An offer you can't (rationally) 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, [Hacking, 1980] 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

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

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 Dr. Nicholas Smith's modified decision theory [Smith, 2014] 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 [Von Neumann and Morgenstern, 1944] 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 [Von Neumann and 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.

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 deci-

sion 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 [Bernoulli, 1738, reprinted 1954]) 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 $\$(\frac{1}{2} \times 2 + \frac{1}{4} \times 4 + ...) = \infty$ [Resnik, 1987], 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' [Hacking, 1980].

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 [Pascal and Havet, 1852]. 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 [Mackie, 1990] or the possibility of infinite utility [McClennen, 1994]. Pascal's Mugging was proposed by [Bostrom, 2009] 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]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?

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.

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

[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.

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 [Mackie, 1990, Diderot and de Bottens, 1746] 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 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.

2.6 Low Utility, High Probability

Before I analyse HULP exploitation, it will benefit readers to address the inverse of HULP problems: Low Utility, High Probability (LUHP) problems. Each HULP problem has a corresponding LUHP problem which can be obtained by

multiplying the utility payoffs in the HULP decision table by -1. This yields problems where, for example, God sends you to hell $U = -\infty$ if you do not worship him, or a mugger threatens to ruin your life to the degree of some large negative utility (perhaps by committing mass genocide).

I propose that ideal rational agents should treat LUHP problems symmetrically to HULP problems. An agent can be exploited equally well by offering them a chane at heaven or a chance to avoid hell. If corresponding HULP and LUHP problems appear different, or generate different intuitions, I suspect it is merely a product of the asymmetric way human brains model reward and loss. Standard decision theory values profit-making and loss-avoidance equally (assuming any diminishing returns are already factored into the agent's utility function). Motivated by this, I will only address HULP problems in this paper. However, LUHP problems could be used equally well to exploit an expected utility maximiser.

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. [Omohundro, 2008, Bostrom, 2012] 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 [Omohundro, 2008] 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." [Yudkowsky, 2012] 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 al-

cohol 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

[4]I credit this argument for preference change to Dr. Mark Colyvan, who explained it during a meeting we had.

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 distribution 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

Expected utility maximisation is a well-regarded norm of decision theory. This is partly because (as discussed earlier) any agent who satisfies a series of reasonable axioms and has a coherent set of preferences is an expected utility maximiser [Von Neumann and Morgenstern, 1944]. Additionally, agents who choose actions with the highest expected utility maximise their long-term utility. As such, standard decision theory broadly expects rational agents to be expected utility maximisers.

Despite this, alternative decision theories which do *not* wholly advocate expected utility maximisation exist. This is partly because of dissatisfaction over the way expected utility maximisation handles the Pasadena paradox. The Pasadena game involves flipping a coin until it lands heads, and pays $\$-1^n.\frac{2^n}{n}$ where n is the number of preceding tails flips. Because this sequence does not converge, the expected utility of playing the Pasadena game is undefined [Nover and Hájek, 2004]. Expected utility theory falls similarly silent on the Altadena game, which is almost identical to the Pasadena game except its payoffs are \$1 higher in each term.

Decision theorists largely agree that the Altadena game is more desirable than the Pasadena game, but expected utility theory cannot account for this intuition because both games have undefined expected utility. However, because each outcome of the Altadena game is higher than the corresponding outcome of the Pasadena game, it *dominates* the Pasadena game[5]. This demonstrates that sometimes the dominance principle offers advice where expected utility falls silent. This has motivated the construction of decision theories which combine dominance and expected utility ([Easwaran, 2009, Colyvan, 2008, 2006].

One might hope that further integrating dominance reasoning into decision theory could allow agents to avoid HULP exploitation. After all, the HULP action doesn't dominate the walk-away action. Perhaps if agents take dominance reasoning more seriously than expected utility, they won't be compelled to take the HULP action. Unfortunately, as we will see, dominance-based agents face a dilemma: they can either entirely disregard expected utility, which leaves them unable to solve many simple decision problems, or to supplement dominance reasoning with expected utility calculation, which reopens the door for HULP exploitation.

This is demonstrated by considering the ways in which dominance and ex-

[5]put Choices definition here

pected utility reasoning can be combined. Firstly, note that each principle falls silent in different problems. I have already discussed a scenario where expected utility falls silent but dominance reasoning offers advice (choosing between Pasadena and Altadena gambles). However, dominance reasoning also falls silent on the majority of decision problems, where there is no dominating action. For example, consider this simple problem:

Simple Problem 1 (SP1)
Action A: 90% chance of \$100, 10% chance of \$2
Action B: 90% chance of \$1, 10% chance of \$3

Because neither action dominates, dominance reasoning offers no advice, but rational agents should still prefer action A due to its higher expected utility. This is the case for most problems of choice under risk, and therefore it is inadvisable to use dominance reasoning alone without consulting expected utility where dominance falls silent.

