An offer you can't (rationally) refuse

Systematically exploiting utility-maximisers
with malicious gambles

Adam Chalmers

October 9, 2016

A thesis submitted in partial fulfilment of the requirements for the degree of Bachelor of Arts (Honors) in Philosophy, University of Sydney, October 2016.

Abstract

Decision theory aims to provide mathematical analysis of which choice one should rationally make in a given situation. Our current decision theory norms have been very successful, however, several problems (such as Pascal's Wager, the St. Petersburg Paradox, and Pascal's Mugging) have proven vexing for standard decision theory. In this paper, I show that these problems all share a similar structure and identify a class of problems which decision theory overvalues. I demonstrate that agents who follow current standard decision theory can be *exploited* and have their preferences reordered if offered decision problems of this class. I show that preference reordering is a serious problem, which motivates the search for a decision theory which is immune to exploitation. I find Dr. Nick Smith's theory of Rationally Negligible Probabilities cannot be exploited in this way and discuss why agents should adopt it.

Acknowledgements

This thesis was written on land which belongs to the Gadigal people of the Eora Nation. Their sovereignty was never ceded, so this land always was and always will be Aboriginal land.

My work on decision theory would not be possible without the help of my supervisor, Dr. Mark Colyvan. Thank you for being my tour guide through the wild world of decision theory, for listening patiently while I explored implausible theories, and for your impeccable notes on my work. I have enjoyed your company immensely throughout this year. Thank you for all your help.

To my wonderful parents: thank you for your constant support through honours, but every part of my education. I wouldn't be capable of writing a thesis like this were it not for the lessons and principles that I have learned from you.

A huge thank you to everyone who read and gave notes on early copies of this thesis: Eleanor Gordon-Smith, Jacob Henegan, Isabel Rae Timmerman, and Daniel Kenny, thank you so much for pouring through sixty pages of an unfamiliar field just to help your friend. To my study crew of Tess Lyon, Emma Balfour, Tahlia Chloe, Oliver Moore, Holly McMath and Gerard Inland: your company in Fisher, Schaeffer, the postgrad dungeons or the Eastern Suburbs kept me from collapsing into a quivering mess. You kept me in good spirits during the most difficult project I've ever completed and I know with absolute certainty that I could not have completed this without you.

Contents

1 What is decision theory and why is it broken?

Normative (or ideal) decision theory is the study of how to make the right choice in a given situation.[1] Economists, politicians, scientists, financial planners and doctors all use decision theory to choose which possible action will best let them achieve their goals.

People generally don't need decision theory if they know all the information relevant to their problem. Imagine a doctor choosing between two medical treatments. If she knew treatment A would work and treatment B wouldn't, then her choice is easy. There's no need to consult decision theory in this case. However, if she's 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 more complicated. In situations like this, decision theory offers precise mathematical analysis of how prudent each action is, based on its probable outcomes. It allows people to replace their intuitions — which are often inaccurate — with mathematical guides that quantify risk and uncertainty. Decision theory has proved incredibly effective at solving real world problems, as evidenced by its popularity among mathematicians, economists and scientists. However, scholars have devised some unusual decision problems to which decision theory seemingly provides misguided solutions.

For example, the St. Petersburg lottery is a thought-experimental game with a very small chance of an incredibly high payoff.[2] 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 overestimates the true value or desirability of the

[1] The related study of *descriptive* decision theory focuses on how people *actually* make decisions, rather than how they *ought* to. This paper is only concerned with ideal decision theory.

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

St. Petersburg game — it *overvalues* it. Hacking (1980) for example 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 decision theory is generally considered to be an accurate and helpful way of analysing decision-making. It is an unusually and laudably practical branch of philosophy. However, if decision theory fails to give sensible advice regarding these 'paradoxical' problems, then it requires revision.

I believe decision theory can be modified to better advise agents facing these paradoxes without losing its effectiveness at solving real-world problems.[3] In this paper, I will examine how certain paradoxes can be used to exploit people who strictly follow decision theory. These unlucky people could, 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 sure 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 risk everything for that infinite money, regardless of how slim the odds of actually winning it are.

