An offer he can't refuse: exploiting expected utility maximisers

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 believe decision theory *overvalues* this gamble. For example, [11] writes "What is the fair price for [entry to the St. Petersburg game]? Some argue that the game is a bargain at any finite price, yet few of us would pay even $25 to enter such a game." Many argue it is foolish to value a game at infinite dollars when there's only a 1% chance of winning more than $128 from it.

If Hacking 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

[1]The details of this game are outlined in section 2.3.1, *St. Petersburg Paradox*

problem in more technical detail. However, in sections 3 and 4 I will review Dr. Nicholas Smith's modified decision theory [20] 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 [21] 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 [21]. 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 [2]) 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 total coin flips.

The expected utility of this gamble is $\$(\frac{1}{2} \times 2 + \frac{1}{4} \times 4 + ...) = \infty$ [18], 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' [11].

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 the chance of entering heaven if God exists [17]. If the utility cost of performing religious duty is a finite negative number C, then Pascal's Wager has the following decision table:

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

If we analyse the expected utility of each action, worshipping God appears far more attractive than 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[2]. 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 [12] or the possibility of infinite utility [13]. Pascal's Mugging was proposed by [3] 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 \$5, 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. Suppose the probability of the Mugger telling the truth is p. If the Mugger offers $\$R$ such that $p(R-5)+(1-p)(-5) > 0$, then the agent's expected utility is maximised by giving \$5 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.

[2]This assumes the probability of God is non-zero, i.e. that God is not logically impossible. TODO: find a citation for this (perhaps from Jaynes) or establish that I don't need one. Mark? Your opinion?

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.

If an expected utility maximiser is offered the chance to play a St. Petersburg game, they should pay any finite utility cost to do so because the game has infinite expected utility[3]. Such an agent should be willing to bear any utility cost—trading all their wealth, or murdering large numbers of innocent people (assuming each human life has a finite value)—in order to play the game. As long as the price is finite, the agent must pay it or violate the expected utility maxim. Viewing the St. Petersburg game in this way shows that overvalued games aren't just interesting mathematical trivia, but proof that standard decision theory offers thoroughly alarming advice in certain situations, and are a serious cause for concern.

Pascal's Wager can similarly be used to compel an expected utility maximiser's options. An agent should easily give up any riches or material goods if such costs are necessary for religious observance. Indeed, religious observance can involve any finitely-large utility cost and still outweigh non-observance, due to the presumption that heaven has infinite utility value. 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 -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 [12, 9] to Pascal's Wager point out that the agent doesn't face a binary choice. There are many possible gods—Quetzalcoatl, Jesus, Poseidon—and many of them offer the possibility of heaven without requiring human sacrifice. 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.

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. It is obvious that infinite expected utility outweigh any finite utility cost. 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

[3]This assumes the agent's utility function for money is strictly increasing. This assumption is unnecessary if we instead deal with a modified St. Petersburg game which awards payouts in utiles instead of dollars. For example, a pharmaceutical company could host a St. Petersburg game where the payouts are life-saving vaccines. These variations could track utility in a more accurate way than dollar payouts.

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.

2.5 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 act towards their utility function should thus be careful to avoid HULP-exploitation, and immunity to HULP-exploitation should be a desirable property of 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 advance their interests, or even taking actions which greatly harm 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 $O, u(O) < 0$, where u is the utility function which maps outcomes to utilities. 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 X to Y such that $u(X) > u(Y)$. As it stands, the agent will prefer A to B. If an exploiter offers a reward R for performing A, then provided

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

the agent will now prefer B to A, having had their preferences reversed by the exploiter.

3.2 Should agents avoid having their desires changed?

We have seen that HULP-exploitable agents can have their preferences reordered by malicious others. 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. [16, 4] 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. As [16] argues that for artificial intelligences, "any changes to [their utility functions] would be disastrous to them. If a malicious external agent were able to make modifications, their future selves would forevermore act in ways contrary to their current values." [22] puts the argument 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 doesn't want to kill people."

These arguments may be true of purely goal-focused agents such as some forms of artificial intelligence. But other decisionmakers might not be so intensely attached to their utility function. Many agents desire to change their utility function, and 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[4]. Alex would probably enjoy having their preference against Duke Ellington altered, because enjoying the music of Duke Ellington would provide them with more 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 videos of animal cruelty. 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 mind 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.

Some may object that HULP problems resemble insurance plans. For example, how does Max paying Eliza for a small chance at a large reward differ from Max paying an insurance company for a small chance of insurance settlement? The key difference is that there is a fixed rational price to pay for a given insurance policy, but no fixed price which is rational to pay in a HULP problem.

When buying insurance against a particular event, the event can only be finitely costly. Even a terrible event like the Fukushima tsunami and subsequent nuclear disaster has a fixed (although exceptionally high) cost. Insurance events like earthquakes or robberies have a corresponding probability distribu-

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

tion over possible outcomes, but this distribution is finite. No matter how bad an earthquake can be, it cannot have a literally infinite damage. Therefore, given the probability of disaster occuring and a finite probability distribution over possible damages, you can calculate a fixed maximum price you should pay for a given insurance policy (ignoring the effects of risk- and loss-aversion, hyperbolic discounting, the value of reassurance, certainty, and all the other reasons humans buy insurance policies).

