Waging War on Pascal’s Mugger[1,2]

Patrick Kaczmarek

Abstract. Fanatics judge a lottery with a tiny probability of arbitrarily high value as better than the certainty of some modest value, and they are prone to getting swindled. You need only make the lie “big enough” to get one over on them. I put forward an elegant solution to the fanatic’s problem. When coming to a fully rational decision, agents may ignore outlandish possibilities.

1 | Introduction

Consider two cases:

> *Mugger*. For my pocket change, a shifty man promises to cast a magic spell that prevents the deaths of a quadrillion people in an alternate dimension.[3]
>
> *Moonshot*. A warlord is withholding medicine that would save the lives of millions of children. For my pocket change, she promises to release the medicine if it snows in Los Angeles at dawn in seven days’ time.

I know what I would do in *Mugger*. I would slowly back away from the man and never give it a second thought.

It isn’t that I lack compassion. I deny neither that it is good to save lives nor that there is a moral duty to perform easy rescues. I am not exercising my agent-centered prerogative. I am aware that there is *some* chance that the dodgy stranger is speaking

[1] An allusion to (Hájek 2003).

[2] Only after submitting this paper for review was I made aware, by an anonymous reviewer, that Martin Smith has just published a similarly motivated paper, Decision Theory and De Minimis Risk (forthcoming at Erkenntnis). We both propose a non-probabilistic criterion for discounting possibilities—one that is designed to capture something like ‘farfetchedness’. However, there are important differences. For example, as I read him, Smith and I would disagree about the rational permissibility of discounting the probability of saving millions of children’s lives in *Moonshot* to zero. More so, my account, I suspect, better handles uncertainty about what counts as ‘normal’ (see discussion at §2.b). Although I hope to revise the paper at some point, and explore what (dis)advantages my account has over Smith's, I am sharing the paper in its current state in case it proves to be useful fodder for those engaging the topic. For helpful discussion and comments, thanks to Niels Brögger, Tim Campbell, Michael Plant, Andrew Warren and Haydn Wilkinson. I am grateful to the anonymous reviewer and Ergo’s editor for the kinds words despite the bad news.

[3] Adapted from (Bostrom 2009). Unlike Bostrom, my case does not proceed by first asking you to determine the probability the mugger will stick with his promise *independently* from what he promises to do.

truthfully (it isn't impossible, however hard to believe).[4] Expected value theory demands that I give the mugger my pocket change.[5] But I won't, and I don't take this to be a rational failing.

I also know what I would do in *Moonshot*. I would play the warlord's game. Indeed, I would claim that any minimally considerate individual would play her game.[6] Even though the odds it will snow at dawn in Los Angeles in seven days are extremely slim, the deaths of millions of children is unspeakable; the loss of my pocket change is dwarfed by the expected goodness of saving the lives of these children.

Many people, I reckon, feel the same about *Mugger* and *Moonshot*.[7] However, this pattern of judgments is in tension. Both gambles involve arbitrarily low downside and a tiny probability of something marvelous. Expected value theory instructs us to treat them alike. That is, it says it is rationally impermissible for me to keep my pocket change if the offer's expected value is greater than the expected value of any available alternative in the choice-scenario, where a given act's expected value is calculated by multiplying the payoff of each potential outcome by its probability and then adding those values together. To simplify, let's stipulate that the expected values of the offers in *Mugger* and *Moonshot* are the same. *So why is it rationally impermissible to turn down the warlord's offer but rationally permissible to reject the mugger's offer?*

There has been, so far as I can tell, no serious attempt in the literature to answer this question. Although considerable ink has been spilled trying to avoid the notorious 'Pascal's Mugging', most of the discussion has been directed at fanaticism, where one is said to be fanatical if he judges a lottery with tiny probability of arbitrarily high value as better than the certainty of some modest value (Wilkinson 2022, 447; cf. Russell 2023).[8] One side argues that outcomes with small enough probabilities should be ignored.[9] Another maintains that probabilities can never be rationally neglected. Both assume fanaticism and muggings stand or fall together.