This is a grave problem for decision theory because exploitation is simple to perform and effectively allows the exploiter to control the victim's actions. This exploitation compromises the victim's agency by forcing them to act towards another agent's interests and against their own. I will show that agents who sys-

[3]In decision theory, an agent is any entity capable of making decisions. This includes people but also many animals as well as some corporations and computer programs.

tematically and unconditionally fall for this kind of exploitation are irrational, and explore ways for rational agents to avoid such exploitation. This section will outline decision theory and its problems. Section 2 will examine the problem in more technical and mathematical detail. In section 3, I will consider several ways agents can protect themselves from this exploitation. In section 4, the best solution (Dr. Nicholas Smith's 'Rationally Negligible Probabilities' notion from Smith (2014)) will be examined in more rigorous detail. Section 5 will address possible objections to my work.

This paper makes four novel contributions to decision theory:

1. It examines connections between several distinct decision theory paradoxes, and shows they are all specific examples of a more general class of problems I call HULP (High Utility, Low Probability).

2. It demonstrates how HULP problems can be used to exploit standard decision theoretic agents by allowing the exploiter to arbitrarily reorder the victim's preferences (I call this HULP exploitation).

3. It critically evaluates several ways agents can protect themselves from HULP exploitation, including one candidate solution (Rationally Negligible Probabilities).

4. It analyses Rationally Negligible Probabilities, combining it with epistemology and statistics to answer some questions about how it is to be applied.

1.1 Expected utility

Normative decision theory is the study of the mathematical processes of making the ideal decision in a range of different scenarios. These scenarios, known as *decision problems*, comprise of:

- An agent (the decision maker), e.g. a person deciding whether or not to wear a raincoat.
- A set of actions the agent can take, e.g. "take a raincoat" and "don't take a raincoat".
- A set of states: possible ways the world could be, e.g. "it's raining outside" and "it's not raining outside".[4]
- A list of outcomes which could occur when the agent takes a given action while the world is in a given state, e.g. "if it's raining and I bring a raincoat, I will stay dry" and "if it's not raining and I bring a raincoat, I will either have to carry it or be hot and sweaty".[5]
- A table of how greatly the agent wants desires each outcome (or equivalently, how valuable the agent thinks each outcome is).[6] Value or desirability are measured in *utility*, e.g. "staying dry is worth 50 utiles, carrying a raincoat is worth -5 utiles, and getting hot and sweaty is worth -20 utiles".[7]

In some situations, the agent knows exactly which outcome will occur when they take a certain action while a certain state holds. These situations are called *decisions under certainty*. In other situations, agents know one of several possible outcomes will result, and have some idea of the probability of each outcome occuring. These are called *decisions under risk*, and they are the focus of this work.

If an agent knows the world's possible states, understands the actions available to them, understands the probabilities of each action producing each out-

[4]Formally, states are mutually-exclusive and exhaustive propositions.
[5]Formally, each outcome is an action-state pair.
[6]Formally, a *utility function* which maps outcomes to real-valued quantities of utility.
[7]In decision theory, the word "utility" is used exclusively to denote a measure of preference. This is distinct from its use in moral philosophy, where it is often a stand-in for pleasure, wellbeing, happiness etc.

come, and has assigned utilities to each outcome, then they can use decision theory to choose which action to perform. 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, if possible,[8] always choose the action with the highest expected utility, which is the average of each possible outcome's utility, weighted by how likely that outcome is. Mathematically, it is defined as

$$EU(a) = \sum_{o \in O(a)} P(o) \times U(o)$$

where a is an action, $O(a)$ is the set of all outcomes which may occur as a result of that action, and $P(o)$ and $U(o)$ are an outcome's probability and utility respectively. Expected utility maximisation is usually considered a *norm* of decision theory: a correct and rational mode of decision-making that agents should try to emulate. This is based on three arguments.

