An offer he can't refuse: exploiting ideal decision theory

Adam Chalmers

2016

1 Abstract

Our best models of ideal decision-making have a deep flaw: they consistently over-value certain lotteries and gambles. I outline a class of lotteries (HULP) whose value is not accurately measured by expected utility. I show that expected utility maximisers can have their agency hijacked and their preferences reordered if offered a HULP gamble. Decision algorithms which are not vulnerable to HULP exploitation are discussed, including Nicholas Smith's Rationally Negligible Probabilities.

2 How to lose all your money by following decision theory

2.1 What is decision theory and why does it matter?

Decision theory is the study of how to make the right choice in a particular situation. Economists, politicians, scientists, financial planners and doctors all use decision theory to choose which possible action will best let them achieve their goals.

Making a plan is simple when one knows all the relevant information. Imagine a doctor choosing between two medical treatments. If we knew which of them would help her patient more, how much each would cost, and exactly what the side effects would be, her decision would be easy. There's no need to resort to decision theory. However, if the doctor is unsure exactly what disease her patient has, or if each treatment has a wide range of possible costs and side-effects, then her decision becomes much more complicated. In situations like this, decision theory offers precise mathematical analysis of each possible outcome and its likelihood. It allows people to replace their intuitions —which are often flawed and deceptive —with mathematical guides that quantify risk and uncertainty. Decision theory has proved incredibly effective at helping people in uncertain situations make important choices.

Decision theory is part philosophy, part economics and part mathematics. It has proved incredibly effective at solving real world problems, as evidenced by its popularity amongst mathematicians, computer scientists, economists and scientists. However, philosophers have devised some unusual decision problems which decision theory seems to provide misguided solutions to.

For example, the St. Petersburg lottery is a thought-experimental game with a very small chance of an incredibly high payoff[1]. Decision theory says this gamble actually has infinite value, because there's no limit to the amount of money you could win (although your chances of winning a particular amount get exponentially smaller as the amount gets exponentially higher). However, many philosophers disagree, believing that decision theory *overvalues* this gamble. It would be foolish, they argue, to value a game at infinite dollars when there's only a 1% chance of winning more than $128 from it.

TODO: cite papers for the above arguemnt

If TODO: CITATION is correct, then decision theory is flawed and requires revision. This is alarming, because many view decision theory as one of philosophy's big success stories due to its success at solving real-life problems since its formulation in the 20th century. However, if it fails at these 'paradox' problems then the fundamental axioms of decision theory might require revision. Perhaps, like Newtonian astrophysics, standard decision theory effective at real-world problems on Earth, but will fail to model more exotic problems we could face in future decades. If so, we may be able to expand decision theory with new rules for these new problems. Or, more alarmingly, we may have to throw it away like Newtonian astrophysics, and begin work on an entirely new framework. Much like the discovery of relativity, this would be a long and difficult task requiring a paradigm shift for multiple academic fields.

Fortunately, I believe decision theory requires only small modifications, and not an entire General Relativity-like revolution. I aim to show that decision theory can be modified to better advise agents facing these paradoxes without losing its effectiveness at solving real-world problems. In this paper, I will examine how some paradoxes can be used to exploit people who strictly follow decision theory. They can, for example, be forced to give free money to exploiters who offer them clearly fraudulent bets and gambles. Even if the decision-theoretic agent is overwhelmingly sure that they're being exploited, decision theory still tells them to hand their money over. Exploiting an agent like this is easy: you simply have to offer them a small chance at an infinite amount of money if they give you a few dollars. No matter how little they believe you, their distrust will never be high enough to cancel out the desirability of an infinite reward. The cost—benefit analysis will always advise them to try for infinite money, regardless of how unlikely it is.

This is a grave problem for decision theory because exploitation is simple to perform and effectively lets the exploiter control the victim's actions. This section will outline decision theory and its problems. Section 2 will examine the problem in more technical detail. However, in sections 3 and 4 I will review

[1]The details of this game are outlined in section TODO: INSERT SECTION HERE

Dr. Nicholas Smith's modified decision theory [11] and show that agents who subscribe to it can escape this sort of exploitation. Section 5 will address possible objections to my analysis and look at my work's broader implications for decision theory.

2.2 Expected utility

Normative decision theory is the study of the mathematical processes of making the ideal decision in a range of different scenarios. Decision theory is applied to decision problems, which comprise of:

- An agent, the decision maker
- A set of actions the decision maker can take
- A set of states: mutually-exclusive propositions which describe ways the world could be
- A set of outcomes that may result from a given action being taken while a given state obtains
- A mapping from outcomes to utilities (numerical measures of desirability)

In decisions under risk, the agent has a probabilistic model which tells the agent the probability that a particular outcome will occur, given the agent takes a certain action while a certain state obtains.