But as *Moonshot* and *Mugger* are meant to illustrate, tying their fates together is worrisome. Those who ignore small enough probabilities will pass on long shots worth taking, and blithely accept gambles that involve arbitrarily high risk for arbitrarily little reward that intuitively they shouldn't (Isaacs 2016). Meanwhile, the fanatic's vulnerability to

[4] I assume that one's credences about the fantastic should not be represented by a non-zero hyperreal (Easwaran 2014).

[5] See (Bostrom 2009) for that argument. Like him, I deliberately avoided introducing infinities. If retold with infinities, any mixed strategy might end up being rationally permissible in *Mugger* (Hájek 2003; cf. Chen and Rubio 2020, §4.1).

[6] Following Caspar Hare, I take it that a moral agent is minimally considerate when, for any states of affairs S, S*, she takes the consideration "you are better off in S than in S*" to be a reason to favor S over S* (Hare 2013, 23).

[7] A few people that I spoke with felt it was rationally permissible to turn down both of these offers; however, they still found something extra counterintuitive about a rational requirement to accept the offer made in *Mugger*.

[8] The terms 'Pascal's Mugging' and 'fanaticism' were introduced into the literature, respectively, by (Bostrom 2009) and (Bostrom 2011). The latter has also been referred to as 'recklessness' (Beckstead and Thomas 2023).

[9] For instance, (Monton 2019; Kosonen 2021). Smith (2014) defends the weaker position that rational actors are rationally permitted, but not rationally required, to discount tiny probabilities to zero.

muggings is embarrassing; indeed, Nick Bostrom intended Pascal's Mugging as a *reductio ad absurdum* of fanaticism.[10]

I will argue in this paper that fanatics are no more threatened by muggers than the rest of us.

The solution to the puzzle lies in an old observation. It has long been recognized that there is an important, even obvious, difference between *Mugger* and *Moonshot*: *only the former stretches reality into fantasy*. The trouble has been utilizing this observation. Specifically, we have thus far treated the fantastic as if it were little more than extremely unlikely. No wonder, then, that muggers need only to make their lies "big enough" to get one over on the fanatic. A large enough reward could, in the eyes of a fanatic, compensate for bad odds.

2 | Possible Worlds

We require a more nuanced take on the fantastic. Probability alone cannot discriminate muggings from the unremarkable and yet highly unlikely. Duncan Pritchard (2015) identifies the missing ingredient in the following passage:

> Although in general close possible worlds will tend to be worlds where high-probability events occur, and far-off possible worlds will tend to be worlds where low-probability events occur, there are exceptions. In particular, there can be close possible worlds where very-low-probability events occur—that is, where such events are easy possibilities, even despite their low odds of obtaining (Pritchard 2015, 443-4).

I will start by unpacking possible worlds and how I am thinking about them in the context of lotteries.

2.a. Presumed distance from the actual world

Possible worlds are ways the world could have been. Some possible worlds are different from the actual world in minor ways, such as the possible world where my left leg is crossed over my right leg rather than the other way around. These are close possible worlds. Possible worlds are farther away from the actual world the more change to the actual world that is required for them to obtain (Lewis 1987).

When analyzing the truth of counterfactuals, the closeness (similarity) of a possible world is determined in reference to the actual world (Lewis 1973, §4.2; Stalnaker 1984).[11] In cases of chance, however, I am unsure which outcome in a lottery describes the actual world. While a fair coin is in flight, for example, I cannot know whether W_{heads} or W_{tails} is the actual world.

[10] More precisely, Bostrom's target was standard expected value theory, which he takes to be committed to fanaticism.

[11] Which possible world semantics should be used for measuring closeness is far from settled (Fine 1975; Lewis 1979; Veltman 2005). I acknowledge this presents a substantial limitation to the project I'm developing in this paper; my discussion of degrees of change to a model, below, inherits many of these concerns.