Secondly, note that the two principles do sometimes offer conflicting advice. For example, in Newcomb's paradox, dominance reasoning leads agents to choose two boxes and expected utility reasoning leads them to choose one[Resnik, 1987, pg. 110]. Given this, a decision theory which includes both dominance and expected utility reasoning cannot merely use one when the other falls silent: it must also specify which principle overrides the other when conflict occurs.

Now we can examine how dominance-based decision theories address HULP problems. If a theory uses the dominance principle without expected utility, then it is immune to HULP exploitation (because neither HULP nor walk-away actions dominate). However, these agents have no justification to choose A over B in SP1 above, which is symptomatic of their general inability to correctly value probabilistic gambles. Excluding expected utility altogether is simply too much of a sacrifice merely to avoid HULP exploitation.

If instead an agent uses dominance reasoning to supplement expected utility when it fails (for example, when choosing between Pasadena and Altadena gambles), then the agent will be HULP-exploitable, because both HULP and walk-away options have well-defined expected utilities.

If agents use the two principles in the reverse manner, first consulting the dominance principle and then using expected utility if dominance offers no advice, then again they become HULP-exploitable. For neither HULP nor walk-away option dominantes, and therefore dominance theory falls silent in HULP

problems. The agent will then consult expected utility and be advised to take the HULP option.

It appears that dominance-based decision theory cannot satisfactorily protect agents from HULP exploitation. Purely dominance-based agents cannot be HULP exploited, but also cannot make reasonable choices about SP1 or other simple gambling problems. Combined dominance- and expected-utility-based theories like those proposed in [Colyvan, 2008], [Colyvan, 2006] and [Easwaran, 2009] will all endorse HULP actions over walk-away actions.

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, [Menger, 1934] demonstrates that if your utility function does not have a limit or maximum value, a modified St. Petersburg game with higher payoff structure (e.g. superexponential) 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. 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 decision-making 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 seems perfectly reasonable for agents to have unbounded utility functions. While bounded utility functions may certainly have benefits, it hardly seems irrational for agents to have unbounded functions. As [Samuelson, 1977] 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?" [Smith, 2014] considers bounded utility functions as a solution to paradoxes involving infinite gambles, but rejects them 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 do —and 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, [Brito, 1975] points out that even immortal agents have finite capacity to derive utility from their goods during each moment of time, and therefore should have bounded utility functions[6]. However, if St. Petersburg payoffs included not only high-utility goods, but increased capacity to consume these goods, then these payoffs grow arbitrarily large in value[7].

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? [Arrow, 1951, pg. 414] considers the idea that "events whose probability is sufficiently small are to be regarded as... impossible." Intuitively, if we decide to consider events with probability less than a very small number ϵ as impossible, then the expectation of the St. Petersburg game becomes finite (because only finitely many terms will have probability above ϵ, i.e. be considered possible). It would also allow agents to

[6] [Cowen and High, 1988] 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 computational resources can only spend a limited amount of resources enjoying goods or computing their utility function in a given length of time.

[7] [Cowen and High, 1988] shows that agents whose utility is bounded 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."

rule out the possibility of God's heaven, Pascal's Mugger, or any other HULP outcome which falls below probability ϵ.

However, Arrow notes this "principle of neglect of small probabilities... seems extremely arbitrary in its specification of a particular critical probabilities" [Arrow, 1951, pg. 414]. His critique continues: for any probability cutoff ϵ, how would one evaluate a decision problem with $n > \frac{1}{\epsilon}$ distinct possible outcomes each with probability $p < \epsilon$? One of these outcomes will definitely occur, but they all seem to be ruled out due to being below the probability cutoff. Indeed, this precise situation occurs whenever we measure a continuous variable —each specific value has probability zero, but one of them always occurs.

This reasoning demonstrates why it would be unwise to exclude from consideration all events below a certain probability. However, [Smith, 2014] proposes a more nuanced system of probability cutoffs called *truncation* which defeats Arrow's arguments.

Smith's theory is motivated by the idea that, like engineering theory, decision theory should not require infinite precision. Instead, all calculations should account for a degree of 'tolerance' or measurement error. Smith combines this with a norm of decision theory he observes: that "decision makers should ignore (i.e. not factor into their decision making) outcomes with zero probability"[8] [Smith, 2014, pg. 472]. The conjunction of these two ideas leads him to conclude that any outcome whose probability is within a small tolerance of zero (i.e. probability ϵ for some ϵ close to zero) should *also* be excluded from consideration. This leads Smith to the *rationally negligible probabilities* (RNP) proposal:

> "For any lottery featuring in any decision problem faced by any agent, there is an $\epsilon > 0$ such that the agent need not consider outcomes of that lottery of probability less than ϵ in coming to a fully rational decision."

