Pascal’s Mugging as a challenge to Expected Utility Theory

Abstract

Expected utility theory (EUT) is the dominant normative decision theory for dealing with uncertain outcomes. However, it entails the *fanatic conclusion* – it permits extremely small probabilities of astronomically large value to dominate our decision, leading to counter-intuitive results. It has been proposed that this problem – colloquially known as “Pascal’s Mugging” – undermines EUT. This paper rejects this view, arguing that expected utility is not rendered redundant by fanaticism. It first examines the foundations of expected utility theory, investigating whether it must necessarily lead to the fanatic conclusion. It shows that EUT cannot be discarded without entailing various issues. In response, this paper proposes the *counterfactual solution*: when conditioning on extremely small probabilities, we find that we do not have to be “mugged” by spurious decisions, because the opportunity cost is high. This cancels out the expected utility of fanatic outcomes. Thus, expected utility theory is not effectively challenged by Pascal’s Mugging. The paper concludes by illustrating the relevance of this solution to the expected utility of reducing existential risk.

Introduction

> **Pascal's Mugging**: Imagine you walk past a darkened alley when a cloaked man approaches you. He demands your wallet, containing £100. If you give it up, he promises to return tomorrow, giving you £1,000 in return. You think he is very unlikely to honour this – there is at most a 1% chance he is telling the truth. Following expected utility theory, you work out that keeping your £100 is better than betting on a 1% chance of £1,000. Seeing that you are unconvinced, he increases his promise to £1,000,000. You still think he's almost certainly lying – with ≈99.999% confidence. The man keeps increasing his promised reward for you. Eventually, the number is so large you end up agreeing to give up your wallet – he has effectively "mugged" you. Did you act rationally?

To many, this will seem an absurd question – of course not. Giving money to men in darkened alleys with an almost guarantee of no reward seems absurd. Yet, under the decision theory of expected utility maximisation, giving him your money would be completely rational. Even worse, the mugger could make the same claims every day of your life, and you would receive nothing, yet expected utility theory would endorse this.

What is going wrong here? That is the subject of this paper. In Section I, I give more context to expected utility theory, explaining how it works in ordinary decision making as well as odd situations such as Pascal's Mugging. Section II generalises the problem. It shows that with three compelling premises, we end up with the *fanatic conclusion:* for any option x, with expected utility $p_x U_x$, there is some extremely unlikely but astronomically valuable option y, such that choosing y is better than x. In section III, I explore various attempts to reject the premises that led us to the *fanatic conclusion*. I find these attempts ultimately unsuccessful. In Section IV, I argue we should embrace fanaticism – but this need not imply that we should give to the mugger. I show this through a *cancellation strategy*, which defuses the bullet-biting nature of the fanatic conclusion. Section V concludes the paper, exploring what this entails for real world high-stakes, low probability decisions.

Section I: Presenting the problem

Expected utility theory (hereafter EUT) is the dominant normative theory of rational choice (Briggs, 2019). It offers us a framework for dealing with uncertainty. Most simply stated, it is the idea that we should multiply the utility of an outcome by its probability.

To illustrate, consider the **Dice Game**: For a £5 entry fee, you can gamble on a fair die. If the die shows a 6, you win £50. Otherwise, you get nothing. Should you play?

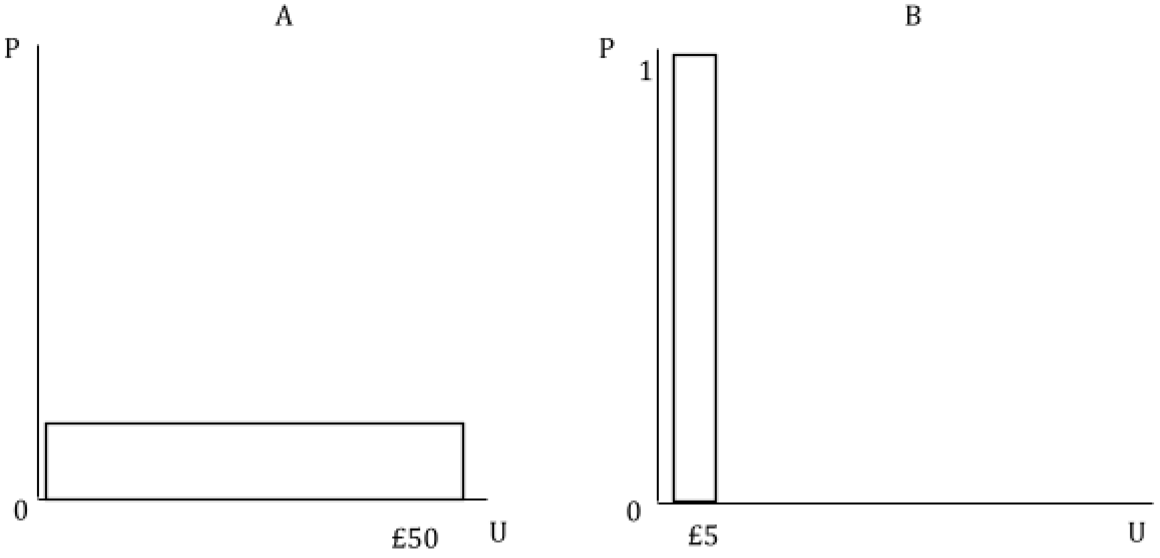
Playing (A) has a $\frac{1}{6}$ chance of winning £50, and a $\frac{5}{6}$ chance of £0. Compare this to not playing (B), where you keep £5 with certainty. Describing these outcomes in terms of expected utility, we get the following:

Play the game: $EU_A = \left(\frac{1}{6}\right)(£50) + \left(\frac{5}{6}\right)(£0) = £8.33$

Do not play the game: $EU_B = (1)(£5) = £5.$

Since $EU_A > EU_B$, we (rationally) should play the game. This is illustrated in Figure 1.

Figure 1



P represents the probability distribution [0,1], whilst U is the utility of each decision. The expected utility of each option is the *summed total area* in each diagram. As we can see, A is larger than B, because $EU_A > EU_B$.

Compare this to **Pascal's Mugging**. First conceived by Bostrom (2009), Pascal's Mugging is a play on the historical case of Pascal's Wager – the idea (originating from Blaise Pascal) that belief in God is rational because the expected utility of belief is higher than disbelief (Hájek, 2018). Taken to its extreme, Pascal's Mugging leaves expected utility maximisers open to exploitation by muggers.

The options Pascal faces are:

A: Give to the Mugger

B: Keep his money.

Suppose the mugger promises to give Pascal £10^{20} tomorrow. Pascal could believe there is a 0.00000001% chance the mugger is telling the truth – he is extremely sceptical this amount of money even exists. On the other hand, he is certain that keeping £100 has an expected utility of £100, whether or not the mugger is lying. Thus, we have the following situation:

$$EU_A = \left(10^{-8}\right)\left(£10^{20}\right) + (1 - 10^{-8})(0) = £1,000,000,000,000$$

$$EU_B = \left(10^{-8}\right)(£100) + \left(1 - 10^{-8}\right)(£100) = £100$$

Because $EU_A \gg EU_B$, Pascal strongly ought to give to the mugger under EUT.

Absurdities like this may motivate one to ask – why should we trust expected utility theory at all? Von Neumann and Morgenstern construct a proof from basic axioms of rationality (transitivity, completeness, independence and continuity) to show that expected utility maximisation is the optimum strategy when dealing with lotteries. Similar representation theorems have been given by Ramsey (1926) and Savage (1972). Additionally, we could play the **Dice game** many times, investigating which strategies give us the greatest pay off. Through the law of large numbers, we would converge on maximising expected utility.

Yet, this strategy also led to Pascal getting mugged. The next section generalises the problem, showing how it leads to the *fanatic conclusion*. Before proceeding, three important clarifications are necessary. Firstly, expected utility theory is used as a *normative*, not descriptive theory. Whilst there is strong evidence to suggest humans do not actually maximise expected utility (Tversky and Kahneman, 1974), this is not the paper's primary concern. Secondly, money is used as a placeholder for utility throughout. However, utility can be defined as anything of intrinsic value: welfare, justice, rights, equality, beauty, or more. One does not need be a utilitarian for this problem to apply. It is only assumed that value can be represented through a cardinal utility function. Thus, expected utility and expected value can be used synonymously. Thirdly, unlike in the original Pascal's Wager, the possibility of infinite utilities is ignored. This is because Von Neumann-Morgenstern's continuity axiom is violated by infinity already, and so it is irrelevant for this paper.

Section II: Generalising the problem

In this section, I generalise the problem of Pascal's Mugging into a set of three compelling premises, which lead to the *fanatic conclusion*, and the implication that Pascal is correct to pay his mugger.

II.1 Utility Function: All outcomes can be represented by an unbounded utility function.

This condition says that utility is a function with no upper bound. That is, the amount of utility in the universe can always increase. There is not cut-off put in place Formally, this can be defined as

$$\forall x \in D, U \in R, \ \neg(|U(x)| \leq |\overline{U}|)$$

Where x is a decision within the set of possible decisions D.

The rationale for imposing no bound is simple – if we compare two options, one with higher utility than the other, ceteris paribus we should always prefer the large one. With money, an upper bound makes sense. For example, diminishing marginal utility of money kicks in and may reach 0 – after possessing £1,000,000,000,000, we may simply not want an extra pound. However, diminishing marginal utility of utility does not make sense – we defined utility as that which is intrinsically valuable. As such, we will always value an extra unit of it.