What can I say about W_{heads} under these conditions? I can speak to the likelihood that W_{heads} is the actual world. For instance, this could be based on the frequency of heads observed in the past when this particular coin was tossed. There is also something that could be said about how close W_{heads} comes to my model of the actual world; that is to say, my best understanding of the laws of nature, history and what the past has set in motion, location of celestial bodies and so forth.[12] This description can be more or less fine-grained, and possible outcomes can fit a model to varying degrees. Some possible outcomes are a natural fit on that model, while others put pressure on it, and if these events were to obtain, then they would force me to revisit, even to throw out, my preferred model. A possible world's *presumed distance* from the actual world is the degree of change it implies to my model.

As Pritchard reminds us, these two things sometimes come apart. Imagine if I were to toss a fair coin ten times. Despite being extremely improbable, a coin landing heads ten times in a row could still *very easily* occur. The coin must simply land one way rather than another, as coins tend to do when thrown in the air. Nor is there anything far-fetched about a coin landing on the same side several times in a row. By this I mean there are no great changes that need to be made to my understanding of the world for the target outcome to occur.

We have our sophisticated take. At first pass, what makes it rationally permissible for me to reject the offer in *Mugger* is that the target outcome is realized in a far-off world, given that interdimensional magic constitutes a significant change to my best understanding of the world.

Specifically, I submit the following principle.

> Far-Enough to Ignore: For any lottery featuring in any decision problem faced by an agent, she may ignore worlds whose presumed distance is great enough in coming to a fully rational decision.[13]

Before moving on to consider problems for this principle, I want to clarify what I mean by 'significant changes'.

What I do *not* mean is that an event is "far enough away" only if it obtaining implies the majority of my particular beliefs are mistaken. For example, there is a possible world where invisible goblins, rather than gravity, pull objects down to the ground. Many of my current beliefs would remain intact if these otherwise harmless goblins were discovered. Rocket ships would still need jet fuel to achieve escape velocity, bees would still pollinate flowers and so on. Little would change in terms of how I navigate the world around me. Still,

[12] Of course, that understanding could be horrendously flawed; certainly, it has been in the past. For example, it was once widely believed there was a real danger of falling off the edge of the world while sailing if one strayed too far from the coastline. We too are bound to get egg on our faces. Nevertheless, hardly anyone denies that it is rationally permissible to act in accordance with one's false and gappy beliefs (Muñoz and Spencer 2021, 77); especially if the agent has the credences they ought, epistemically, to have (Sepielli 2017).

[13] This formulation takes its inspiration from Nick Smith's 'Rationally Negligible Probabilities'. The key difference is that his principle is couched in talk of probabilities rather than possible worlds (Smith 2014, 472).

if there really were invisible goblins, I would be fundamentally confused about the nature of my world.

Recall, a possible world's presumed distance from the actual world is the degree of change it implies to my understanding of the world. Following David Lewis' suggestion, I'll now add that the most significant changes to that understanding concern the laws of nature (Lewis 1979, 472).[14] Although many of my particular beliefs are consistent with invisible goblins, their discovery would imply that my model of the actual world was mistaken about the laws, and so the possible world where invisible goblins play the role of gravity is very far away, presumably.

However, I do not think contradicting my best understanding of the laws is essential to the outlandish. For example, there is a possible world in which I am a brain in a vat and the overwhelming majority of my particular beliefs are mistaken. There is no dog sleeping at my feet, no radio streaming Spotify and so on. If we grant that this possibility is consistent with the laws, then I would not be fundamentally confused about how the world works. Rather, I would be mistaken about my personal history and location in spacetime. I would be wrong about less fundamental things but still wrong about an awful many important things.

2.b. Refining the principle

There is a complication.

Just as someone can be uncertain about which world is actual, they can be uncertain between multiple models of the world. For example, String Theory and Loop Quantum Theory are differing explanations of quantum gravity. If String Theory forms part of my model of the world, then many possible worlds consistent with Loop Quantum Gravity involve "big, widespread, diverse violations of law" (Lewis 1979, 472). Because this upsets my model, possible worlds consistent with Loop Quantum Gravity will be designated as very far-off.