Firstly, agents who maximise their expected utility (and whose beliefs obey the laws of probability theory) will obtain maximal utility in the long run. This is demonstrated in Von Neumann & Morgenstern (1944) which shows that maximising expected utility maximises actual utility. If an agent assigns utilities to outcomes in a way that accurately reflects their desires, expected utility maximisation will effectively show agents the best possible strategy towards achieving their desires.

Secondly, Von Neumann & Morgenstern (1944) show that any agent who meets four simple, uncontroversial axioms of rational choice should be an expected utility maximiser. Therefore, if an agent is not an expected utility max-

[8] In some decision problems, e.g. the Pasadena paradox, actions with undefined expected utility exist (Nover & Hájek, 2004).

imiser, they must be violating one of these axioms of reasonable behaviour. Agents who violate these axioms are vulnerable to Dutch Books and are therefore usually considered irrational.[9] This means most models of rationality conform to the four axioms and are therefore expected utility maximisers.

Thirdly, expected utility has been very successful in the real world. It seems to correctly guide people in a wide range of probabilistic activities — gambling, medical interventions, business decisions, etc. Its wide success in a number of fields and disciplines presents empirical evidence for its usefulness.

For these three reasons, expected utility is a well-regarded norm of decision theory — to the point where "expected utility maximisation" is sometimes used interchangeably with "standard decision theory" (e.g. in Colyvan (2008, pg. 38)). However, scholars have devised many decision problems where expected utility seems to dramatically overvalue certain actions and endorse seemingly irrational choices. I believe this overvaluing is a serious problem for expected utility maximisers, because it leads agents to pay irrationally large amounts for certain games. This means that, despite expected utility maximisation's practical and theoretical attractiveness, it is not a complete theory of rational choice. As I will soon explain, this is because expected utility maximisers can be exploited by offering them an overvalued gamble. Later in this paper, I will discuss ways to modify expected utility maximisation to avoid this exploitation. For now, I will analyse these paradoxical gambles and draw out how they can be used for exploitation.

[9] Hájek et al. (2008) defines a Dutch Book as "a set of bets bought or sold at such prices as to guarantee a net loss." If an agent's beliefs violate the laws of probability theory or their preferences violate the Neumann-Morgenstern axioms, then an agent would willingly pay money for a set of bets which would necessarily lose them money. Vulnerability to a Dutch Book is therefore considered a sign of irrationality.

1.2 Decision theory 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 examples of HULP problems, demonstrate what they all have in common, and use these commonalities to define what a HULP problem is.

1.2.1 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 $\$(\frac{1}{2} \times 2 + \frac{1}{4} \times 4 + ...) = \infty$. Because it has infinite expected utility, expected utility maximisation advises agents to pay any finite amount to purchase it (Resnik, 1987). 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 less than $33 from the gamble. There's only a 1 in 1000 chance the player will make more than $1000. Because it's so unlikely the agent will win significant money, many philosophers feel "few [agents] would pay even $25 to enter such a game" (Hacking, 1980). Given this, it seems patently irrational to pay billions or trillions of dollars for a St. Petersburg ticket — yet, that is exactly what expected utility maximisation advises agents to do. No finite cost is too high to pay, in the eyes of an expected utility maximiser. Expected utility seems to fundamentally overvalue the true worth of the St. Petersburg game.

1.2.2 Pascal's Wager

In Pascal & Havet (1852), religious worship is treated as a decision problem. An agent's decision to commit to (or abstain from) religious observance is motivated by the chance of entering heaven if God exists. 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. An agent's belief in God, however, is fixed at some real number $0 < P(\text{“}God\text{''}) < 1$, therefore the expected utility of worship will always be infinite (because multiplying a real number and infinity yields infinity). This remains true no matter how tiny the agent's estimate of God's existence is.[10] It remains true even if the religious observance is costly or demanding — even if the agent has to pay a large (but finite) amount of utility to perform religious observance, worshipping still maximises expected utility.