II.2 Probability Function: All empirical uncertainty can be represented by a precise probability within the interval (0,1)

This premise states that the probability of any outcome must be between 0 and 1, not inclusive. Our initial probability axioms originate from Kolmogorov (2018), which state that:

1. All probabilities lie within the set [0,1], and
2. All probabilities sum to 1.

Why the move from [0,1] to (0,1)? This is because we cannot know empirical facts about the world with absolute certainty or impossibility. Thus, we cannot assign a probability of 0 or 1 to non-logical statements. For example, Pascal cannot believe with absolute certainty that the Mugger is lying, because his belief is rooted in empirical, falsifiable claims. Formally, $\forall p_i \in (0, 1)$, where p_i is the probability of some non-logical statement being true.

> II.3 Expected utility maximisation: The correct decision is that with the highest sum of multiplied utility-probability pairs.

This premise is the expected utility decision rule. This states that we should simply multiply the utility of each outcome with its likelihood of occurring. We can then sum these outcomes to get a complete expected utility. We should then pick the decision with the highest expected utility. This is the decision rule followed in the **Dice Game** and **Pascal's** Mugging. Von Neuman-Morgenstern's proof shows that expected utility is the optimum decision rule. An agent whose preference are:

(i) complete ($x \succcurlyeq y$ or $y \succcurlyeq x$ or $both$),

(ii) transitive ($x \succcurlyeq y,\ y \succcurlyeq z \Rightarrow x \succcurlyeq z$),

(iii) continuous ($x \succcurlyeq y \succcurlyeq z,\ \exists p \in [0, 1] \Rightarrow y \sim px + (1 - p)z$), and

(iv) independent ($\forall z \in R,\ p \in (0, 1], x \preccurlyeq y\ iff\ p(x) + (1 - p)z \preccurlyeq py + (1 - p)z$)

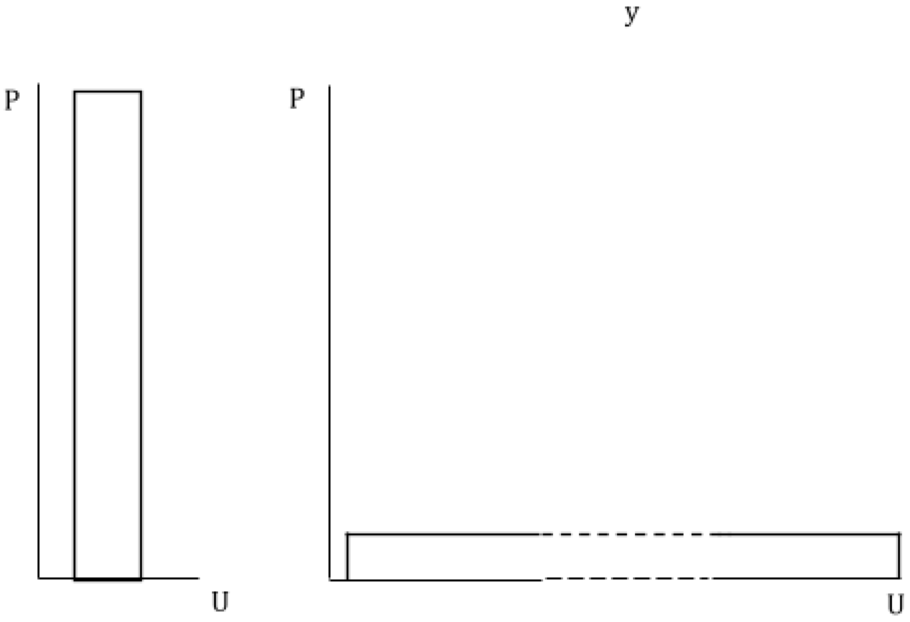
must exhibit expected utility maximising behaviour. As these are reasonable axioms of rational choice, a normative claim can be generated from this – that agents *ought* (rationally) to maximise expected utility. This crosses the is-ought divide, giving agents *reason* to follow this decision theory. For example, it gives normative reasons for Pascal to give to his mugger.

From these three premises, we arrive at the fanatic conclusion:

Fanatic conclusion: For any ordinary decision x, with expected utility $p_x U_x$, there is some alternative y with extremely low probability p_y and astronomically large utility U_y, such that choosing y is better than x.

Mathematically: $\forall x, y \in D, \exists y \; s.t \; y \succ x$, where D is the set of available decisions. We can represent this situation in Figure 2:

Figure 2



This *prime facie* innocuous conclusion is simply a natural result of expected utility theory. If utility is unbounded ($U \in R$), astronomically large amounts of utility are admitted into our expected utility equation. Thus, in our decision set there is some outcome (y), where the utility is so large that we end up picking it over x, regardless of how unlikely it is. This is analogous to the repugnant conclusion – where for any population x of size n, there is some much larger population with lower utility y, such that $y \succ x$ (Parfit 1984). The repugnant and fanatic conclusion are both counter-intuitive, despite originating from plausible utilitarian axioms. The former involves an ever-increasing unbounded population size overwhelming the utility function, whilst the latter involves an ever-increasing unbounded utility overwhelming the probability function.