If an agent knows the world's possible states, understands the actions available to them, the probabilities with which each action produces each outcome, and has assigned utilities to each outcome, then the agent is able to apply decision theory to their current situation. A decision algorithm takes these facts as inputs and ranks each action in order of how useful they are to achieving the agent's goals.

Expected utility maximisation is a specific decision theory algorithm which states that agents should always choose the action with the highest expected utility. An action's expected utility is the average of each possible outcome's utility, weighted by how likely that outcome is. Mathematically, it is defined as

$$EU(a) = \sum_{\substack{s \in S \\ o \in O(a)}} P(o|a \wedge s) \times U(o)$$

where a is an action and o is any outcome which may arise as a result of that action. Expected utility maximisation is often suggested to be a *norm* of decision theory: a correct and rational mode of decision making that agents should strive to emulate. This is based on three arguments.

Firstly, agents who maximise their expected utility will obtain maximal utility in the long run. This is demonstrated in [12] which shows that maximising expected utility maximises utility. If an agent assigns utilities to outcomes in

a way that accurately reflects their desires, expected utility maximisation will effectively guide agents towards achieving their desires.

Secondly, expected utility theory is a logical implication of basic axioms of rational choice [12]. This means any agent who doesn't take an action of maximal expected utility is violating one of the axioms of rational choice, which are all intuitively reasonable.

Thirdly, expected utility has been very successful in the real world. It seems to correctly value probabilistic actions such as gambles, medical interventions and business decisions. It has been widely adopted in a range of fields and businesses, which provides a degree of practical justification for it.

These three reasons mean expected utility is a well-regarded norm of decision theory. However, as previously discussed, philosophers have devised many decision problems where expected utility seems to dramatically overvalue particular actions and endorse irrational choices. I believe this evaluation is a serious problem for expected utility maximisers, and that therefore despite its theoretical soundness and real-world practicality, it requires modification. This is because expected utility maximisers can be exploited by offering them an overvalued gamble. Now that our readers have a basic understanding of decision theory, I will discuss these overvalued gambles and paradoxes in more depth, so that we can refer to them in our later discussion of exploitation.

2.3 Overvalued paradoxes

Expected utility theory systematically overvalues a class of decision theory problems I call High Utility, Low Probability (HULP) problems. I will briefly explain three exemplars of HULP problems and then discuss the features they share which are constitutive of HULP problems.

2.3.1 St. Petersburg Paradox

In the St. Petersburg paradox (first described in [1]) an agent is offered a gamble where a fair coin is repeatedly flipped until it lands tails-up. The agent is then paid $\$2^N$, where N is the number of heads that were flipped.

The expected utility of this gamble is $\$(\frac{1}{2} \times 2 + \frac{1}{4} \times 4 + ...) = \infty$ [10], and therefore an expected utility maximiser should be willing to pay any finite amount to purchase it. However, this seems like a gross overvaluation because the incredibly high-value outcomes of this lottery have incredibly low probability of occurring. 97% of the time, the agent will make $32 or less from the gamble. As Ian Hacking wrote, 'few of us would pay even $25 to enter such a game' [5].

2.3.2 Pascal's Wager

Blaise Pascal proposed treating theism as a decision problem: an agent's decision to participate in religious observance is motivated by their desire to enter heaven, conditional on God existing [9]. If the cost of performing religious duty is C, then Pascal's Wager has the following decision table:

	God exists	God doesn't exist
Worship	$\infty - C$	$-C$
Don't worship	0	0

By this, worshipping God dominates non-worship. Heaven is presumed to have infinite utility, therefore the expected utility of worship will always be infinite (because an agent's disbelief in God could only ever be finite). This remains true no matter how tiny the agent's estimate of God's existence is. It also remains even if the religious observance is costly or demanding—even if the agent has to pay a large (but finite) amount of utility to perform religious observance, worshipping still maximises expected utility.

2.3.3 Pascal's Mugging

Many objections to Pascal's Wager dispute specific properties of God ([6]) or the possibility of infinite utility ([7])). Pascal's Mugging was proposed by [2] in order to focus criticism towards the decision-theoretic aspects of Pascal's Wager and away from metaphysical or mathematical analysis.

In Pascal's Mugging, the agent is confronted by a Mugger who claims to be a wizard whose powers can magically multiplying money. She asks the agent for a loan of $10, promising to use her magical powers to give the agent a fantastic sum of money in return. The agent has no reason to believe in magic and the Mugger offers no evidence of her wizardry, and so it appears rational for the agent to reject her offer.