The similarity of possible worlds depends on which of these two theories goes into my model. Yet, neither String Theory nor Loop Quantum Theory has a clear upper hand in my beliefs. And so, if there is uncertainty about which theory to incorporate into my model of the actual world, then, because the similarity ordering of worlds depends on my model, there is uncertainty about which possible worlds qualify as outlandish. Nor is quantum gravity the only open question concerning the nature of the actual world; there are many others. This suggests that my uncertainty about which possible worlds are very far-off could be considerable.

This itself presents little problem. We can measure the *expected* distance of possible worlds by multiplying across the models in which one finds purchase and devotes some of her credence. To illustrate, suppose a given event E_1 is inconsistent with a law of nature if Quantum Loop Theory were true. As such, E_1 occurs in a far-off possible world. To make this concrete, let's arbitrarily assign it a distance of 100 conditional on Quantum Loop Theory. Next, let's take stock of being equally torn between String Theory and Quantum Loop Theory, and that E_1's distance from the actual world on the former is 1. If String Theory and Quantum Loop Theory are the only two theories that I am torn between, then the expected distance of E_1 from the actual world is 50.5.

[14] I leave open how we define 'law of nature'. For discussion, see (Lewis 1973; Armstrong 1978; Maudlin 2007).

However, if we accept Far-Enough to Ignore, then it might seem as if my uncertainty between String Theory and Quantum Loop Theory implies that I can ignore E_1 in my practical deliberations. After all, halfway to outlandish still seems pretty outlandish. This *is* problematic. Suppose E_1 is the destruction of the world with a device that I made in my garage. If Quantum Loop Theory is true, turning on the device will do nothing. It seems wrong to turn on the device, given my belief there is a 50/50 chance of String Theory being true.

There is a further problem. If it is rationally permissible to ignore far-enough away worlds, then it could be rationally permissible to bet the farm on an event with arbitrarily low probability of occurring.

Suppose that there are three possible events and you are uncertain between three models. According to Model 1, E_1 has a distance of 1 and the alternatives are arranged linearly with a distance of 99 from E_1 to E_2, a distance of 100 from E_2 to E_3 and a distance of 199 from E_1 to E_3. According to Model 2, E_3 has a distance of 1 and the possible worlds are arranged linearly with a distance of 99 from E_3 to E_2, a distance of 100 from E_2 to E_1 and a distance of 199 from E_3 to E_1. Meanwhile, according to Model 3, E_2 has a distance of 1 and both E_1 to E_3 have a distance of 199 from E_2. Suppose you have 0.02 credence in Model 3, 0.49 credence in each of Model 1 and Model 2 and furthermore that the cutoff for permissibly ignoring a possible outcome is 100.

The expected distance of E_1 is 102.49. E_3 has the same expected distance as E_1 does. Meanwhile, E_2 has an expected distance of 98.2.[15] Applying Far-Enough to Ignore, it is rationally permissible to ignore E_1 and E_3, and thereby to treat E_2 as if it were sure to be the case, even though you assign the model that supports this event happening a probability of 0.02. This seems bizarre.

What the above cases highlight is just how the strange the real world is (or at least, appears to be). Much of what gets discussed seriously in modern physics is wild. More so, the theories meant to elaborate on that weirdness often violently disagree about fundamental matters. Indeed, if Quantum Loop Theory were proven true by physicists tomorrow, then in relatively short order String Theory would be viewed, not simply as mistaken, but absurd; something that made us blush for ever having believed it in the first place.[16]

We need to account for this phenomenon, and we do so by recognizing that our definition of ‘outlandish’ cannot be completely divorced from probability. Although E_1 and E_3 *might be* very far-off possible worlds, we have significant credence in the models supporting them. Just as we originally observed that it doesn’t seem rational to neglect extremely unlikely events that happen in close possible worlds, it seems we shouldn’t ignore potentially far-away worlds supported by models in which we have significant credence. Although E_1 obtains in a possible world that is farther away than W_{heads} in expectation, it is nonetheless within the range of what should be tolerated. And it falls in that range precisely because the supporting model isn’t improbable.