To illustrate the issue with the *Fanatic conclusion* for Pascal, we must add the additional premise:

> II.4 Lemma: The expected utility of giving to the mugger is higher than keeping the money.

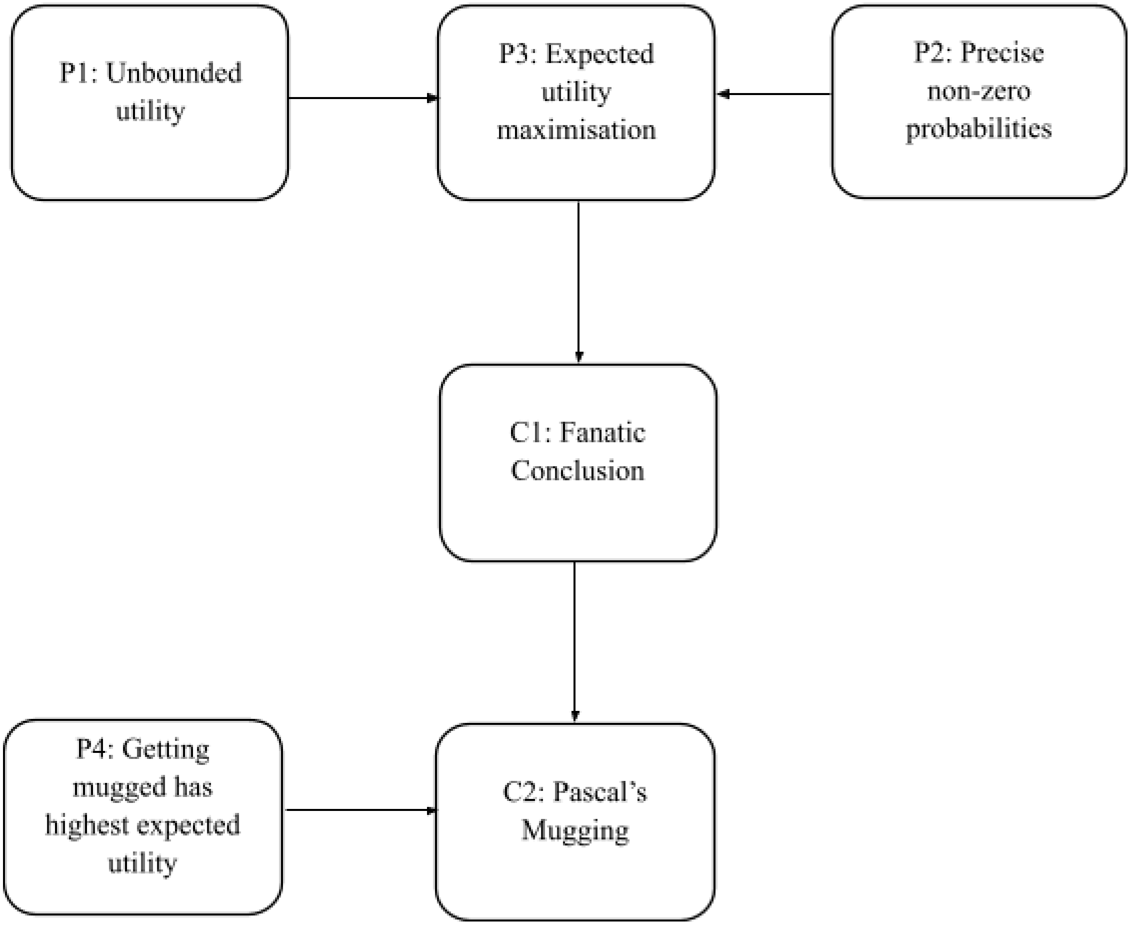
With the numbers used in Section I, this is true. This leads to the unpalatable conclusion:

> *Pascal's Mugging*: Pascal ought to give to his mugger.

Clearly, this is an undesirable outcome, because it leaves Pascal open to exploitation. More generally, it means that whenever we face uncertainty in our daily lives, we are vulnerable to similar muggers offering absurd options we cannot rationally refuse. This presents a serious challenge to expected utility theory, potentially giving us grounds to reject it entirely as a normative decision theory.

Figure 3 summarises Section II, illustrating how Pascal gets mugged.

Figure 3


P1: Unbounded utility
P3: Expected utility maximisation
P2: Precise non-zero probabilities
C1: Fanatic Conclusion
P4: Getting mugged has highest expected utility
C2: Pascal's Mugging

Section III: Potential solutions

This section considers various attempts to avoid mugging scenarios. It considers a) whether premises P1-P3 are true, and b) whether the *fanatic conclusion* that follows is valid. It rejects all, showing that the fanatic conclusion is in fact true.