1.2.3 Pascal's Mugging

Pascal's Wager was originally proposed as an argument for belief in God, not as a decision theory paradox. As such, 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) to strip

[10] This assumes the probability of God existing is nonzero. Arguments against Pascal's Wager based on the impossibility of God certainly exist (Oppy, 1991) but are not relevant to this discussion.

back the metaphysical elements of Pascal's Wager, so that scholarship could be focused on its implications for decision theory.

In Pascal's Mugging, the agent is confronted by a mugger who claims to be a wizard with powers that can magically multiply money. He asks the agent for a loan of \$5, promising to use his magical powers to give her a fantastic sum of money in return. The agent has no reason to believe in magic, and the mugger offers no evidence of his wizardry, so it appears rational for the agent to reject his offer.

However, the mugger can promise an arbitrarily large amount of money, and our agent's belief in this claim, while low, should be greater than zero. After all, there is *some* chance the mugger is telling the truth — it's incredibly unlikely, but not impossible. 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 his claims.

1.3 Unifying overvalued paradoxes

These three problems share some similarities. Each gamble's expected utility is far higher than its intuitive, reasonable value, i.e. expected utility overvalues them. All involve a choice between two actions:

- An action which has a possible outcome O such that:
  - The utility of O is arbitrarily or infinitely high
  - The probability of O is low

  e.g. buying the St. Petersburg lottery, worshipping God, and paying the mugger.

- An action with a single outcome which has zero utility, e.g. not buying the St. Petersburg lottery, not worshipping God, and not paying the mugger.

I will call the former an (arbitrarily) High Utility, Low Probability action (hereafter "HULP" action), and the latter a "walk away" action. I will also define a HULP problem as any decision problem which contains both a walk away and HULP action. The St. Petersburg paradox, Pascal's Wager, and Pascal's Mugging are all HULP problems. In each HULP problem, the HULP action has arbitrarily or infinitely high expected utility, and the walk away option has zero expected utility.

1.4 HULP problems

HULP problems are interesting because in most decision problems, choosing the action which maximises expected utility is the rational thing to do. Rationality and expected utility maximisation generally go hand in hand. However, in HULP problems, the HULP action maximises expected utility while also seeming irrational. For example, paying $2,000,000 to play the St. Petersburg game maximises one's expected utility, but seems wildly irrational. The definition of the HULP class of decision problems is interesting because it reveals structural similarities between the three previously-discussed paradoxes, and shows how rationality and expected utility maximisation come apart in the same way for each problem.

In each of the above problems, expected utility maximisation instructs agents to choose the HULP action over the walk-away action. However, agents who consistently choose the HULP action in HULP problems leave themselves open to alarming consequences.

If an expected utility maximiser is offered the chance to play a St. Petersburg game, they should pay any finite utility cost (monetary or non-monetary) to do

so, because the game has infinite expected utility.[11] Such an agent should be willing to bear any utility cost — trading all their wealth, or murdering large numbers of innocent people — in order to play the game.[12] As long as the price is finite, the agent must pay it or violate the expected utility maxim, which (according to orthodox decision theory) would be irrational. Yet it appears entirely rational to refuse this steep price for the St. Petersburg game. This demonstrates that expected utility maximisation is *not a universal norm of rationality*, as it advises agents to take irrational actions in HULP problems.

Pascal's Wager can similarly be used to compel an expected utility maximiser's action. 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.

Consider Pascal's Mugging. If the mugger-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.

[11] 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.

[12] Assuming each human life has a finite value.

None of these individual considerations is new. It is obvious that infinite expected utility outweighs any finite utility cost. One could simply bite the bullet and claim an agent still ought to take the gamble, despite the intuition to the contrary. Later in this paper I will argue that biting this bullet and taking the HULP action leaves the agent open to both systematically poor outcomes and also to arbitrary preference reordering, which compromises agency in an unacceptable way.