However, with a HULP problem, there is no upper limit to the amount of money an agent should pay to take the HULP action, whether that is buying a St. Petersburg ticket, donating to Church, paying a Mugger, etc. No matter what price you are prepared to offer, the infinite expected utility of the outcome means a rational agent should be willing to pay more. Insurance policies therefore lack the coercive element of HULP-exploitation.

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.

4 Avoiding exploitation

I have shown that agents who value HULP gambles at their expected utility are vulnerable to HULP exploitation, where their preferences can be arbitrarily reordered by an adversarial agent. In this section I will consider the characteristics of agents who cannot be HULP-exploited, and assess whether any of these agents are "strictly better" than standard expected utility maximisers. I will consider an agent's decision theory "strictly better" than expected utility maximisation if it endorses the same actions as expected utility maximisation for non-HULP problems, and avoids endorsing the HULP action for HULP problems.

4.1 Agents who don't maximise expected utility

[TODO: look up EUM in Choices and rework the following paragraph]

Expected utility maximisation is a well-regarded norm of decision theory. As discussed above, part of expected utility maximisation's appeal is that any agent who satisfies a series of reasonable axioms and has a coherent set of preferences over outcomes is an expected utility maximiser[5]. Additionally, agents who choose actions with the highest expected utility maximise their long-term utility[6].

For these reasons, the dominant norm of contemporary decision theory is that agents should maximise expected utility. However, alternative decision theories which do *not* advocate expected utility maximisation have been formu-

[5]TODO: CITATION
[6]TODO: CITATION

lated, partly in response to perceived inadequacies in the way expected utility maximisation values the St. Petersburg and Pasadena games. Works such as [10], [7] and [6] position dominance as an alternative or superior principle to base decisions on.

The dominance principle states that action A is preferred to action B just in case that, for every state S, the utility of the outcome resulting from action A in state S is greater than or equal to the utility of the outcome resulting from action B in S, and that for at least one state A's outcome has strictly greater utility than B's outcome [TODO: CITATION].

Dominance reasoning and expected utility maximisation often endorse the same action. However, there are problems (such as the Pasadena paradox and its variants) where expected utility maximisation offers no advice at all and "falls silent" [15], but dominance reasoning can still accurately evaluate the relative attractiveness of related gambles[7]. There are other problems, such as the simple example below, where neither action dominates the other, but expected utility reasoning shows action A is clearly better than B

Action A: 90% chance of \$100, 10% chance of \$2
Action B: 90% chance of \$800, 10% chance of \$1

Worse still, there are problems where expected utility and dominance reasoning give conflicting advice, such as Newcomb's Paradox [TODO: CITATION]. Some decision theorists, motivated by expected utility's silence on Pasadena problems and conflict with Newcomblike problems, have developed decision theories which therefore either disregard expected utility, or consider it an alternative, perhaps subordinate concern to dominance. For example, [6] espouses a pluralistic approach to decisionmaking where either dominance reasoning or expected utility is used whenever the other theory "falls silent". In both [10] and [7] dominance is the primary concern of decisionmaking, and expected utility concerns can be derived from a generalised decision mechanism.

In HULP problems, the HULP action does not dominate the walk away action. This means agents who follow dominance-based decision theories may not be compelled to choose the HULP action, and thus avoid HULP-exploitation. Unfortunately, because neither HULP nor walk away options dominate the other, the dominance-based theories above will defer to expected utility to decide btween the actions. This means that agents who follow the dominance-based theories above remain vulnerable to HULP-exploitation. Agents who rely solely on dominance (and disregard expected utility entirely) avoid HULP-exploitation, but at the rather devastating cost of being unable to navigate having no grounds to choose B over A in the gamble above. By disregarding expected utility en-

[7]For example, the Altadena game has payoffs one dollar higher than each corresponding payoff in the Pasadena game. Intuitively the Altadena game is more valuable than the Pasadena game, but expected utility theory can't say why. But as [6] demonstrates, the Altadena game dominates the Pasadena game, demonstrating that sometimes "dominance reasoning gives clear advice: choose the Altadena game. Standard (expected utility) decision theory, on the other hand, offers us no advice."

tirely, agents will fail to maximise their profit when faced with many non-HULP decision problems.

4.2 Agents with bounded utility functions

Initially, scholars attempted to solve the St. Petersburg paradox by claiming that a linear increase in wealth should only elicit a logarithmic increase in utility. This would mean the expected value of a St. Petersburg gamble converges on a finite value. However, [14] demonstrates that if your utility function does not have a limit or maximum value, a modified St. Petersburg game with higher (e.g. exponential or superexponential) payoff structure which yields infinite expected value can be constructed. The possibility of these Super-Petersburg games demonstrates the need for *bounded* utility functions.