III:1. Rejecting unbounded utility

One response, taken by Sprenger and Heesen (2011), is to argue that maximum possible utility does in fact have a limit. One reason why this may be the case is practical – there may simply be a limit on the total amount of possible utility in the universe. With roughly 10^{80} atoms in the Universe, there is a hard limit on what is actually possible. If the mugger suggests he will offer more than £10^{80}, there are grounds to not just consider this unlikely, but literally impossible, assigning a probability of 0. Thus, the fanatic conclusion becomes false, because for an option of £10^{80} (x), there is no possible higher utility, lower probability option y such EU(y)>EU(x). Thus, imposing a bound avoids the fanatic conclusion whilst preserving expected utility.

There are two issues with this. Firstly, it cannot be generalised. It only offers expected utility theory an escape in this particular universe, where there may be an upper limit to physical matter. But this does not provide a general sound decision theory in the abstract, which can be applied in every possible universe. For example, it is plausible that we exist in a simulation, or in a multiverse. In these cases, imposing a bound cannot work in the same manner, and so expected utility entails the fanatic conclusion. Therefore, this not a general solution. Furthermore, because empirical facts about even the size of the universe are uncertain, utility does not have a clear bound at this point. If there is a small chance that the universe is infinite, applying a probabilistic approach to knowledge implies that, in expectation, there is infinite potential utility. Thus, there is no clear rationale for imposing a utility bound non-arbitrarily.

In response, one may argue that for practical purposes, imposing a bound on potential utility may always be arbitrary, but this is more acceptable than allowing rational agents to get mugged. However, there is a more serious issue than this – that of *evaluative compositionality*. Raised by Smith (2014), this is the idea that practical rationality is not responsible for making claims about intrinsic goals, only instrumental ones. For example, if one's goal is to maximise utility, practical rationality is defined as that which best achieves this. Imposing a limit on utility tampers with our intrinsic goals. Thus, it is not rational to impose a bound of utility because it is outside the realms of practical rationality itself. Therefore, we are not rationally permitted to impose a limit on utility – leaving us unable to reject P1.

III.2: Rejecting precise non-zero probabilities

A second response concerns the probability we put on extremely unlikely outcomes, like the chance the mugger is telling the truth. *Prime facie*, it is very counter-intuitive to let our decision be dominated by an outcome which does not occur in 99.9999…% of cases. However, having non-zero precise probabilities demands that this is possible.

In response, one option is to reject the claim that empirical claims cannot receive a probability of zero. Some claims are so absurd as to validate giving them absolute zero weight. Whilst tempting, this leads to severe problems in certain scenarios. For example, imagine Pascal assigns a probability zero to the mugger's promises of giving him £1,000,000. But, to try and convince Pascal, the mugger could then show Pascal a suitcase with £1,000,000 inside it, and let him inspect all the notes, and get them all independently verified as legal tender. He could leave the suitcase overnight, protected and sealed, and do everything else he can to convince Pascal. With strong evidence like this, it now makes sense for Pascal to start believing the man. Yet, after imposing a credence of zero on the claim "the mugger will give me £1,000,000", it is impossible for him to change his degree of belief up from this. This seems absurd – when shown such strong evidence, it makes sense for Pascal to change his mind. A credence of zero forbids this.

A second option is to accept that all claims do have a non-zero chance of being true, but claim that we can rationally ignore claims below some threshold $p = \varepsilon$. Smith (2014) takes this approach, arguing that rationality does not require us to factor in arbitrarily low probabilities. This would allow us to dismiss the mugger, because we put such low weight into his claims that we end up ignoring them.

However, this response suffers from *partition dependency.* That is, the outcomes we designate as above or below a particular threshold depend on how we classify the probability of certain states. For example, a dice roll has a probability 1/6 of landing any way up. However, the probability that it lands in *any particular place*, to the subatomic level, is extremely low. This number is lower than ε, so we are permitted to ignore it. Yet every single possible dice roll has this feature – every roll has less than ε chance of landing in a particular way. We would thus dismiss every single possibility, including highly likely ones. This problem occurred because the outcomes we considered depended on the way in which we partition it. This problem afflicts all decisions, including Pascal’s Mugging. Thus, it is difficult to ignore some probabilities without ignoring all of them – which leads to absurdity.

A third option deals with precise probabilities. It may be argued that assigning a precise number to the chance of every outcome is simply impossible. For example, it is very challenging to non-arbitrarily make precise statements about the chance the mugger is lying. It could be 10^{-5} or 10^{-15}, and it is hard for us to give good reasons to distinguish between these two, despite the enormous impact it has on our expected utility calculation. Thus, we are in a position of Knightian ignorance[1], with uncertainty about uncertainty

In response, this is a valid concern, but only for descriptive decision making. Humans face bounded rationality, meaning we cannot precisely compute all factors. However, this is irrelevant to normative decision theory – how one *should* make decisions. If we possessed the computing power required, we could precisely define the chance the mugger is telling the truth through a Bayesian approach. This may be practically challenging, but this is not a weakness of precise probabilities as an ideal. Thus, all uncertainty can be described through precise non-zero probabilities.