The fact that these HULP actions appear irrational in a wide range of paradoxes hints that there is a deeper, systematic irrationality to taking HULP actions. We have not identified three separate problems where the expected-utility-maximising action appears irrational. Rather, the HULP class of problems illustrates a fundamental problem with expected utility. To demonstrate this, I will show how expected utility maximisers can be exploited by agents who can present them with HULP problems.

1.5 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 expected 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 thinks there's only a 1 in 10^{20} chance she's telling the truth, the expected utility of paying her is still infinite regardless of her honesty.

$$EU(\text{Pay Eliza}) = (10^{-20} \times (\infty - 100)) = \infty$$

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

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 for Max neither of these are true, and Eliza can thus compel any action from him by offering him the chance at a St. Petersburg lottery in return.[13]

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

[13] One may argue that this demonstrates nothing more than the capacity for informed agents to profit from uninformed agents. Was Eliza successful merely because Max had uncalibrated probabilities? After all, he estimated there was a 10^{-20} chance she was telling the truth, but in fact, she was not. If Max should have assigned probability 0 to her telling the truth, then this becomes a normal case of an informed agent using their superior knowledge to outwit an uninformed agent. However, I believe this is invalid reasoning, because Max should *not* assign probability 0 to Eliza telling the truth. If Max's initial belief in Eliza is nonzero (a reasonable assumption) and he updates his belief in Eliza's truthfulness by Bayesian conditionalisation, then his belief in her, however tiny, will always be greater than zero. This is rational.

Just because Eliza is certainly lying (after all, she does not have enough money to pay off any result from a St. Petersburg game) does *not* mean Max should assign probability 1 to her lying. If that was the case then the only rational probabilities one should assign to propositions about deterministic systems are 0 and 1 — after all, the states either hold with probability 0 or 1.

to carry out these actions, Max would 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.

1.6 Negative HULP problems

Before analysing HULP exploitation, I will note the existence of inverse-HULP problems: High Negative Utility, Low Probability (HNULP) problems. Each HULP problem has a corresponding HNULP problem which can be constructed by multiplying the utility payoffs in the HULP decision table by -1. This yields decision 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 HNULP problems symmetrically to HULP problems. An agent can be exploited equally well by offering them a chance at entering heaven or a chance at avoiding hell. If corresponding HULP and HNULP 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). As my work only deals with *ideal, normative* decision theory, I will only address HULP problems in this paper. However, HNULP problems could be used equally well to exploit an expected utility maximiser.

2 Exploitation

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

- The victim 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 preferences reordered at no cost to the exploiter. A decision theory which forces agents to abandon, invert or arbitrarily mutate their preferences fails at the purpose of decision theory, which is to help agents achieve their preferences to the greatest extent possible.

Agents should therefore avoid HULP exploitation, and immunity to HULP exploitation should be a desirable property of decision agents. If expected utility maximisation leaves agents vulnerable to arbitrary preference reordering, then it fails to guide agents towards achieving their preferences and therefore fails as a norm of rationality.

2.1 How does exploitation work?

HULP exploitation can lead to the victim agent modifying and reordering their preferences in a way which may greatly harm their interests. This is informally demonstrated in the Max/Eliza story above. 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. This is a self-defeating and unacceptable consequence for expected utility maximisation. HULP exploitation allows other agents to alter your preferences to match theirs, so that you thereafter maximise someone else's expected utility and not your own. The entire appeal of expected utility maximisation is that it helps an agent achieve their values, whatever those values are. But if expected utility maximisation renders agents vulnerable to this sort of preference-hijacking, then it fails to fulfill its primary purpose of helping agents achieve their values.

The following section formalises how an exploiter can compel or prevent any action from the victim at will, at no personal cost.

2.1.1 Formalising exploitation

Suppose the victim has some action A available to them, which is certain to cause the undesirable outcome O where $u(O) < 0$. The victim will prefer not to act upon this action.[14]

The exploiter then offers the victim a desirable reward R where $u(R) > 0$ 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:

[14] u is the utility function. An outcome O has utility $u(O)$.