However, the Mugger can promise an arbitrarily large amount of money, and our agent's skepticism, while strong, is fixed and non-zero in accordance with norms of rationality. So if the Mugger offers [equation here] such that [equation here], then the agent's expected utility is maximised by giving $10 to the Mugger, despite having no reason to believe her claims.

2.4 HULP Problems

These three problems all share some similar features. In all of them, a gamble's expected utility and its actual intuitively reasonable value appear to differ. All involve a choice between two options:

- A walk away option with certain chance of zero payoff (and therefore expected utility of zero). Examples of this include not buying the St. Petersburg lottery, not worshipping God, and not paying the Mugger.
- A HULP option (High Utility payoff with Low Probability) with high expected utility. This includes buying the St. Petersburg lottery, worshipping God and paying the Mugger.

Expected utility maximisation instructs agents to choose the HULP action over the walk-away action because the HULP action has higher expected utility. However, agents who consistently choose the HULP action in HULP problems leave themselves open to alarming consequences. Here are some of them.

The St. Petersburg gamble can easily be modified to award payouts in utiles instead of dollars[2]. If an expected utility maximiser is offered the chance to play a Utility St. Petersburg gamble, they should pay any finite utility cost to do so. Such an agent should be willing to bear any utility cost—trading all their wealth, or murdering large numbers of innocent people—in order to play the gamble. As long as the price is finite, the agent must pay it or violate the expected utility maxim.

Pascal's Wager can similarly be used to compel an expected utility maximiser's options. An agent should easily give up their riches and material goods if such costs are necessary for the agent's actions to qualify as religious observance. Indeed, religious observance can involve any finitely-large utility cost and still dominate non-observance, due to the presumption that heaven is valued at infinite utility. Suppose an agent was considering the worship of Quetzalcoatl the Aztec serpent god. Quetzalcoatl's worship involves human sacrifice, which our agent thinks is abhorrent and values at, say, -9000 utiles. An expected utility maximiser should still prefer to perform human sacrifice and be rewarded with heaven rather than not worship, because a large finite utility cost still does not lessen the infinite utility of heaven.

Of course many responses to Pascal's Wager point out that the agent doesn't face a binary choice. There are many possible gods—Quetzalcoatl, Jesus, Poseidon—and not all of them demand human sacrifices. Many gods offer a chance at infinite utility. Alternative gods who do not require human sacrifice exist, and the decision maker is free to choose a god whose worship is more pleasing, since differences in each god's likelihood and worship-cost are cancelled out by the infinite utility reward in the expected utility calculations [4].

Pascal's Mugging cannot be resolved by the many gods objection. If the wizard offers to grant arbitrarily high utility instead of money, then an expected utility maximiser should be willing to pay any amount of utility to appease the mugger. By similar reasoning to the previous examples, an expected utility maximiser would sacrifice their family or perform any other arbitrarily undesirable deed, because the Mugger can offer them utility high enough to perfectly balance out the expected utility equation.

None of these individual considerations is new. That an infinite utility gain can outweigh a finite utility cost is obvious. One could simply bite the bullet and claim expected utility maximisation is a correct norm of decision theory. However, the fact that this same problem keeps rearing its head in different decision theory paradoxes hints that there is a deeper, more systematic problem. These are not three seperate edge cases each designed to be overvalued by expected utility theory. Rather, they provide insight into a deeper problem with expected utility. To illustrate this, I will show how expected utility maximisers can be exploited by agents who can present them with HULP problems.

[2]Imagine a sociopathic scientist is offering St. Petersburg gambles that pays out not dollars, but saved human lives (perhaps he knows the secret to manufacturing a much-needed vaccination, or perhaps he has some doomsday device that can kill an arbitrarily large number of humans).

2.5 Systematically exploiting expected utility maximisers

Suppose there are two agents, an expected utility maximiser called Max, and an exploitative agent called Eliza. If Eliza knows Max is an expected utility maximiser, she can force him to undertake arbitrarily (but finitely) unpleasant actions by appealing to the norms of his decision theory.

Here is a specific example. Eliza would like $100 from Max, and knowing that he is an expected utility maximser, offers to sell him a St. Petersburg lottery for $100. Max knows buying the gamble would maximise his utility in this situation, but he doubts she has the financial backing to guarantee he'd receive 2^n dollars after flipping n heads. Eliza retorts that, even if Max's belief in her is in 10^{-20}, the expected utility of paying her is still infinite regardless of her honesty.

$$EU(\text{Pay Eliza}) = (10^{-20} \times (\infty - C)) = \infty$$

$$EU(\text{Don't pay}) = 0$$

As an expected utility maximiser Max is forced to agree that giving Eliza $100 is the highest-utility option available to him, and hands it over. Eliza, of course, reveals she was lying and walks away with Max's $100.