III.3: Rejecting Expected Utility maximisation

In response to Premise 3, some argue that orthodox expected utility theory is the incorrect approach to combining probabilities and utilities. For example, Buchak (2013) argues in favour of a risk-weighted decision theory, where one is rationally permitted to choose an option with a lower expected utility over a higher one. For example, compare £10 with certainty (A) to a 10% chance of £100 (B). Whilst expected utility would say we should be indifferent, Buchak’s risk weighted decision theory permits choosing A over B. Applied to Pascal’s Mugging, this would allow Pascal to walk away with his money, since this is the more risk averse option compared to getting mugged, even if the latter’s expected utility is higher. On this account, we can refute the fanatic conclusion, because for some

[1] See Knight (1921) for more on Knightian uncertainty

option x, a higher utility but lower probability option y is riskier, and hence should not be chosen, even if its orthodox expected utility is higher.

However, this view suffers from a violation of the *sure thing principle*. This states that if a decision maker would do A if X is true, and would also do A if ¬X is true, then they should do A. This weak condition is violated by risk-weighted expected utility, as illustrated by Briggs (2015). This is intuitively irrational, and something which conventional expected utility maximisation does not suffer from. Therefore, rejecting Premise 3 causes more irrationalities than it solves.

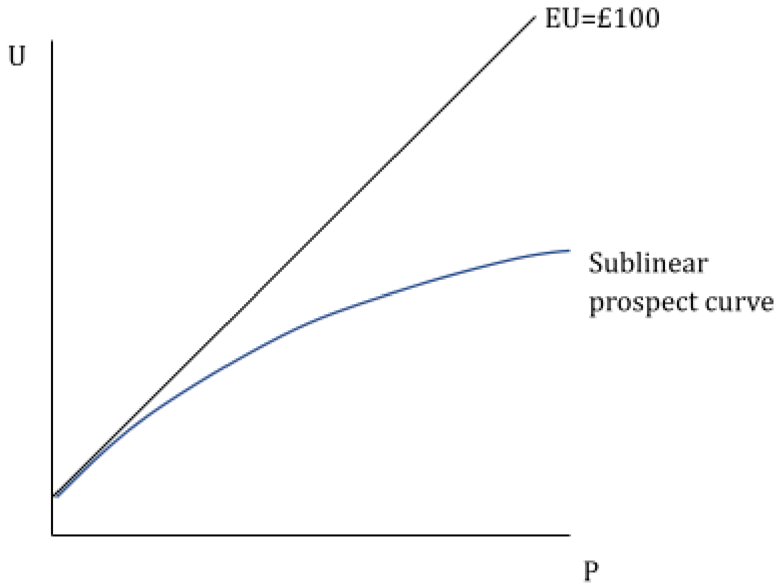
III.4: Rejecting the validity of the conclusion

Another method for rejecting the fanatic conclusion is to show that the argument is invalid. This would mean accepting (II.1) unbounded utility, (II.2) precise non-zero probabilities and (II.3) orthodox expected utility theory, but not accepting the *fanatic conclusion* – that there is some absurdly unlikely but hugely valuable option y which is better than an ordinary choice x. One could do this through arguing that as the purported utility of an outcome increases, our confidence level that it will occur will decrease at a faster rate. Thus, $EU(x) > EU(y)$.

To demonstrate, consider once again the original Pascal's Mugging case. Suppose initially, the mugger promises to give Pascal $U_x = £10^3$. Pascal thinks that there is a $p_x = 10^{-2}$ chance the mugger will deliver on this. In Bayesian lexicon, this is Pascal's "prior" – the probability the mugger telling the truth before receiving any evidence. Thus, $EU(x) = £10$. This is worse than keeping £100 with certainty.

In response, the mugger changes his claim: he can give Pascal $U_y = £10^6$. Without updating his prior of the mugger's claims, Pascal would now give his money away, since the expected utility is $£10^4$. However, it seems reasonable for Pascal to be more suspicious. Muggers giving away £1,000 in the street already was already unlikely – but giving away £1,000,000 seems ludicrous. It therefore makes sense for Pascal to update his prior probability downwards, in response to the increase in utility. The crucial question is – by how much?

Those who reject the fanatic conclusion argue that one must decrease their probability *faster* than the utility function increases – a sublinear relationship. This is plotted below:



If this argument holds, then as the utility increases, the expected utility decreases. Because Pascal's prospect curve is below the indifference curve $EU = £100$, Pascal never has to give to the mugger, even as utility tends to infinity. Thus, one can accept unbounded utility functions, precise non-zero probabilities and expected utility maximising, without agreeing that the Fanatic conclusion must follow.