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

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. This means any agent regardless of their wealth or power can exploit the other victim.

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 B, then provided

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

the agent will now prefer choosing B to A, having had their initial preferences reversed by the exploiter. Note that this process is fully general — any number of preferences can be reordered like this. Note also that an agent's strength of preference for an outcome (measured by its utility) is only defined in relation to their prefernce for other outcomes (Von Neumann & Morgenstern, 1944). Want-

ing or not-wanting an outcome is just a matter or preferring it to one's current situation. This means the ability to reorder any two preferences is equivalent to the ability to redefine an agent's utility function. An exploiter can simply swap the victim's preferences until the victim ranks outcomes from best to worst in exactly the way the exploiter would like.

2.2 Should agents avoid having their desires changed?

We have seen that HULP-exploitable agents can have their preferences reordered by malicious others. It may be unclear at first how large a problem this is. 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. HULP-exploitable agents risk having their preferences reordered. But why is this bad?

In computer science and artificial intelligence, 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. Omohundro 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."

As a simple example, imagine an agent who only values animal welfare.[15] The utility she assigns to a state of the world is just the sum of each animal's welfare, and therefore her actions are chosen in order to minimise the world's animal suffering. This agent would greatly resent anyone modifying her utility function U to U', because if she valued animal welfare less in the future,

[15]This example would work just as well if she valued animal welfare primarily, but not exclusively. However, for simplicity and ease of explanation I have used this simpler example.

then she will likely contribute to less animal welfare than if U had been left unchanged. Thus, her intrinsic preference for animal welfare gives rise to an instrumental preference for preserving her utility function.[16] This is why she would greatly wish to avoid HULP exploitation. Reordering her preferences (for example, swapping the utilities she assigns to the current state of the world, and the state of the world where one less animal is in pain) would be unacceptable to her.

Note that this is a purely pragmatic, prudential argument, *not* a moral or ethical one. An agent whose singular value is the creation of paperclips will have just as strong a desire to protect his utility function as the animal welfare activist. Both will resent having their current utility function changed because it sets back the goals valued by that very utility function.

These arguments may be true of purely goal-focused agents (such as some forms of artificial intelligence). But other decision makers might not be so intensely attached to their utility function. Even the most rational humans often find preference change to be desirable and meaningful. For example, few people enjoy the taste of wine when they first try it, but many consider developing such a taste to be desirable and good. Many nicotine addicts wish to remove their preference for nicotine, and spend considerable resources trying to do so. If some powerful agent could reach into their mind and alter their nicotine preference, they wouldn't resent it: they'd be very thankful.

However, these examples do not show that preference changes are unconcerning. While many people wish to develop a preference for wine or remove a preference for nicotine, few people in their right mind wish to develop a preference for eating faeces or mistreating their children. Many types of agents identify with their preferences. Humans define themselves by their values, but

[16] Of course, she may agree to change her utility function if the change secures a commitment from some other powerful agent to produce lots of animal welfare, take lots of animal-welfare-increasing actions, etc.

even certain AIs such as, say, self-driving car programs exist to satisfy their utility function (i.e. to safely drive cars). Often, an agent's preferences are inextricably bound up with the very identity of that agent. HULP exploitation can compel an agent to abandon any principle no matter how cherished, or take up any cause no matter how vile.

Decision theory exists to help agents achieve their goals. Adopting a decision theory should not cause an agent's goals to arbitrarily change whenever a sham priest or shady gambler offers them a far-fetched deal. When agents are HULP exploited, their exploiter can force them to completely restructure their preferences. The notion of agency intrinsically includes acting towards one's goals or preferences. If a decision theory allows the subversion of one's agency, then agents should not adopt that decision theory. Thus, the fact that expected utility maximisation opens an agent up to HULP exploitation is a fundamental flaw in expected utility maximisation.

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 alter the norms of decision theory, keeping the benefits of expected utility maximisation where they can be safely applied, while also protecting agents from HULP exploitation.