(RNP) is justified by the idea that "in any actual context in which a decision is to be made, one never needs to be infinitely precise in this way —that it never matters" [Smith, 2014, pg. 474]. Clearly, infinite precision matters in some sense, as its presence or absense may change which action one's decision theory endorses. For example, standard expected utility endorses trading any finite

[8]Later in this paper, Smith justifies the idea that "ignore outcomes with zero probability" is a decision theory norm by observing its formalisation within the mathematics of expected utility. Outcomes with zero probability have no effect on a gamble's expected utility and are therefore excluded from consideration under standard decision theory.

sum of money for entry to the St. Petersburg game. However, (RNP) implies there is some maximum amount above which the RNP-following agent should not pay. Clearly, adopting or disavowing (RNP) may change which action an agent chooses.

Smith acknowledges this and claims "factoring in outcomes of lower and lower probability ad infinitum does not make ones decision any better, any more rational," and that allowing or forbidding tolerances on decision making will "lead to different decisions being made —but... they will not be any more rational" [Smith, 2014, pg. 475]. One may object to this on the grounds that this reasoning allows two ideal agents to rationally decide on different outcomes to the same decision problem. A similar objection exists that an agent who changes their tolerance on a decision problem leaves themselves vulnerable to a Dutch Book.

Smith claims neither of these consequences are direct evidence that (RNP)-agents are irrational. I suspect that the rationality of (RNP)-agents depends on their choice of ϵ, which Smith's theory provides no guidance for choosing appropriate values of. In the following chapter I will discuss several ideas to choose ϵ in a principled, non-arbitrary manner, which I hope provides rational grounds for agents to choose different tolerances.

Arrow's objection, discussed above, does not apply to (RNP). Smith requires that $0 < \epsilon \leq \hat{L}$, where $\hat{L}$ is "the (equal-) highest probability assigned to any outcome by [the lottery] L" [Smith, 2014, pg. 479]. This constraint means that if a gamble has n outcomes of probability $\frac{1}{n}$ each, $\epsilon \leq \frac{1}{n}$ and therefore none of the outcomes can be excluded from consideration. Furthermore, ϵ may be chosen after considering a gamble's specific features, whereas Arrow seems to be criticising the idea of choosing a global cutoff point in advance, regardless of the specific problem. Smith agrees with Arrow that "choosing a critical probability threshold once and for all is a bad [idea]"[9].

Agents who adopt (RNP) cannot be HULP-exploited, because they may treat any outcome with probability less than a chosen ϵ as zero. This allows them to value infinite gambles the same as truncated, finite ones, so the St. Petersburg game takes on a finite expected value. Carefully-chosen values of ϵ may defang Pascal's Wager or Mugging by excluding the HULP option from consideration by choosing an ϵ larger than the probability of the HULP reward eventuating. If the possibility of heaven (or of arbitrarily high reward from the

[9]This quote and discussion are credited with gratitute to personal correspondence with Dr. Smith.

Mugger) is treated as zero, an agent's preferences cannot be arbitrarily reordered by offering them a chance at these outcomes.

Adopting (RNP) seems to handle HULP-exploitation better than dominance theories. As we discussed before, dominance theories are forced to either admit HULP-exploitation or fall silent where no action dominates. Fortunately, RNP-agents choose their action by maximising expected utility (of the truncated gamble) and therefore do not fall silent when no option dominates. Instead, they can choose the high-utility action. However, unlike dominance theories, falling back to expected utility reasoning does *not* make them vulnerable to HULP-exploitation, because (as discussed in the preceding paragraph) under the truncated gamble, the HULP action may not be the high-utility action.

(RNP) seems superior to bounding utility, too. RNP offers protection from HULP-exploitation to *all* agents, not just agents who happen to have bounded utility functions, and therefore has no need to argue that bounded utility functions become a norm of rationality. As discussed, limiting utility functions in this way seems to be arbitrary and lack solid motivation. (RNP) is motivated by more general skepticism about infinite-precision decision theory, whereas bounded utility was primarily offered as a specific solution to St. Petersburg and other HULP problems. Therefore I think (RNP) escapes Bounded Utility's charge of being unmotivated. (RNP) can, however, be seen as arbitrary, especially regarding the choice of ϵ. One's choice of ϵ appears more arbitrary than one's choice of an upper bound on their utility function —at least a bound on utility can be (theoretically) derived from the agent's limited capacity to appreciate value and limited time to exist. As mentioned earlier, I believe ϵ can be rationally derived from more general epistemological concerns. If my analysis of ϵ in the following chapter is correct, then (RNP) escapes this criticism.

Furthermore, (RNP) agents cannot be HULP-exploited by offering them a reward in lifespan or computational resources (as bounded utility agents can). This is because any reward, no matter how large, will go unconsidered if its likelihood falls below ϵ.

It seems (RNP) offers a way for agents to avoid HULP-exploitation without incurring the costs or theoretical constraints that adopting dominance theories or bounded utilities does. I will consider RNP to be a candidate solution to HULP exploitation, and in the next section, will continue my analysis of RNP.