Whilst this does offer an escape from the Fanatic conclusion, it entails issues of its' own. Most worryingly, it implies bizarre results when the numbers get *really* large. Suppose the mugger promises to give Pascal $U = £10^{9999}$. For example, perhaps the mugger has access to another universe where he can extract all it's value, and give it all to Pascal, so long as he returns to the same spot tomorrow. Under a sublinear prospect curve, Pascal reasons this is less likely than $p = 10^{-9999}$. Thus, he thinks he has escaped the mugger. With a credence of $p < 10^{-9999}$, there is no amount of evidence that could change Pascal’s mind. This causes a problem if the mugger then starts revealing compelling evidence – for example, he could show Pascal around this alternative universe, explain in detail the physics, and how tomorrow all this potential utility will be his if Pascal hands over £100 today. If we were to undergo this experience, it would be clear *ex post* that you should hand over your wallet. However, within a Bayesian epistemology, a prior of $p < 10^{-9999}$ cannot be updated on fast enough to accept the mugging, even when rationally one ought to[2]. As a result, having a probability function always smaller than the utility function is large does not solve the problem of extremely large numbers – in fact, it adds issues.

Thus, precise non-zero probabilities and unbounded utilities, combined using expected utility theory, must lead to the *fanatic conclusion*. It is tempting to abandon expected utility theory entirely. However, in the next section, I argue we can have our cake and eat it – we can preserve orthodox expected utility without getting mugged.

Section IV: The counterfactual solution

In this section, I argue that even if expected utility theory entails the fanatic conclusion, this need not be a *reductio ad absurdum*. Specifically, it does not imply that Pascal should give to his mugger. More generally, I show that when faced with similar “fanatic” options, we can defuse their potency through the counterfactual solution. This allows us to avoid fanatic decisions often, but not always. I believe this is the optimum solution, as we preserve expected utility theory whilst not getting mugged.

To show why Pascal should not give to his mugger, reconsider premise IV: The expected utility for Pascal of giving to the mugger is higher than keeping his money. Also recall the *sure thing principle*:

> *Sure thing principle*: If an individual should do A|X being true, and should do A|¬X being true, then they should do A.

[2] For more on why the probability function cannot update fast enough, see Yudkowsky (2013) or Kokotajlo (2018)

We can break down the problem to see whether giving to the mugger fulfils the *sure thing* principle. In this scenario, X is the condition of the mugger being honest. When the mugger is lying, it is obvious he should just keep his £100, rather than give it away with no utility at all. Suppose Pascal thinks there is a 10^{-5} chance the mugger is honest. When the mugger is telling the truth, the potential utility is astronomically large – perhaps £10^{10}. Meanwhile, the utility of just keeping the money, conditional on the mugger telling the truth, is assumed to be £100. It initially appears that in expected utility terms we should give to the mugger. Table 1 shows these outcomes:

Table 1	P(Honest)=0.00001	P(Lying)=0.99999	Expected Utility
U(Keep)	£100	£100	£100
U(Give)	£10000000000	£0	£100000

However, there is a crucial mistake leading to us giving to the mugger: the claim that U(Keep)|p(Truth) =£100. Why is this misleading? Suppose that the mugger's claims are true – that we live in a universe in where muggers carry around suitcases with £10^{10} in them. We must then ask – does it follow that we should give our wallet to this particular mugger, rather than do something else? For example, perhaps a second man may approach us, and can offer us £10^{10} +1 utility. If so, we should give to the 2nd over the 1st mugger. It would be a mistake to give to the first because there is a high *opportunity cost*. More generally, conditional on the mugger telling the truth, there may be plenty of other opportunities for utility. These opportunities may supersede the first mugger's claims. Thus, we do not have good reasons to give to *this particular mugger*. The value of keeping our £100 wallet has increased dramatically. This means that, conditional on the mugger's claims being true, we should not necessarily give to *that same mugger*. Returning to the *Sure thing principle*:

> If we should (keep our money| The mugger being honest), and should (keep our money| the mugger lying), then we should keep our money

Following the *sure thing principle*, we should keep our money. Why? – because if the mugger is telling the truth, we can get more utility from keeping our wallet than giving it to him. Thus, we have managed to escape the pull of Pascal’s Mugging, without abandoning expected utility theory. Instead, we have recognised that the expected utility of getting mugged away is lower than keeping our money, even when the mugger makes claims about astronomical value.

One may object that this is not a complete solution, because it still allows for the possibility of us getting mugged. For example, U(get mugged)|Honest may be greater than U(keep)|Honest. In this scenario, we are unable to invoke the *sure thing principle*, because our answer to what we should do switches depending on the evidence. There will be occasions where we end up giving to the mugger, even on some extremely small probability that they are telling the truth.

In response, I would contend that this is an acceptable outcome. If we consider the world conditional on the mugger’s claims being true, there will be instances where it is best to give to the mugger, rather than do anything else. However, this is not unlikely, due to the higher value of keeping the wallet. Once one has considered all the pros and cons of giving to the mugger conditional on them being honest, we may settle on the conclusion that in such a hypothetical world, we would rationally give to the mugger. If this is the case, it is not unacceptably fanatical to give to the mugger, because there are good reasons to do so.