A bounded utility function has some maximum value. An agent with such a utility function can be *maximally satisfied* such that receiving additional goods does not result in increased utility[8]. Agents with bounded utility functions need not assign infinite or arbitrarily high utility to entering heaven or to the payoff of a St. Petersburg gamble, because at a certain point their utility function becomes (or approaches) saturation. Thus, these agents appear invulnerable to HULP-exploitation.

Bounded utility functions have some intuitive appeal – most agents have a limited capacity to use their goods, and limited mental capacity to reflect upon, perceive, or enjoy the satisfaction of their preferences. However, there are some problems with requiring ideal decisionmaking agents to have bounded utility functions, even if it does allow them to avoid having their preferences arbitrarily reordered.

The first objection is that it hardly seems irrational for agents to have unbounded utility. As [19] writes, "models should adapt to people, not people to models... I know one Paul who, on reflection, does not enjoy linear utility. But why couldn't he have done so?" [20] considers and rejects the need for bounded utility functions on similar grounds, writing that "from a technical view this solution is very attractive... [but] unmotivated: it cuts the utility function to fit the decision theory, whereas what we want is a decision theory which tells us what a rational agent would doand there seems to be nothing irrational about having an unbounded utility function."

A second objection is that even bounded-utility agents can be promised arbitrarily-large rewards if the rewards help extend their bounds. For example, [5] points out that even immortal agents have finite capacity to derive utility from their goods during each moment of time[9]. However, if St. Petersburg pay-

[8] Such a bound may be asymptotic. A bounded utility function may either reach a natural maximum or may asymptotically approach a maximum as the inputs to the utility function approach infinity.

[9] [8] summarise the argument for bounded utility as: "even if money could be spent ínfinitely fast, the human mind still has a limited capacity to process pleasure or enjoyment within a limited space of time." More generally, any agent with finite computation resources can only spend a limited amount of resources enjoying goods or computing their utility function in a given length of time.

offs included not only high-utility goods, but increased capacity to consume these goods, then these payoffs grow arbitrarily large in value[10].

My conclusion is that although bounded-utility agents will be invulnerable to many forms of HULP-exploitation, the notion of agenthood is still entirely compatible with unbounded utility functions. There is no principled reason to exclude unbounded agents from decision theory. We should be able to find a solution to HULP problems for these agents, especially because some bounded-utility agents can still be HULP-exploited given a creative enough decision problem.

4.3 Agents with low-probability cutoffs

Bounding utility attacks the High Utility aspect of HULP problems. Can we attack the Low Probability aspect instead? [1] considers the idea that "events whose probability is sufficiently small are to be regarded as morally impossible," and concludes that its application would yield a finite expected value for the St. Petersburg problem. However, he notes this "principle of neglect of small probabilities... seems extremely arbitrary in its specification of a particular critical probabilities." His further criticism is as follows: for any probability cutoff ϵ, how would one evaluate a decision problem with $\frac{1}{\epsilon}$ distinct possible outcomes each with probability ϵ? One of these outcomes will definitely occur, but they all seem to be ruled out of our consideration by the probability cutoff.

References

[1] Kenneth J Arrow. Alternative approaches to the theory of choice in risk-taking situations. *Econometrica: Journal of the Econometric Society*, pages 404–437, 1951.

[2] Daniel Bernoulli. Exposition of a new theory on the measurement of risk (reprint). *Commentaries of the Imperial Academy of Science of Saint Petersburg, reprinted in Econometrica: Journal of the Econometric Society*, pages 23–36, 1738, reprinted 1954.

[3] Nick Bostrom. Pascal's mugging. *Analysis*, pages 443–445, 2009.

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

[5] Dagobert L Brito. Becker's theory of the allocation of time and the st. petersburg paradox. *Journal of Economic Theory*, 10(1):123–126, 1975.

[10] [8] shows that agents whose utility is bound by their remaining lifespan or temporal resources will value a modified St. Petersburg game at infinite utility if "the individual is given the option of playing the game for both money and time. Along with each dollar the individual wins, he is also given an additional minute of life."

[6] Mark Colyvan. No expectations. *Mind*, 115(459):695–702, 2006.

[7] Mark Colyvan. Relative expectation theory. *The Journal of Philosophy*, 105(1):37–44, 2008.

[8] Tyler Cowen and Jack High. Time, bounded utility, and the st. petersburg paradox. *Theory and Decision*, 25(3):219–223, 1988.

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

[10] Kenny Easwaran. Dominance-based decision theory. *Unpublished manuscript. Retrieved from http://www. ocf. berkeley. edu/˜ easwaran/papers/dec ision. pdf*, 2009.

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

[12] John L Mackie. *Miracle of Theism*. Oxford University Press, 1990.

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

[14] Karl Menger. Das unsicherheitsmoment in der wertlehr. *Zeitschrift fr Nationalkonomie*, 51:459–485, 1934.

[15] Harris Nover and Alan Hájek. Vexing expectations. *Mind*, 113(450):237–249, 2004.

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

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

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

[19] Paul A Samuelson. St. petersburg paradoxes: Defanged, dissected, and historically described. *Journal of Economic Literature*, 15(1):24–55, 1977.

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

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

[22] Eliezer Yudkowsky. Singularity, 2012.