5 RNP and principled choices of epsilon

In previous chapters we have seen that HULP-exploitation leaves standard decision theoretic agents vulnerable to having their preferences arbitrarily reordered. This would, of course, greatly interfere with the agent's goals and interests. Therefore, if a principle allows agents to avoid HULP-exploitation without other adverse consequences, it should be adopted as a norm of rationality. I have proposed Dr. Smith's Rationally Negligible Probabilities as a hypothesis which meets these criteria better than competing theories, such as bounded utilities. In the previous chapter, I claimed that bounded utilities were arbitrary and unmotivated, however, similar charges could be levelled against (RNP) in two ways.

Firstly: is the adoption of RNP itself arbitrary? Why insist decision makers should use tolerances in their decision making? In the case of the St. Petersburg paradox, infinite precision can be obtained easily by summing a sequence, a feat which requires only high school level mathematics. Why should infinite precision not be used when convenient?

Secondly: is the choice of epsilon arbitrary? Is there a principled way to choose an appropriate epsilon? If not, then RNP should be criticised for having a free parameter which can be set to arbitrary values in order to produce endorsements of whatever action seems intuitively reasonable to decision theorists.

This chapter will explain why I do not find these critiques convincing. I will discuss two motivations for RNP, which defend it from the first charge of arbitrariness. Each motivation leads to an account of how to choose epsilon in a systematic way, which defends RNP from the second charge of arbitrariness.

5.1 Finite deliberation

It is *prima facie* tempting to believe the following proposition: "all rational agents should consider each hypothesis which could affect their future." This makes a normative claim about rationality, which I will call Consider Each Hypothesis (CEH)[10]. This seems reasonable —surely rationality requires careful consideration of each possible state that could hold, and each state's consequences for the agent's action. One of the strongest objections to RNP would

[10]It is unclear whether the duty is only to consider hypotheses which might affect the agent's future. For example, a photon outside an agent's light cone can have no causal effect on the agent's future on our current best physical theories. However, agents operating with poorer physical theories —or agents who reason with uncertainty over physical theories —may need to consider the photon even if no causal influence is possible.

be that it violates CEH, and that of the two, CEH is a more defensible principle of rationality. To defend RNP I will show that CEH violates an even stronger norm of rationality and therefore should be disregarded.

While CEH appears reasonable, I believe it to be false, because it conflicts with a norm of rationality which is often assumed but rarely explicated: that *decision making procedures should terminate in a finite amount of time.* Call this the *termination principle.* An agent's deliberation should conclude in a finite amount of time —after all, to spend the remainder of one's days pondering a decision is rarely the correct action in a situation.

If an agent obeys the termination principle, they can only do a finite amount of cognitive work during the decision-making process. No matter how intelligent, knowledgeable or quick-thinking they may be, all agents have a finite amount of cognitive resources, and therefore can only run finite computations in a given moment of time. If every hypothesis has some computational cost, then examining an infinite number of hypotheses would violate the termination principle[11].

I believe the termination principle to be an essential norm of rationality. A decision procedure which does not terminate is no decision procedure at all —it provides no guidance, not even claiming ambivalence to all options. It falls silent. Given this, if CEH violates the termination principle then I feel we should disregard CEH. Therefore, any criticism of RNP which claims agents should consider all relevant hypotheses must also be disregarded.

This reasoning not only justifies RNP, but also clarifies it, because it guides us towards a choice of epsilon. If agents can't consider all hypotheses, then naturally, there must be some ordering according to which hypotheses are considered, and a point after which hypotheses are no longer considered. How is this termination point to be chosen?

Consider an agent who is travelling down a road, considering which of two upcoming forks to take. Naturally, environmental circumstances will affect how much time they allocate to making their decision. If the agent is racing in a four horse chariot down rocky terrain, they will need to make their decision quickly to avoid crashing. Alternatively, if the agent is walking at a leisurely pace with no pressing urgency, they can stop and consider the decision for a much longer time.

The available time with which to make a decision will limit how many hy-

[11] Agents with infinite computational resources could perhaps consider infinite hypotheses (assuming their resources and the number of hypotheses have the same cardinality), however, an account of such agents is beyond the scope of this work.

potheses the agent may consider. The chariot-riding agent might only decide based on how wide each fork's path is. The leisurely-strolling agent can consider a range of hypotheses about each path —where it leads to, how long it might be, how scenic a route it might encompass, etc. This suggests one way to choose a cutoff for hypothesis-consideration: based on the available time to decide each hypothesis. If an agent has n seconds to make a decision, and each hypothesis takes approximately t seconds to consider, then the agent can consider an average of $\frac{n}{t}$ hypotheses. I would suggest hypotheses be considered in order of prior likelihood, because (on our best laws of physics) there are infinite hypotheses like "all the atoms constituting the decision agent rearrange into a puddle with *this* specific shape," "all the atoms constituting the decision agent rearrange into a puddle with *that* specific shape," etc, all of which have miniscule probability and merit little consideration.

I suggest that the hypothesis-considering cutoff and the RNP probability cutoff should align: an agent's choice of epsilon should not be significantly lower than the probability of the lowest-probability hypothesis the agent can consider in their available decision time. After all, if an agent isn't going to consider a hypothesis, they can hardly motivate their actions by appeal to that hypothesis' consequences.

