question of the odds of *my* coming through with the best offer that I might give you. You didn't want to answer that, and instead you gave me a probability function. But now you are giving me a different probability function. Maybe you should have thought of some of these possibilities earlier. What would *motivate* choosing one function rather than another?

**Pascal**: Well, I have the intuition that *all* your offers, even the one you made along with your Boss, aren't worth it, and that's how to make sense of that intuition.

**Mugger**: But what would *independently* motivate using the new probability function rather than your original function? Why not just give up on your intuition that our offer is a bad one? I started making these, umm … deals with people in public back almost 15 years ago and no one before *you* has ever objected to the idea that there is some non-zero lower bound on the probability of my fulfilling any promise that I might promise. And now, you first wanted to give me a simple reciprocal function, and then you put the number of days as an exponent in the denominator?

**Pascal**: Hmm.

**Mugger**: Also, you might not want to hear about more ways that my Boss and my friends might be able to give you even more rewards. What are you going to say next? We might be here a while, and I'm getting impatient.

https://doi.org/10.1017/S0012217323000392 Published online by Cambridge University Press

**Pascal**: OK, OK, let me think. I confess that I don't have a clear idea, independent of thinking about avoiding the utility function diverging, of what the probability function should look like.

**Mugger**: Aha! So, you admit that you are just trying to rationalize your resistance to giving me your wallet!

**Pascal**: Well, it's just not clear to me that what is happening here with your Boss and your fellow dimension-mates is not a more dressed-up version of what we were talking about earlier in our conversation. I thought that there was an implicit assumption when you specified that you would restrict your rewards to days of bliss that there weren't these other muggers or other crazy business. That's why I assented to giving you the probability distribution that I did. If I had known about your Boss, I wouldn't have.

**Mugger**: Now you're being unfair to me. First of all, nothing I have said about my Boss violates my initial deal with you. And you didn't really know what else was going to come along, did you? You gave me a probability distribution for some really great rewards. You seemed satisfied with that. But now that I told you about my Boss — whose possibility you might wish you had considered beforehand — you want to change your probability function.

**Pascal**: This is the kind of thing that I wanted to avoid getting into in the first place! Maybe I should have refused to give you a probability function, just as I refused to answer your first question.

**Mugger**: But that is also *ad hoc*. A probability distribution is just an assessment of the chances of a range of possible states of affairs. You should not be turned off from giving one just because you are worried about some promises some people in the street might make afterwards. What *other* states of affairs might you refuse to give probabilities for, and for what reasons? That is not a stable strategy.

**Pascal**: Hmm. Let me think about this. [Thinks.] I've tried four different kinds of responses to you, and to each, I must admit, you've had a good reply. Since I am an ardent devotee of classical decision theory, it seems that you have put me in the following quadrilemma. Either (1) I give you a function of the odds of you coming through with any promise that you might make that has a non-zero lower bound, in which case you and your associates will promise me such a big reward that, according to expected value theory, I should give you my wallet; or (2) there is some probability function that starts out lower than the possible rewards you might give me and decreases at the *same* rate as the rewards increase, in which case we have something similar to a St. Petersburg game (except a less risky, and omnipresent one) — and I should give you my wallet; or (3) I can propose a decreasing probability function that decreases exponentially faster than the possible rewards that you might give me increase, but then change my probability function depending on what you or your Boss say in response, but that would be deeply *ad hoc*; or (4) I can refuse

https://doi.org/10.1017/S0012217323000392 Published online by Cambridge University Press

to give you a probability function, but this refusal also doesn't have any independent motivation, and will make me wonder whether I can give probability functions for anything — and that of course is going to have some bad consequences.

**Mugger**: Yes, Yes. You got it!

**Pascal**: [Starts, hesitatingly, to reach for his wallet.] You really shouldn't get away with this!

**Mugger**: Very good, Pascal. Don't worry, it's for your own (expected!) good. [Reaches out hand; others approach closer.]

* * *

## Part VII. In Which a Stranger Arrives and Gives the "Many-Muggers" Defence

[A Mysterious Stranger emerges from the shadows of the alley.]

**Stranger**: Hold on just a moment, everyone. I've been observing what has transpired here. M. Pascal — look carefully. The "Boss" and the others there, how do we know for certain that they are working *with* the Mugger, and not at cross-purposes?