Why was Eliza successful? If Max was not an expected utility maximser, or didn't value the St. Petersburg gamble at infinite expected utility, he would have grounds to deny Eliza. Unfortunately, he is neither, and Eliza can thus compel any action from him by offering him the chance at a St. Petersburg lottery in return.

Note that Eliza could have performed a Pascal's Mugging just as easily by promising to use her magical powers to grant Max a large reward, the size of which would be calculated to dominate Max's skepticism. In fact, constructing any HULP problem will allow Eliza to exploit Max and force his action. This is because in all HULP problems, the HULP action has higher expected utility than the walk-away action regardless of which particular large finite cost is attached to performing the HULP action. Eliza could ask Max to do anything and rest assured that, despite the incredibly large utility costs Max would pay in carrying out these actions, Max would be compelled to obey her if he is a genuine expected utility maximiser.

These concerns demonstrate why expected utility maximisation is inadequate as a norm of decision theory. Any expected utility maximising agent can have their agency hijacked by an exploiter who is able to offer them HULP problems, and as we have seen, such problems are trivial and cheap to offer. In the next chapter I will formalise my notion of HULP problems and go over the mathematics behind HULP exploitation.

3 Exploitation

In the previous section, I provided an intuitive explanation of HULP exploitation. To summarise:

- The victimised agent has some available action A which is undesirable because of its low (perhaps negative) expected utility
- The exploiter agent offers the victim an infinite or arbitrarily large reward just in case the victim performs A.
- The victim now evaluates that performing A has infinite (or arbitrarily large) utility, and performs A.

This section formalises the notion of HULP exploitation and explains why immunity to such exploitation is a desirable norm of rationality. I will begin by outlining the generic form of HULP exploitation, and showing how the Max-Eliza exploitation in the previous section is a specific instance of this general exploitation method. I will then show that HULP-exploitable agents can have their agency hijacked and their preferences reordered at no cost to the exploiter. Agents who wish to preserve their utility function should thus be careful to avoid HULP-exploitation, and we should enshrine immunity to HULP-exploitation as a norm of rationality and desirable property of any decision agents.

3.1 How does exploitation work

Agents should avoid specific situations of HULP-exploitation, and if possible, avoid being HULP-exploitable at all. This is because HULP-exploitation can lead to the victim modifying and reordering their preferences in a way which leads to them avoiding opportunities to greatly advance their interests, or even taking actions which greatly set back their interests. This is informally demonstrated in the Max-Eliza story above, where Max initially prefers keeping his money to giving it away, but after Eliza offers him the gamble, he prefers giving his money away to keeping it.

Formally: HULP-exploitation allows the exploiter to arbitrarily reorder the victim's preferences by presenting offers which modify the expected utility of the victim's actions. The following section formalises how exploiter can compel or prevent any action from the victim at will, at no personal cost.

3.1.1 Formalising exploitation

Suppose the victim has some action A available to them, which is certain to cause the undesirable outcome $O, u(O) < 0$. The victim will prefer not to act upon this action.

The exploiter then offers the victim a desirable reward $R, u(R) > 0$ if the victim performs A. The victim estimates the probability that the exploiter's offer is honest, $p(H)$, and calculates the new expected utility of performing A:

$$EU(A) = p(H).((u(R) + u(O))) + (1 - p(H)).u(O)$$

Rearranging this yields

$$EU(A) = p(H).u(R) + u(O)$$

A is therefore desirable (i.e. $EU(A) > 0$) exactly when

$$p(H).u(R) + u(O) > 0$$

i.e. when

$$u(R) > \frac{-u(O)}{p(H)}$$

Therefore, if the exploiter offers a sufficiently-high reward, they can turn the initially unattractive action A into a desirable one. Crucially, this requires no cost or commitment from the exploiter. The exploiter merely has to put the offer to the victim in order to flip their preference regarding A.

More generally, suppose the victim has two possible actions A and B which result in outcomes X and Y, and prefers Y to X such that $u(X) < u(Y)$. If an exploiter offers a reward R for performing A, then provided

$$u(R) > \frac{u(B) - u(A)}{p(H)}$$

the agent will prefer A to B.

