Philosophy Honours Thesis

Adam Chalmers

2016

1 Expected Utility and HULP Problems

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 decisionmaker
- A set of actions the decisionmaker can take
- A set of states: mutually-exclusive propositions which describe ways the world could be
- A set of outcomes that may result from each action, and each action's utility (a numerical measure of each action's relative desirability)

In decisions under risk, the agent has a probability distribution that relates their actions to outcomes. This distribution describes the probability that a certain outcome arises as the result of a certain action, conditional on each particular state that may or may not obtain.

If an agent knows the world's possible states, understands the actions available to them, the odds 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.

Many specific decision algorithms exist, including causal decision theory, evidential decision theory, the dominance principle and expected utility maximisation. Expected utility maximisation 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

$$EU(a) = \sum_{o} P(o|a).U(o)$$

where a is an action and o is any outcome which may arise as a result of that action. In economics, mathematics and computer science, expected utility maximisation is often suggested to be a *norm* of decision theory: a correct and rational mode of decisionmaking that agents should strive to emulate. This is based on two arguments.

Firstly, agents who maximise their expected utility will, by definition, obtain maximal utility in the long run, and given that utility is a measure of an outcome's desirability, expected utility maximisation therefore effectively guides agents towards achieving the most of their desires.

Secondly, expected utility theory is a logical implication of basic axioms of rational choice (citation Von Neumann & Morgenstern, 1944). 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.

In many real-world situations expected utility theory seems to correctly value probabilistic actions such as gambles, medical interventions and business decisions. However, philosophers have devised many decision problems where expected utility seems to dramatically overvalue particular actions and endorse irrational choices. Even worse, this paper shows that expected utility maximisers can be systematically and repeatedly exploited by malicious agents who put them into particular decision problems.

1.1 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.

1.1.1 St. Petersburg Paradox

In the St. Petersburg paradox (citation Bernoulli), an agent is offered a gamble where a fair coin is repeatedly flipped until it lands tails-up. The agent is then paid $\$2^N$ (footnote re dollars), where N is the number of heads that were flipped.

The expected utility of this gamble is $\$(\frac{1}{2}\times 2+\frac{1}{4}\times 4+...)=\infty$, and therefore an expected utility maximiser should be willing to pay any finite amount to purchase it. Millions or billions of dollars are worth almost nothing compared to something with infinite expected utility. However, this seems like a gross overvaluation because the rewards from this gamble depreciate so steeply. 97% of the time, the agent will make $32 or less from the gamble.

1.1.2 Pascal's Wager

(citation 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.

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

By this decision table, 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.

1.1.3 Pascal's Mugging

Many objections to Pascal's Wager dispute specific properties of God or particular accounts of infinity in probability theory. Pascal's Mugging was proposed by (citation Bostrom) in order to focus criticism towards the decision-theoretic aspects of Pascal's Mugging 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. 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.

1.2 HULP Problems

These three problems all share some similar features. In all of them, a gamble's expected utility and its actual intuitive reasonable value appear to diverge. 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 it 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 utilons instead of dollars. 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 utilons. 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 doesn't cancel out 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 decisionmaker 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. (citation)

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 are new. That an infinite utility gai 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.

1.3 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. While Max is walking down the street, Eliza appears and asks him for $100. Max politely declines. Eliza then remembers Max is an expected utility maximser, and offers him the chance to play a St. Petersburg lottery in exchange for $100. Max declines again, explaining that even though this would be a utility-maximising trade for him, he doubts she actually possesses the material backing to host a genuine St. Petersburg gamble (footnote or citation). Eliza counters by explaining that, even if Max's belief in her is 1 in 10^{20}, his tiny confidence is cancelled out by the infinite expected utility of the St. Petersburg gamble if she is telling the truth.

$$EU(payEliza) = (10^{-20} \times (\infty - C)) = \infty$$

$$EU(dontPay) = 0$$

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 either 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 - promising to use her magical powers to grant Max a large reward calculated to dominate his 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 dominates the walk-away action regardless of which particular large finite cost is attached to performing the HULP action. Eliza could ask Max to rob a bank, torture his family, kidnap the president or mutilate himself 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.