In summary, a fuller picture of the expected utility calculation reveals that the value of keeping one’s money is vastly greater in the world where the mugger is honest. It may often be more valuable than giving to the mugger. In these cases, we should refuse to pay up. However, in some scenarios, it is rational to give to the mugger when they truly offer the highest utility.

Section V

In this section, I consider what the practical applications of this result is, and conclude by summarising how Pascal’s Mugging fails to challenge expected utility theory.

In the real world, there are some opportunities we face to create extraordinary value, even if it is very unlikely. For example, in the next billion years over ten quadrillion humans could inhabit the Earth, each with intrinsic value. However, extinction this century could put this enormous utility at risk. Thus, even the smallest chance of preventing this occurrence has tremendous expected utility. For example, Bostrom (2013) argues that on conservative estimates, the expected value of reducing existential risk by one millionth of one percent is the equivalent of saving 100 million people. In response to arguments such as these, some dismiss them as being a "Pascal's Mugging" (Matthews, 2015), a *reductio ad absurdum* to ignore them. My paper makes the case that dismissing such opportunities is wrong. There is no good reason to reject fanatical options from our decision theory. Moreover, I show that sometimes, when the evidence is strong enough, there are good reasons to allow ourselves to "get mugged". Efforts to reduce the probability of existential risk are one such example – compared to any counterfactual alternatives, reducing existential risk is the best way to improve humanity's trajectory. Hence, my paper gives evidence for why one should not dismiss the slim possibility of improving the far future.

In conclusion, this paper has explored Expected Utility theory, and the issues that can arise when taking it to the extreme. It has considered the challenge of the *fanatic conclusion*: that for any option x, with expected utility $p_x U_x$, there is some extremely unlikely but astronomically valuable option y, such that choosing y is better than x. It has explored mechanisms for avoiding this conclusion, such as modifying the utility function, probability function, decision rule and normative requirements. However, it found none of them convincing. Instead, this paper opted to embrace the fanatic conclusion, but demonstrated that this need not imply we should give to Pascal's mugger. It showed that when the mugger is telling the truth, it does not necessarily follow that giving them our money is the best option. However, it has not unilaterally ruled out the possibility of paying the mugger, such as when they present particularly compelling evidence that paying them will best increase utility. Finally, this paper has considered what this entails for practical decisions such as existential risk reduction. It has found that we can have our cake and eat it – we can retain expected utility theory without giving in to Pascal’s Mugging. Therefore, Pascal’s Mugging is not a successful challenge to expected utility theory.

Words: 5259

Bibliography

Bostrom, N., 2009. Pascal's mugging. *Analysis*, *69*(3), pp.443-445.

Bostrom, N., 2013. Existential risk prevention as global priority. *Global Policy*, *4*(1), pp.15-31.

Briggs, R., 2015. Costs of abandoning the Sure-Thing Principle. *Canadian Journal of Philosophy*. Cambridge University Press, 45(5-6), pp. 827–840.

Briggs, R. A., 2019. *Normative Theories of Rational Choice: Expected Utility*, *The Stanford Encyclopedia of Philosophy.* Available at: https://plato.stanford.edu/entries/rationality-normative-utility/

Buchak, L., 2013. *Risk and rationality*. Oxford University Press.

Hájek, A., 2018. *Pascal’s Wager*. *The Stanford Encyclopedia of Philosophy*. Available at: https://plato.stanford.edu/entries/pascal-wager/

Knight, F.H., 1921. *Risk, uncertainty and profit* (Vol. 31). Houghton Mifflin.

Kokotajlo, D., 2018. *Tiny Probabilities of Vast Utilities*. [online] Effective Altruism Forum. Available at: https://forum.effectivealtruism.org/posts/nMfhXPiqPRDtEdauY/tiny-probabilities-of-vast-utilities-bibliography-and

Kolmogorov, A.N. and Bharucha-Reid, A.T., 2018. *Foundations of the theory of probability: Second English Edition*. Courier Dover Publications.

Matthews, D., 2015. *I spent a weekend at Google talking with nerds about charity. I came away … worried.*. [online] Vox. Available at: https://www.vox.com/2015/8/10/9124145/effective-altruism-global-ai

Parfit, D., 1984. *Reasons and persons*. OUP Oxford.

Ramsey, F.P., 2016. Truth and probability. In *Readings in formal epistemology* (pp. 21-45). Springer, Cham.

Savage, L.J., 1972. *The foundations of statistics*. Courier Corporation.

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

Sprenger, J. and Heesen, R. 2011. The bounded strength of weak expectations. *Mind*, vol. 120, pp. 819-832.

Tversky, A. and Kahneman, D., 1974. Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), pp.1124-1131.

Yudkowsky, E., 2013. *Pascal's Muggle: Infinitesimal Priors and Strong Evidence - LessWrong*. [online] Lesswrong.com. Available at: https://www.lesswrong.com/posts/Ap4KfkHyxjYPDiqh2/pascal-s-muggle-infinitesimal-priors-and-strong-evidence