TODO: finish this section, strengthen it, I'm not sure if it's true or not.

5.2 Finite-resolution decision theory

In this section, I propose a method for choosing epsilon on purely epistemic and empirical grounds without recourse to expected utility.

I begin by foregrounding how how decision theory presupposes the existence of a deciding agent, a decision problem, mathematical truth, etc. Most of the time, these can all be assumed, however skeptical and empirical reasoning implies there is always a tiny non-negative chance that these conditions fail to hold. Clearly, decision theory would fail to provide helpful, actionable advice if these foundational assumptions were untrue. I call the failure of this assumption the Cognitive Skepticism Hypothesis (CSH). Say these assumptions fail with probability e (for a particular agent or class of agents). Decision theory will fail to provide helpful, actionable advice regarding events with likelihood below e, because it is more likely that CSH holds (and therefore the agent's decision theory fails) than the event actually occurs. Say this other event occurs. We should use decision theory to choose how to respond to it. Yet, if this event with likelihood

$p < e$ occurs, it is more likely that CSH occurred and some kind of decision theory failure has ensued ($p = e$).

I develop the analogy of a microscope or a "probability event horizon" to illustrate my point —just as an electron microscope's mechanism prevents it from examining anything smaller than an electron, decision theory cannot reasonably talk about events whose probability is surpassed by a failure of decision theoretic assumptions. Similarly, just as light cannot escape a black hole's event horizon once crossing it, a hypothesis which falls below the "probability event horizon" cannot ever be reconfirmed by empirical evidence, because it has become less likely than a violation of the assumptions which make empirical methodology possible.

By analogy to the finite resolution of a microscope, I claim each agent's decision theory will have a finite probability-resolution below which it ceases to operate effectively, and as such we should follow the Rationally Negligible Probabilities hypothesis and choose epsilon equal or close to that value. This is justified by examining the mathematical phenomena of "dead hypotheses" discussed in [Jaynes, 2003], which explains how if hypothesis A makes identical predictions to hypothesis B, but has lower prior probability, no amount of evidence could confirm A over B. If some unlikely event E's predictions can be explained equally well by CSH, and CSH has a prior likelihood, then nothing can ever confirm E over CSH, which I believe is good reason to stop considering E.

5.2.1 Presuppositions of decision theory

Before beginning this section, it is important to acknowledge the difference between decision theory and decision making. Decision theory is, as per its name, the *theory* of making decisions. It is not a decision making process itself, rather, it is a collection of rules, norms, algorithms and processes to guide agents in choosing actions which achieve their goals. It is important to recognize this distinction so we may acknowledge that rational choices can be made without decision theory. Even the most rational agents will often make decisions without decision theory —for example:

- If a decision problem is underdetermined, e.g. "we flip a coin. On heads, you win ten dollars. On tails, something else happens."
- If the possible outcomes of an action are unknown, e.g. "we don't know

what this device does. Should we activate it?"

- If an agent has no time to deliberate before making a decision, e.g. a child suddenly runs in front of your car while you're driving down the road.
- If the decision maker doesn't currently have enough mental energy or cognitive resources to spend rationally analysing their choices, e.g. a spy has been captured and tortured for hours before trying to formulate an escape plan.
- If the agent is (perhaps temporarily) unable to think rationally, e.g. someone trying to plan their walk home while on powerful psychadelic drugs.

These examples are intended to show the reader that, in some situations, decision making agents might have both a perfect grasp of decision theory, and a genuine intention to use it towards rational choice, but simply not meet the prerequisite conditions for using decision theory. In other words, *decision theory presupposes certain facts about the decision-maker and their environment.* A non-exhaustive list of these presuppositions includes:

- The agent knows and understands the full specification of the decision problem.
- The agent has both the capability and cognitive resources to consider all relevant states, actions and outcomes and to carry out any calculations their decision theory requires.
- The agent is aware of their utility function.
- The agent has a reliable mechanism for producing thought (e.g. a working brain for a human, reliable computer hardware for an artificial intelligence, functioning biology for an alien, etc).

These are all necessary preconditions for consulting decision theory, and as such most philosophical research into decision theory takes for granted that they hold. However, I believe any agent with physically-real existence must entertain some small degree of uncertainty about whether or not they meet all these conditions. This follows from skeptical reasoning. Imagine an ideal decisionmaking agent (a wise and learned decision theory professor, an alien superintelligence, an AI), who for some reason wanted to double-check and ensure these conditions were met. What could convince the agent that these conditions hold?

Our agent might introspect on their thought process and check that their thoughts are proceeding in a reasonable and correct manner. Unfortunately, this is clearly a circular process because an agent whose thought processes are malfunctioning can't trust those thought processes to notice their own malfunction. As Yossarian reasons in Catch-22, few insane people are aware of their insanity. Similarly, our ideal agent would be justified in believing they know and understand the full specification of, say, Simple Problem 1 above. But we should acknowledge the miniscule chance of error. Perhaps they misread a 1 as a 7, or perhaps some cognitive malfunction (an undiagnosed tumour, misfiring neuron or cosmic ray hitting their CPU) interfered with their calculation of expected utilities. These events are possible, but overwhelmingly unlikely. It seems our ideal agent cannot prove with absolute certainty that the meet our decision theory presuppositions.