**Pascal**: What do you mean?

**Stranger**: [Addresses Mugger.] Let's say that Pascal does try to sincerely answer your original question and gives you a non-zero probability. Let's say that he tells you that the probability of your coming through with the best offer that you might give him is one in 10 quadrillion, and then you tell him that you are offering him *1,000* quadrillion days of bliss in exchange for his wallet. And imagine that he then *declines* your offer.

**Mugger**: Obviously, that would be a huge mistake, but I can imagine it. Humans often make mistakes like that!

**Stranger**: Well, not so fast. Bear with me. Let's say that, after you walk away dejected, your next "friend" there [points] comes up to Pascal. And your "friend" tells him that, precisely because he declined *your* offer, that *they* are just going to give him 1,000 quadrillion days of bliss.

**Mugger**: Outlandish! That won't happen.

**Pascal**: [Irritated, chimes in.] Really? I thought we are loath to give zero probability to metaphysically possible events.

**Stranger**: Indeed! So tell me: what are the chances, in your mind, that that might happen?

**Mugger**: [Pauses to think.]

**Stranger**: And I see your *other* friends. They all might have their own agendas, and perhaps their own offers for Pascal. There could be many muggers out here in this alley![12]

[12] Cf. Cargile (1966), Hacking (1972); also see Peterson (2018) for a related argument. Eliezer Yudkowsky, in his original formulation of Pascal's Mugging (Yudkowsky, 2007, the blog post upon which Bostrom's article is based), says that Pascal's Mugging is *not* resolved in this manner, though Yudkowsky ends by noting that he has not come up with a satisfactory resolution.

https://doi.org/10.1017/S0012217323000392 Published online by Cambridge University Press

**Mugger**: I resent that characterization! I was just trying to make a deal with M. Pascal that benefits both of us!

**Stranger**: Yes, I hear you. And I understand you not wanting to give *me* a single answer to *my* question.

**Mugger**: Right, fair's fair. Pascal didn't give me a single answer to my original question, after all!

**Stranger**: Ah, but can you give us a probability distribution for all of the various *other* offers that Pascal might receive (and have come through for him) if he *declines* your offer? Remember, expected value theory tells us that he shouldn't decide on an offer just on the basis of what happens if he accepts it, but also in consideration of what happens if he declines.

**Mugger**: I can't give you a probability distribution on the various outcomes — there are too many to think of. That's not a fair question to ask me!

**Pascal**: I would have thought that *you*, of all people, claiming to be from the Seventh Dimension, would be able to answer that!

**Stranger**: If you don't want to give us a probability distribution, can you give us a *lowest bound* on the odds of some really wonderful other things that Pascal might hope for and might happen if he declines your offer?

**Mugger**: Hmm.

**Stranger**: And is that low bound any lower than the one Pascal might've given for the likelihood that *you'll* come through for him?

**Mugger**: Well, *I* am here, and he already has my word that I will give him all of those days of bliss. Those other possibilities about my friends are just idle speculations.

**Pascal**: [Picks up on Stranger's line of thought.] You have indeed made the possibility of your giving me 1,000 quadrillion days of bliss decision-theoretically *salient* to me ….

**Mugger**: What does that mean?

**Pascal**: A possibility is decision-theoretically salient when one has taken account of the possibility in one's decision structure. And it doesn't even mean that I brought the possibility to mind individually. It just means that it has entered my utility calculations, perhaps as one of a very large series of possibilities to which I have given a probability distribution function. Of course, if a possibility (or a set of possibilities) is not *psychologically* salient to one, that could affect whether one takes account of it in one's decision making — it could affect whether it is decision-theoretically salient.

**Mugger**: OK, go on.

**Pascal**: Now, I confess, I just started speculating about how your *friends* might help me out a minute ago when this Mysterious Stranger came over. But I don't think that the possibility of them coming through with a great deal for me (conditional on my *declining* your offer) is *any* less likely than *you* coming through for me with a great deal (conditional on my *accepting* your offer).

**Mugger**: Well, that's harsh! Doesn't my giving you my word give you some greater credence in me, Pascal?

**Pascal**: Earlier, I said that I was torn between giving an always-decreasing probability function, and one that had a low bound. What might justify having a low bound? Maybe because the story about you coming from the Seventh Dimension and giving me something really great has about as much chance as *any* scenario that undermines what I take to be the laws of nature and my basic experience. And your being from the Seventh Dimension is like that.