What falls outside of this range? To my mind, a truly outlandish event is an event that obtains only in models that I scarcely believed (and which together receive a tiny portion of my credence) and might not have placed any stock in if not for the injunctions of

[15] Expected distance of E_1 (/E_3) is 1*(0.49) + 200*(0.51) = 102.49. Expected distance of E_2 is 1*(0.02) + 100*(0.98) = 98.2

[16] Much like happened with phlogiston.

practical rationality.[17] By contrast, although we have comparable credence in the target outcomes of *Mugger* and *Moonshot*, the latter is a rare event but obtains in high-credence models.

I take this to be the true mark of the outlandish, and how each of us immediately distinguishes *Mugger* from run-of-the-mill long shots, such as winning big on roulette and guessing the weather.

2.c. Restating the principle

We can restate my proposed principle as follows.

> Outlandish Possibilities are Negligible: For any lottery featuring in any decision problem faced by an agent, she may ignore outlandish outcomes in coming to a fully rational decision.

Let's return now to *Mugger* and *Moonshot*.

Each of these target outcomes is extremely unlikely. However, the successful rescue of a quadrillion people in an alternate dimension using magic is consistent only with low-credence models (which together I give little credence). Witnessing the dodgy man deliberately rescue people from death with a magic spell would compel me to scrap my best understanding of the world. By contrast, successfully thwarting the warlord's evil plans in *Moonshot* is consistent with the cluster of high-credence models that receives most of my credence.

Outlandish Possibilities are Negligible dictates that *Mugger* can be permissibly treated as a guaranteed loss of my pocket change. In other words, I may proceed as if I knew that none of the possible worlds where the shifty man was telling the truth was the actual world. The same cannot be said about *Moonshot*. There is nothing outlandish about it snowing in Los Angeles. Rather, weather is fickle and cold snaps do happen, even in Southern California. I must consult expected value as to whether to play the warlord's game.

Outlandish Possibilities are Negligible delivers the intuitively correct verdicts in *Mugger* and *Moonshot*, and furthermore it avoids the unsettling implications developed in §2.b.

3 | Concluding Remarks

There are two last points that I wish to address before closing.

First, as I have formulated Outlandish Possibilities are Negligible, an agent is rationally permitted to roll the dice if she is so inclined on gambles such as *Mugger*. A stronger restatement of this principle would condemn her choice to pay the mugger as irrational. I have no knock-down argument to give in support of either formulation. Instead,

[17] Andrew Warren points out this position commits me to denying the possibility of miracles on models. As they put it to me: small credence in magical powers is negligible at the level of models but not at the level of events. This is indeed how I am imagining models, as exceptionless hypotheses of how the world is arranged.

I will share that I myself lean towards the permissibility of taking chances on the fantastic.[18] After all, magic and interdimensional causation are still genuine possibilities (though I am reluctant to appeal, on this particular occasion, to the well-worn adage that "stranger things have happened").

Second, I do not claim that Outlandish Possibilities are Negligible makes you immune to exploitation. Clearly, the warlord in *Moonshot* could be lying. She could dump the medicine in the ocean whether or not it snows at dawn in Los Angeles seven days from now. Rather, my principle removes a particular weapon in the mugger's arsenal to which the fanatic was dangerously exposed: increasing the size of the promised reward until the lie is big enough to compensate for its implausibility (cf. Baumann 2009). Outlandish Possibilities are Negligible blocks this move, making it significantly harder to swindle a fanatic. Even if the mugger starts off with a mundane-sounding mechanism for delivering the reward, at some point she will need to make a claim that violates one's basic understanding of the world, since that understanding imposes a strict upper bound on what any individual can realistically do (cf. Balfour 2021). And once the mugger dips into the fantastic to sweeten the deal, the fanatic is permitted to treat the mugger's offer as what it almost certainly is: a barefaced lie.