Nor should they have to. This entire argument runs similarly to the argument against skepticism —yes, the agent *could* be hallucinating their every experience. They *could* have misread the problem specification, no matter how many times they double-checked their understanding. Their cognition *could* be defective enough that they fail to recognize their cognitive defects. But it is overwhelmingly unlikely that these events have all occurred in conjunction. Even if they have, this hypothesis doesn't guide the agent's action. If our agent knew of their mental failure, what could they possibly do to fix it? The alternate hypothesis —that the agent has reliable mental processing, accurate introspection and comprehension —has the advantage of being both more likely and being *action-guiding*.

Let's consider this formally. Say an ideal rational agent considers the following hypothesis

Cognitive Skepticism Hypothesis (CSH): my reasoning skills are persistently, systematically defective and my understanding of my environment is persistently, systematically wrong.

The more an agent believes this hypothesis, the less useful decision theory will be to them. After all, if an agent doesn't trust their understanding or practice of decision theory, they'll place less trust in the outputs of their decision theory, and be less inclined to take its advice seriously. I think most agents should assign (CSH) very low prior probability, for reasons both practical and theoretical.

Firstly (and practically), cognitive failures of this magnitude are very rare. [Perälä et al., 2007] estimates 3.48% of people experience some sort of psycotic

episode in their lifetime. Only a tiny fraction of these episodes would involve systematic loss of rationality of the (CSH) kind —say, one in a thousand —and far fewer would be persistent enough to cause lasting skepticism in one's decision-making. Based on this, I would conservatively estimate the frequency of CSH-like episodes among humans at $p = 10^{-6}$. Among ideal decision-making agents the frequency is likely to be far lower, because such agents would be better able to discern and process evidence of any partial cognitive defect incurred from psychosis or malfunction.

Secondly (and theoretically), it's a highly complex hypothesis which requires the simultaneous conjunction of many seperate events. Given this, it should be given a low prior probability under many formalizations of epistemology, such as Solomonoff induction or Kolgomarov complexity.

Given this, I believe most agents would have low belief in (CSH), which justifies their continued use of decision theory.

5.2.2 CSH as a lower bound on decision theory

Q: You've shown that agents shouldn't use DT if CSH is true, and that CSH is unlikely. Why should we disregard any event E less likely than CSH? A1: It's more likely that DT become useless (a consequence of CSH) than the event occur. Given this, you should be more likely to distrust DT than to believe the event occurs. So if it occurs, you should be questioning your perceptions or sense of logic more than your A2: If E is less likely than CSH, and CSH makes the same predictions as E, then you could never get evidence for E. It would always turn out to be evidence for CSH.

We've seen that if an agent believes CSH, decision theory becomes useless to them. I believe this is reason to believe the following:

Disregard Sub-Skepticism Hypotheses (DSH): agents should not consider any proposition which is less likely than CSH.

If DSH is true, then the probability of CSH is a good choice for RNP's ϵ value, because considering hypotheses below this likelihood adds no useful information to the agent's decision-making process. It does not improve the rationality of their choice, nor does it further guide their action. This aligns perfectly with [Smith, 2014, pg. 475], "specifying zero tolerance [i.e. disregarding RNP] ... will, in general, lead to different decisions being made —but (the idea goes) they will not be any more rational."

Why do hypotheses less likely than CSH have no useful effect on an agent's

decision? Put simply: if you're willing to consider some hypothesis H less likely than CSH, you're obligated to also consider CSH. But if CSH holds, you don't meet the necessary conditions for using decision theory. The more seriously you take CSH, the less you can trust decision theory, and if you take H seriously enough to let it affect your decision-making, you must take CSH even more seriously, as it is more likely to be true. However great an effect H has on your decision making, CSH must have a larger effect, due to it having both higher probability and higher disruption to your decision theory if it is true. However, agents should never take CSH seriously when making decisions, because it implies the agent can't use decision theory or trust its advice. So if an agent takes H seriously, they should distrust decision theory. This is an excellent reason not to take H seriously. Considering H implies we should distrust decision theory, so given that we are committed to making a decision-theoretic analysis of a situation, we should exclude both CSH and H from consideration. This is exactly the advice of DSH —that we should disregard H on the basis that it is less likely than CSH.

DSH suggests RNP agents should choose $\epsilon = p(CSH)$. I suggest $p(CSH)$ is a lower bound on ϵ rather than a strict equality, because agents may have problem-specific reasons for disregarding a particular outcome O with $p(O) > p(CSH)$. I do not, however, think there's any point setting $\epsilon < p(CSH)$ because, as discussed, any idea less likely than CSH is hardly worth considering because it implies the agent should seriously consider disregarding decision theory. What would be the point of taking this seriously?