**Mugger**: But it wouldn't be rational to give the *same* low probability to all of these scenarios. Some of them are ….

**Stranger**: Yes, conjunctions of others, and thus should have even lower odds, equal to the product of the two — assuming that they are independent.

**Mugger**: Yes.

**Pascal**: But, bending the laws of the universe for two scenarios means that they are not independent of the other, even if the laws of the universe are bent in different ways for the two scenarios ….

**Stranger**: [Interrupts.] Pascal, let me help you out. You seem inclined to say that there is a set of possible outcomes that are such that whatever you do, each outcome in the set is equally likely to occur no matter what you decide to do. And I can see how that would be motivated by the fact that, for many scenarios, the probability hits some kind of lower limit. But our Mugger friend is correct that that would violate principles of aggregating the probabilities of outcomes.

**Mugger**: Aha!

**Stranger**: But Pascal can put things differently. Pascal can say that, for any decision, there is a set of unlikely outcomes that are such that, for each of them, there is a member of another set of unlikely outcomes with the same probability and the same reward. And because of that, we can ignore all of the outcomes in both sets.

**Pascal**: Are they not the *same* set?

**Stranger**: Well, if you give the Mugger your wallet, he might take it home, and then it might grow wings and flap them and land in the sink. The chances of it landing in his sink will be higher if you give it to him than if you keep it. But, the chances that it would fly into *your* sink if *you* keep it is, for all we can tell, the same as the chances of it flying into his sink if he took it home.

**Mugger**: I think *his* sink is probably bigger — business has been slow for me the last couple of years!

**Stranger**: That may be. But *all in all*, the total expected value of all of these unlikely but possible things happening, in one of the sets, is going to be equivalent to the total of the other set, whatever you do. And, if not, then maybe he *should* take your offer. But that's not the case here.

**Mugger**: *I'm* not sure of that!

**Stranger**: Perhaps, in some objective sense, some of these possibilities are more likely than others. But, from Pascal's *own* position, there is no reason for him to assign a greater overall expected value to the members of one set (collectively) rather than to the other.

**Mugger**: Hmm ….

**Pascal**: That sounds right. So, no deal.

**Mugger**: [Scowls.]

* * *

## Part VIII. In Which the Stranger Clarifies That We Should Not *Discount* in General and That Classical Decision Theory Is Preserved

**Mugger**: So, what you're saying is, you are going to *discount* the expected value of my coming through on my offer to you down to zero?[13] In other words, you are not even going to include, *at all*, the value of the possibility of my coming through for you in your utility calculation of accepting my offer because you take the probability to be very small? For several reasons, discounting is irrational.[14]

**Pascal**: No, no, I agree — discounting in most cases is wrong. That's why I was so interested in your offer in the first place.

**Stranger**: You should have been more wary, Pascal!

**Pascal**: If I get in my car and emit some greenhouse gases, there is a slight risk that that will make the world worse. (There is also a slight chance that it will make the world better, but there is more of a chance that it'll make things worse.) And I shouldn't discount that slight marginal risk. That's why there are reasons not to go for frivolous drives.[15]

**Mugger**: Nice to hear that you care about the environment. But I saw you driving earlier today!

**Pascal**: I only do it because other benefits from my driving (like getting to my job) outweigh the expected harms. (Or so I hope. Maybe I'm a bit selfish in considering my own goods above the harms to others. But that's a topic for another day.)

**Mugger**: Ok, I'll grant you that. But let me get this straight. You are now discounting the small chance that *my* offer will be a good one, but you don't think discounting in general is acceptable?

**Pascal**: Yes. But perhaps we can distinguish the tiny probabilities we *can* rationally discount from the ones we can't.

**Stranger**: The reason for discounting is not the tininess of the probability or the tininess of the expected value (which might not actually be tiny!) itself. Instead, when a new possibility becomes salient, one can discount it if

[13] See, e.g., Kosonen (2021), Monton (2019), Schwitzgebel (2017), and Smith (2014) for discussions of discounting small probabilities.