If I am right, fanatics don't make easy marks. Accepting fanaticism doesn't leave one uniquely vulnerable to getting fleeced by scoundrels.

Bibliography

Armstrong, David (1978). *A Theory of Universals*. Cambridge: Cambridge University Press.

Balfour, Dylan (2021). Pascal's Mugger Strikes Again, *Utilitas* 33(1): 118-124.

Baumann, Peter (2009). Counting on Numbers, *Analysis* 69(3): 446-448.

Beckstead, Nick and Thomas, Teruji (2023). A Paradox for Tiny Probabilities and Enormous Values, *Noûs* 00: 1– 25. https://doi.org/10.1111/nous.12462

Bostrom, Nick (2009). Pascal's Mugging, *Analysis* 69(3): 443-445.

——————— (2011). Infinite Ethics, *Analysis and Metaphysics* 10: 9-59.

Chen, Eddy Keming and Rubio, Daniel (2020). Surreal Decisions, *Philosophy and Phenomenological Research* 100(1): 54-74

[18] I hesitate largely because of the following problem. Suppose you are offered two outlandish bets. They are equally unlikely but the first bet doubles the size of the reward of the second bet. Outlandish Possibilities are Negligible makes it permissible to opt for the second bet, since you are permitted to ignore the first outlandish outcome but needn't ignore the second outlandish outcome. Yet, opting for the second bet is irrational (*qua* violates statewise dominance). If one is going to take either bet, they should take the first bet. This problem disappears if we accept the stronger formulation, such that it is rationally impermissible to roll the dice on wonderful yet outlandish events. There may be other ways to block this implication.

Easwaran, Kenny (2014). Regularity and Hyperreal Credences, *The Philosophical Review* 123(1): 1-41.

Fine, Kit (1975). Critical Notice of Counterfactuals, *Mind* 84(335): 451-458.

Hájek, Alan (2003). Waging War on Pascal's Wager, *The Philosophical Review* 112(1): 27-56.

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

Isaacs, Yoaav (2016). Probabilities Cannot Be Rationally Neglected, *Mind* 125(499): 759-762.

Lewis, David (1973). *Counterfactuals*. Cambridge, Massachusetts: Harvard University Press.

——————— (1979). Counterfactual Dependence and Time’s Arrow, *Noûs* 13(4): 455-476.

——————— (1987). *On the Plurality of Worlds*. Oxford: Blackwell.

Kosonen, Petra (2021). Discounting Small Probabilities Solves the Intrapersonal Addition Paradox, *Ethics* 132(1): 204-217.

Maudlin, Tim (2007). *The Metaphysics Within Physics*. New York: Oxford University Press.

Monton, Bradley (2019). How to Avoid Maximizing Expected Utility, *Philosophers’ Imprint* 19(18): 1-24.

Muñoz, Daniel and Spencer, Jack (2021). Knowledge of Objective ‘Oughts’: Monotonicity and the New Miners Puzzle, *Philosophy and Phenomenological Research* 103(1): 77-91.

Pritchard, Duncan (2015). Risk, *Metaphilosophy* 46(3): 436-461.

Russell, Jeffrey Sanford (2023). On Two Arguments for Fanaticism, *Noûs* 00: 1-31. https://doi.org/10.1111/nous.12461

Sepielli, Andrew (2017). How Moral Uncertaintism Can Be Both True and Interesting. In M. Timmons (ed.) *Oxford Studies in Normative Ethics, Volume 7* (pp. 98-116). Oxford: Oxford University Press.

Smith, Nicholas J. J. (2014). Is Evaluative Compositionality a Requirement of Rationality?, *Mind* 123(490): 457-502.

Stalnaker, Robert C. (1984). *Inquiry*. Cambridge, Massachusetts: MIT Press.

Veltman, Frank (2005). Making Counterfactual Assumptions, *Journal of Semantics* 22(2): 159-180.

Wilkinson, Hayden (2022). In Defense of Fanaticism, *Ethics* 132(2): 445-477.