3.2 Should agents avoid having their desires changed?

We have seen that HULP-exploitable agents can have their preferences reordered by malicious agents. At first glance, it is unclear how serious of a problem this poses. Decision theory tells agents how best to achieve their goals, but it has nothing to say about the stability or value of those goals themselves.

Two extreme views on preference-modification are tenable. In computer science and artificial intelligence, goal-driven agents are often thought to value stability of preference. [8, 3] argue that powerful agents will generally go to great depths to preserve their utility function U (i.e. their preferences), because allowing it to be modified to some other U' would be unlikely to achieve progress towards the original U. Informally: "if you offered Gandhi a pill that made him want to kill people, he would refuse to take it, because he knows that then he would kill people, and the current Gandhi doesnt want to kill people" [13].

I feel the computer-scientific model of utility functions fails to capture the possibility of desirable preference change that many decisionmakers naturally go through. Even the most rational humans often find preference change to be desirable and meaningful. For example, few people enjoy the taste of alcohol when they first try it, but many consider developing such a taste to be desirable and good. Many preference changes function like this, regardless of whether or not the change is consciously chosen or self-directed. Consider a person Alex who has just started to enjoy jazz music, but can't stand Duke Ellington[3]. Alex would probably enjoy having their preference against Duke Ellington altered, because enjoying the music of Duke Ellington would provide them with more

[3]I originally heard this argument for preference change from Dr. Mark Colyvan [NOTE: Mark should I cite you formally on this?]

chances to enjoy music and a deeper appreciation of jazz. Whether Alex grows to enjoy Duke Ellington naturally or by someone else's doing (suppose Alex's friend forced them to listen to Duke Ellington records until it developed a familiar comfort), the end result is a pleasant, desirable preference shift.

However, this example does not show that preferences are arbitrary, or that preference changes are unconcerning. While many people wish to develop a taste for alcohol, or jazz music, few people in their right mind wish to develop a taste for feces, or the sound of nails on a chalkboard, or for snuff videos. It seems there are some preference changes we do not wish to undergo. I feel this is sufficient grounds to claim that HULP-exploitation is a serious problem, because it allows the exploiter to reorder *any* preference regardless of how strongly attached the agent is to it. An exploiter could indeed compel an agent to do any of the distasteful actions above, regardless of how strong the agent's initial preference to avoid them.

I believe HULP-exploitation is worrying because many agents do have strong values which they see as part of their identity, and exploiters can subvert these goal structures. Exploiters can compel agents to act in the exploiter's interests and against the agent's own goals. Although decision theory is not concerned with stability or choice of an agent's values, it seems difficult to conceptualise an agent which wouldn't lament its goals being subject to arbitrary modification. Acting for a purpose is part of what it means to *be* an agent, and if this purpose can be revised by literally any other agent capable of presenting a gamble, then our notion of agenthood requires substantial reworking.

How can an agent possibly avoid HULP-exploitation, given that it seems to arise as a natural consequence of expected utility maximisation? I believe we can very slightly alter the norms of decision theory in a way which keeps expected utility maximisation largely intact, and which does not reject either infinite utilities or unbounded utility functions. In the next section I will outline this new approach.

References

[1] Daniel Bernoulli. Exposition of a new theory on the measurement of risk. *Econometrica: Journal of the Econometric Society*, pages 23–36, 1954.

[2] Nick Bostrom. Pascal's mugging. *Analysis*, page 443445, 2009.

[3] Nick Bostrom. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, 22(2):71–85, 2012.

[4] Denis Diderot and Geo Polier de Bottens. *Pensées philosophiques*. Librairie philosophique, 1746.

[5] Ian Hacking. Strange expectations. *Philosophy of Science*, pages 562–567, 1980.

[6] John L Mackie and JL MacKie. *Miracle of Theism*. Oxford University Press, 1990.

[7] Edward F McClennen. Pascal's wager and finite decision theory. *Gambling on God: Essays on Pascals Wager*, pages 115–37, 1994.

[8] Stephen M Omohundro. The basic ai drives. In *AGI*, volume 171, pages 483–492, 2008.

[9] Blaise Pascal and Ernest Havet. *Pensées*. Dezobry et E. Magdeleine, 1852.

[10] Michael D. Resnik. *Choices: An introduction to decision theory*. U of Minnesota Press, 1987.

[11] Nicholas JJ Smith. Is evaluative compositionality a requirement of rationality? *Mind*, 123(490):457–502, 2014.

[12] John Von Neumann and Oskar Morgenstern. Games and economic behavior. *Princeton, N.J.*, 1944.

[13] Eliezer Yudkowsky. Singularity, 2012.