[14] See Barnett (2018), Barrington (ms.), Isaacs (2016), and Wilkinson (2022) for detailed arguments. We agree that the kind of discounting discussed in these articles is irrational.

[15] *Contra* Kingston & Sinnott-Armstrong (2018) and Sinnott-Armstrong (2005), and in accord with Broome (2019) and Hiller (2011).

https://doi.org/10.1017/S0012217323000392 Published online by Cambridge University Press

one realizes that there exists a *symmetrical* previously non-salient possibility that cancels out the expected value/disvalue from the one you are considering.[16] Or, if it's a *series* of tiny possibilities that become salient, then one can discount them if one notices that there is a previously unrecognized, symmetrical series with which one can form a one-to-one function to the newly salient one. And, with a tiny probability like yours, there are basically an unlimited number of symmetrical possibilities. Your trick was that you made the remote possibilities on *one* side of the choice salient, but in a way that kept the symmetrical remote possibilities on the other side non-salient. Discounting, when rational, has just as much to do with the tiny probabilities that are salient parts of the option being considered in the choice as it does with previously non-salient tiny probabilities that are parts of *other* options in the choice.

**Mugger**: OK, but can't one partition *any* large probability into a bunch of tiny ones?[17] For instance, you might reasonably worry that I am about to knock you out in a few seconds to *steal* your wallet, but maybe you shouldn't worry: it might happen 10 plus one quadrillionth of a second from now, 10 plus two quadrillionths of a second from now, etc., and each of those has a vanishingly small chance. Perhaps even smaller than what you think the chances are of me coming through on any promise I might make.

**Pascal**: But there is a smaller *overall* chance of my getting knocked out if I move away from you. So I am not going to discount those possibilities, or ones in the climate change case.[18] [Puts wallet back in pocket and quickly takes several steps back.]

**Mugger**: Wait. Even if we grant that there is a smaller overall chance that you'll get knocked out if you move away, that doesn't satisfy me because each of the individual possibilities would still be negligible. And you can't just say that they are non-negligible only when, if they are combined with other possibilities, they sum together to be significant. That would violate a criterion of independent assessment of possibilities.

**Stranger**: Yes, Pascal, our colleague here is right. But what you could have said is that the value of *each* of the possibilities of you knocking him out (at

[16] See Schwitzgebel (2017, p. 283) on symmetry: "In general, I'm not inclined to think that my prospects will be particularly better or worse due to their influence on extremely unlikely deities, considered as a group, if I [perform some action] than if I do not."

[17] See Barrington (ms., §2), Hiller (2011, p. 355).

[18] Balfour (2021) argues that Pascal's Mugging-type considerations show that according to expected utility theory, one ought to dedicate all one's decisions, including very minute ones, for the sake of long-term human wellbeing. (This is Yudkowsky's ultimate concern as well.) The considerations in this article rebut Balfour's argument only insofar as the tiny risks to future people are indeed symmetrical to hitherto non-salient scenarios. For reasons that are beyond the scope of this article, we, in fact, think that at least some of the risks discussed by Balfour *are* symmetrical in that way, and thus we do not accept Balfour's conclusion. Furthermore, Balfour's (2021, pp. 123–124) Mugger suggests that if we must change our lives so significantly due to these longtermist concerns, it would instead be a reason to abandon expected utility theory. Again, we disagree, but will reserve discussion for another occasion.

https://doi.org/10.1017/S0012217323000392 Published online by Cambridge University Press

time 10+*n* quadrillionths of a second) is *lower* if you move away than if you stay close to him. That makes it different from the consideration of possibilities that grandly violate what we take to be the causal order of things. There is no symmetrical pair of knock-out possibilities depending on which choice you make (of moving away or standing still).

**Pascal**: OK.

**Mugger**: So, is *this* the only case where you are going to discount the tiny probability? Do you really think that little of my honesty?

**Pascal**: Well, there are pragmatic constraints built into how we apply decision theory in the real world that need to be looked at on a case-by-case basis. In the case of you and your friends, if you (or anyone else) is, according to you, always likely to give people days of bliss in exchange for their wallets, it gets me thinking, with our Stranger, that there is an equal potentially ever-present possibility that everyone can hold onto their wallets while still getting the rewards. So, I admit that I think little of your honesty. But the point here is not that your giving me quadrillions of days of bliss has a tiny chance of some hard-to-determine amount; the point is that, now that you mention that possibility, it seems just as likely to me that I will get that bliss whether or not I give you my wallet. Everything that happens has a miniscule subjective chance of happening, if it is described in fine enough detail. But not everything that happens is symmetrical to scenarios like what you are talking about.