5.2.3 Skepticism and dead hypotheses

DSH is supported by [Jaynes, 2003]'s notion of *dead hypotheses*. A dead hypothesis is one which an agent can never confirm over a competitor. Jaynes uses the example of [Soal and Bateman, 1954], which analyses experimental evidence for extra-sensory perception. In one specific example, a woman named Mrs. Stewart correctly guesses the value of randomly-chosen cards far more often than could be reasonably expected if her guesses were formed at random. This experiment confirms the hypothesis H_f (that Mrs. Stewart has psychic powers) over H_{null} (that Mrs. Stewart's guesses were based on random chance), and concludes that therefore psychic powers exist.

However, Jaynes points out that this naive use of probability theory neglected a third hypothesis H_d —that Mrs. Stewart was deceiving the experi-

menters. The observed data has equal likelihood under either H_d or H_f (i.e. both hypothesis are equally capable of explaining or producing the observed data), so if H_d has higher prior probability, it will have higher posterior probability after observing the data too. More generally, if A has higher prior likelihood than B, and both theories make the same predictions, then no observation could ever make an agent update to believe B over A.

What kind of practical insights can agents draw from this? Imagine that one night you have the fairly ludicrous experience of an angel descending from heaven to tell you God has chosen you to be the Messiah. What hypotheses could explain this observation? The hypothesis that you are actually God's annointed Messiah, H_M, is incredibly unlikely. It's a very complicated hypothesis that flies in the face of all observed evidence to date, and on most models it should have a very low prior probability. However, it explains your vision much better than the null hypothesis, so observing the angel should increase your belief in H_M slightly.

Or should it? You could be having a hallucination or psychotic episode. This hypothesis H_P, while still unlikely, ought to have higher prior likelihood than H_M (most agents will grant psychotic episodes are both simpler explanations and more frequent occurrences than real angels delivering genuine messages from God). Both theories are equally capable of explaining your experience, and in fact, equally capable of explaining many subsequent experiences (newfound religious zeal, further visions, dreams of Jerusalem) that may follow. Given this, the angel observation O should indeed transfer probability mass from H_{null} to H_M —but it should transfer even more from H_{null} to H_P, so that $p(H_M|O) < p(H_P|O)$. If you believe the observations of miracles could be explained equally well by psychosis or by divine intervention, then the higher prior probability of psychosis means you should never believe divine intervention over psychosis. H_M becomes a *dead hypothesis*, one that can never rationally rise to consideration above H_P.

This reasoning sheds some light on why CSH should exclude less likely theories from consideration. If a hypothesis H is less likely than CSH, and could be explained by some form of cognitive failure —persistent deception about the world by a malicious actor, hallucination, Descartes' evil demon —encompassed by CSH, then no evidence could convince you of H over CSH. Even if one of H's predictions occurred, it is more likely to have occurred due to cognitive failure (CSH) than H.

If a hypothesis H is dead with regards to CSH (i.e. makes all the same

predictions as CSH but has lower prior probability), then no observation could ever convince you of H over CSH. Even if you observe, say, an angel declaring you the Messiah, you should disbelieve your own eyes and instead believe that you are hallucinating.

This explains why we should disregard hypotheses less likely than CSH. If you observe the data predicted by one such unlikely hypothesis H, it should not promote H over CSH —in fact, it will actually increase the likelihood of CSH. As discussed above, this will result in you lowering your trust in decision theory. This makes considering H unhelpful, because even if you do observe any evidence in favour of it, all you will really achieve is increasing your own uncertainty. You will become less capable of making decisions, as you will have to entertain more and more doubt about your own powers of reasoning.

To put it simply: it's impossible to confirm H. Any evidence for H will raise your belief in CSH, lowering your trust in decision theory. Therefore, rationally planning on the basis that H becomes self-defeating. Decision theory can and does frequently analyse low-probability hypotheses like H, but if H is dead with respect to CSH, then decision theory *does not yield useful advice*. Its advice becomes self-defeating. This is an excellent reason to avoid considering any hypothesis with probability less than CSH's: considering it cannot yield useful decision-theoretic advice. This is why I believe $p(CSH)$ is a good value for ϵ.

5.2.4 Microscopes and the limits of decision theory

Many readers may find the idea that decision theory works poorly on small probabilities very concerning. I suspect my theory clashes with two intuitions readers might have.

Intuition 1 is that the laws of decision theory are supposed to be mathematical truth, and their primitive operations (multiplication, summation, selecting a maximum) work equally well for all numbers, no matter how small. How can we derive a fundamental discontinuity in the way decision theory works when there's no matching discontinuity in its mathematics?

Intuition 2 is that decision theory should work equally well as long as decision agents beliefs and desires meet the necessary structural requirements, i.e. their beliefs obey the laws of probability theory and their utility functions obey the von Neumann-Morgenstern axioms. Having a tiny belief in a very unlikely hypothesis violates neither of these requirements, so this belief should not prevent an agent from using decision theory.

I too possess these intuitions, but I don't believe my theory clashes with them in any serious way. I don't see RNP or the existence of a privileged ϵ as a discontinuity in decision theory; rather I see it as a fuzzy limit on its usefulness. Decision theory is analogous to a microscope. A microscope augments a human; it surpasses the limits of human vision. With a microscope we can see things which are far too small for our limited eyes to perceive. But even the best microscope has limits. Microscopes work by manipulating photons, and at a certain zoom level, the diffraction limit of photons means they are (roughly speaking) too large to resolve an image properly.

That microscopes have finite resolution does not imply light works differently at small scales, or that that our theory of microscopy is flawed, or that they use low-quality parts that aren't precise enough to further resolve small images. It just means the properties of light have certain real-world consequences for microscope design. Light doesn't behave differently beyond the resolution limit, it just stops being useful for this application.

I view $p(CSH)$ as an analogous resolution limit for decision theory. Discovering our decision theory has a resolution limit doesn't imply that probabilities work differently at small scales, or that our decision theory is flawed, or that our mathematics isn't precise enough to further reason about unlikely events. It just means the application of probability theory to the risk of cognitive failure have unintuitive consequences. Neither decision theory nor probability behaves differently beyond $p(CSH)$, they just stop being useful for our applications.

One could also view $p(CSH)$ as a sort of "event horizon". In physics, a black hole is an infinitely dense, maximally small (i.e. point-sized) space where physical force works in a qualitatively different way. Around this tiny point is a large region of space —the event horizon —where the laws of physics make a curious prediction: once an object crosses the event horizon, it can never cross back. I suggest $p(CSH)$ acts like an event horizon: a large region of probability-space around a point ($p = 0$). Once a hypothesis H crosses the event horizon (i.e. once $p(H) < p(CSH)$) it can never cross back out (i.e. we can never again rationally believe $p(H) > p(CSH)$). The curious behaviour of a black hole's event horizon comes from the law of gravity; the curious behaviour of a probabilistic event horizon $p(CSH)$ comes from Jayne's notion of dead hypotheses. Once an object enter's a black hole's event horizon, gravity overpowers even the highest exit speed. Once a hypothesis falls below $p(CSH)$, CSH's higher prior and identical predictions overpower even the strongest evidence.

Neither black holes nor my RNP + CSH theory break the underlying theory

of physics or probability. They merely draw out unusual consequences in exotic real-world applications of the theory to extreme circumstances. Because black holes don't arise in ordinary, Earthbound human existence, their existence was a surprise to the physics community, despite the fact that the well-known general relativity equations already allowed for them. Similarly, because hypotheses below $p(CSH)$ rarely require real-world plans based on their likelihood, the existence of a probability event horizon may be a surprise to decision theorists, including myself, despite the fact that the well-known laws of statistics and probability theory already allow for it.

With this understanding of my theory, I will address the two intuitions I outlined above, which seem to contradict my work. Intuition 1 says we can't derive a disconnect in the application and use of decision theory without a disconnect in its underlying mechanism (probability theory, utility theory and perhaps statistics). If we think about event horizons, we can see that a simple, general, unified theory like General Relativity can predict wildly different behaviour on opposite ends of a region in space. On one side of an event horizon, an agent's future is boundless; on the other side, an agent's future is solely contained within the event horizon. I suggest decision theory is like general relativity: it is not a flaw that the theory yields vastly different behaviours on different sides of the inequality $p(H) < P(CSH)$. This is merely a logical consequence of our theory which we hadn't come across. The same mathematical operations are being applied, but they translate into vastly different physical or practical outcomes for the humans consulting the calculations.

Intuition 2 says that decision theory should work well for all agents whose beliefs are suitably coherent and utilities are suitably structured. Under my theory, decision theory still *works*, it just becomes *unhelpful*. If you are skeptical that you meet the presuppositions of decision theory (having a functioning mind, understanding the decision problem, etc), then naturally you should trust decision theory less.

I have sketched my defense of (RNP) and explained how to choose a value of ϵ which is neither arbitrary nor based on expected utility. In the next section, I will further detail how RNP solves the problem of HULP exploitation, answer questions readers may have and defend my work from some anticipated critiques.

References

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

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.

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

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

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.

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

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

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

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

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

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

Edwin T Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003.

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

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

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

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

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

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

Jonna Perälä, Jaana Suvisaari, Samuli I Saarni, Kimmo Kuoppasalmi, Erkki Isometsä, Sami Pirkola, Timo Partonen, Annamari Tuulio-Henriksson, Jukka Hintikka, Tuula Kieseppä, et al. Lifetime prevalence of psychotic and bipolar i disorders in a general population. *Archives of general psychiatry*, 64(1):19–28, 2007.

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

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

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

S. G. Soal and F. Bateman. *Modern experiments in telepathy*. New Haven, Yale University Press, 1954. URL `http://hdl.handle.net/2027/mdp.39015039379972`. xv, 425 p.

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

Eliezer Yudkowsky. Singularity, 2012. URL `http://www.yudkowsky.net/singularity`.