**Mugger**: My Boss isn't going to like this. [Boss nods in agreement, but disapprovingly.]

**Stranger**: And about your Boss: as Bayes tells us, the probability of the hypothesis that your Boss is limited in power just by the laws of metaphysics, given the evidence that both (A) you are telling Pascal this, and (B) you and your Boss are here in this alleyway trying to get Pascal to give you his wallet, shouldn't just be considered on its own, but varies in proportion to the chances that (A) and (B) would happen *given* that your Boss really is limited in power just by the laws of metaphysics. And the chances of *that* are, again I'm afraid, no greater than the chances of Pascal's getting bliss from not giving up his wallet. Most likely, a Boss with such expansive powers wouldn't need to bother with a trifle like Pascal's wallet![19] So,

[19] The probability that the Mugger has transdimensional superpowers of such a great extent so as to give Pascal enormous numbers of days of bliss, given that he is trying to mug Pascal is:

$$\frac{\mathrm{P}\ (\text{mugging}\ |\text{superpowers})\ \times \mathrm{P}\ (\text{superpowers})}{(\mathrm{P}(\text{mugging}\ |\ \text{superpowers}) \times \mathrm{P}(\text{superpowers})) + (\mathrm{P}(\text{mugging}\ |\ \text{no superpowers}) \times \mathrm{P}(\text{no superpowers}))}$$

And, in addition to the *prior* probability of the Mugger having transdimensional superpowers being very small, the probability of the Mugger trying to mug Pascal conditional on the Mugger possessing transdimensional superpowers is intuitively much lower than the probability of the Mugger trying to mug Pascal conditional on his lacking superpowers. In other words, the very fact that the Mugger is saying both that he has superpowers and wants Pascal's wallet is evidence that the Mugger does *not* have superpowers. If so, then contrary to what Bostrom (2009, pp. 444–445) and Yudkowsky (2007) suggest, the Mugger's testimony [*breaks the fourth wall and addresses reader*] no one needs to give away their wallet to the next person they see, because even if they *were* going to get some days of bliss from doing so, they most likely wouldn't be hearing about it from an imagined Mugger in the pages of a philosophy journal. They're just as likely (or unlikely) to get bliss from going about their ordinary business. And so, there is nothing wrong with how classical decision theory handles any of these cases.

* * *

## Part IX. In Which Pascal and the Stranger Take Their Leave

**Mugger**: So, you're really not going to give me your wallet after all?

**Pascal**: No. But it was a … an informative time with you both.

**Stranger**: The beauty and danger of your grift consists in the fact that anyone at any time can promise, in the shadiest of ways, an unlimited possible reward in exchange for a finite up-front expense, and the product of this utility will look amazing. Pascal was a perfect mark for you, given his attraction to this kind of argument. But your argument applies even more broadly than just to people who are enamoured by thoughts of an ever-powerful God. It could appeal to anyone with any imagination plus a knowledge of the rudiments of expected value theory. The central trick/flaw in the strategy is that it focuses on the option that is salient, and not on all of the other options that are not salient, but ought to be.

**Pascal**: I think I might head back to St. Petersburg — I have a few things to say to those folks about their game. The more I think about some of the tiny probabilities and huge rewards that they mentioned, the more I wonder whether, at a certain point, I am just as likely to get the reward regardless of how the coins flip. If I were to flip a thousand or even just fifty heads in a row, I would be more worried that I was crazy or that I was in a simulation with at least one glitch (and likely more to follow!) than excited about my winnings. Well, that's something I'll have to talk over with them.

**Acknowledgements.** Thanks to two anonymous referees for helpful comments. Thanks also to Karen Gross for a discussion of the "mysterious stranger" motif and to Nicolas Delon for help with the French translation of the abstract. This article originated from a discussion of Pascal's Mugging on the Board Certified Epistemologists Facebook group.

**Competing interests.** The Authors Declare None.

## References

Armendt, B. (2019). Causal decision theory and decision instability. *Journal of Philosophy*, *116*(5), 263–277. https://doi.org/10.5840/jphil2019116517

Balfour, D. (2021). Pascal's mugger strikes again. *Utilitas*, *33*(1), 118–124. https://doi.org/10.1017/S0953820820000357

should not increase Pascal's credence in its content. (It should also be noted that, in Yudkowsky's formulation, the Mugger explicitly claims to have powers from outside "The Matrix.")

Barnett, Z. (2018). No free lunch: The significance of tiny contributions. *Analysis*, *78*(1), 3–13. https://doi.org/10.1093/analys/anx112

Barrington, M. (ms.). Ignoring the improbable. Available at https://www.mitchellbarrington.com/research/

Bostrom, N. (2009). Pascal's mugging. *Analysis*, *69*(3), 443–445. https://doi.org/10.1093/analys/anp062

Broome, J. (2019). Against denialism. *The Monist*, *102*(1), 110–129. https://doi.org/10.1093/monist/ony024

Cargile, J. (1966). Pascal's wager. *Philosophy*, *41*(157), 250–257. https://doi.org/10.1017/S003181910005871X

Hacking, I. (1972). The logic of Pascal's wager. *American Philosophical Quarterly*, *9*(2), 186–192. https://www.jstor.org/stable/20009437

Hájek, A. (2014). Unexpected expectations. *Mind*, *123*(490), 533–567. https://doi.org/10.1093/mind/fzu076

Hájek, A., & Nover, H. (2006). Perplexing expectations. *Mind*, *115*(459), 703–720. https://doi.org/10.1093/mind/fzl703

Hiller, A. (2011). Climate change and individual responsibility. *The Monist*, *94*(3), 349–368. https://doi.org/10.5840/monist201194318

Isaacs, Y. (2016). Probabilities cannot be rationally neglected. *Mind*, *125*(499), 759–762. https://doi.org/10.1093/mind/fzv151

Jeffrey, R. C. (1983). *The logic of decision* (2nd edition). University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/L/bo3640589.html

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–292. https://doi.org/10.2307/1914185

Kingston, E., & Sinnott-Armstrong, W. (2018). What's wrong with joyguzzling? *Ethical Theory and Moral Practice*, *21*(1), 169–186. https://doi.org/10.1007/s10677-017-9859-1

Kosonen, P. (2021). Discounting small probabilities solves the intrapersonal addition paradox. *Ethics*, *132*(1), 204–217. https://doi.org/10.1086/715276

Monton, B. (2019). How to avoid maximizing expected utility. *Philosophers' Imprint*, *19*, 1–25. https://quod.lib.umich.edu/p/phimp/3521354.0019.018/--how-to-avoid-maximizing-expected-utility?view=image

Peterson, M. (2018). The anti-nihilist wager. *Dialectica*, *72*(4), 597–602. https://doi.org/10.1111/1746-8361.12254

Peterson, M. (2020). The St. Petersburg paradox. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2020 ed.). Stanford University. https://plato.stanford.edu/archives/fall2020/entries/paradox-stpetersburg

Schwitzgebel, E. (2017). 1% skepticism. *Noûs*, *51*(2), 271–290. https://doi.org/10.1111/nous.12129

Sinnott-Armstrong, W. (2005). It's not *my* fault: Global warming and individual moral obligations. In W. Sinnott-Armstrong & R. Howarth (Eds.), *Perspectives on climate change* (pp. 221–253). Elsevier.

Smith, N. J. J. (2014). Is evaluative compositionality a requirement of rationality? *Mind*, *123*(490), 457–502. https://doi.org/10.1093/mind/fzu072

Wilkinson, H. (2022). In defense of fanaticism. *Ethics*, *132*(2), 445–477. https://doi.org/10.1086/716869

Yudkowsky, E. (2007, October 19). Pascal's mugging: Tiny probabilities of vast utilities. *Less Wrong*. https://www.lesswrong.com/posts/a5JAiTdytou3Jg749/pascal-s-mugging-tiny-probabilities-of-vast-utilities

**Cite this article:** Hiller, A., Hasan, A. (2023). How to Save Pascal (and Ourselves) From the Mugger. *Dialogue* 1–17. https://doi.org/10.1017/S0012217323000392

https://doi.org/10.1017/S0012217323000392 Published online by Cambridge University Press