3 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 a malicious 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.

3.1 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, whose beliefs obey the laws of probability, and whose preferences are coherent is an expected utility maximiser (Von Neumann & 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, there exist alternative decision theories which do not *wholly* advocate expected utility maximisation. Interest in these alternative decision theories surged after the discovery of a decision problem with undefined expected utility, 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 & 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 (Nover & Hájek, 2004, pg. 241), 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.[17] This demonstrates that sometimes the dominance principle offers advice where expected utility can't. 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, neither HULP nor walk-away action dominates the other. Perhaps if agents take dominance more seriously than expected utility, they won't be compeled 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 expected 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 falls silent on the many decision problems which have no dominating action. For example, consider this simple problem:

Simple Problem 1 (SP1)

[17] Resnik (1987, pg. 9) defines dominance as follows: "an act A dominates antoher act B if, in a state-by-state comparison, A yields outcomes that are at least as good as those yielded by B and in some states they are even better. The dominance principle tells us to rule out dominated acts."

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 decisions 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 falls back to dominance reasoning when expected utility fails (for example, when choosing between Pasadena and Altadena gambles), then the agent will be HULP-exploitable. Both HULP and walk-away options have well-defined expected utilities, and the HULP option's is higher, so this strategy will endorse the HULP action.

If agents use the two principles in the reverse manner, first consulting the dominance principle and then using expected utility if dominance offers no ad-

vice, then again they become HULP-exploitable. For neither HULP nor walk-away option dominates, 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. Theories which combine dominance and expected utility, like those proposed in Colyvan (2008, 2006); Easwaran (2009) will all endorse HULP actions over walk-away actions.

3.2 Agents with bounded utility functions

Early scholarship 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. super-exponential) which yields infinite expected value can be constructed. The possibility of these Super-Petersburg games motivated some scholars to suggest that agents should use 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 reaches (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 protect them from HULP exploitation.

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, pg. 26) 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, pg. 496) 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." Rationality's primary goal is to help agents act towards their utility function. It is not generally supposed to tell agents *which* utility functions to adopt.

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.[18] However, if St. Petersburg payoffs included not only high-utility goods, but increased capacity to consume these goods, then these

[18]Cowen & 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.

payoffs grow arbitrarily large in value.[19]

This means some agents with bounded utility functions are still HULP-exploitable. Consider a human whose utility function is bounded by her finite lifespan. She assigns finite value to the St. Petersburg game, because after a certain number of coin flips she will die of old age. Her utility function is bounded by her finite lifespan, and therefore she escapes HULP exploitation. However, if the St. Petersburg game payoffs include advanced life extension treatments in addition to money, then the game becomes infinite valuable again, because she can live for longer and longer and therefore enjoy more and more utility.

I conclude that although bounded-utility agents are invulnerable to many forms of HULP exploitation, the notion of agency 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.

3.3 Agents with low-probability cutoffs

Bounding utility attacks the High Utility aspect of HULP problems. Can we try attacking 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 positive 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 rule out the possibility of God's heaven, Pascal's mugger, or any other HULP

[19] Cowen & 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."

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 *Rationally Negligible Probabilities* (RNP) 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 the decision theory norm "decision makers should ignore (i.e. not factor into their decision making) outcomes with zero probability" (Smith, 2014, pg. 472).[20] 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 positive ϵ close to zero) should be excluded from consideration. This leads Smith to the rationally negligible probabilities 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 out-

[20]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. Note, however that zero-probability outcomes can affect one's decisions by dominance reasoning: a St. Petersburg game whose payoffs triple should be preferred to a St. Petersburg game whose payoffs double. Although they both have infinite expected utility, the tripling game dominates the doubling game.

> comes of that lottery of probability less than ϵ in coming to a fully rational decision." (Smith, 2014, pg. 472)

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 absence may change which action one's decision theory endorses. For example, standard expected utility endorses trading any finite 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 agents with Rationally Negligible Probabilities are irrational. I suspect that the rationality of RNP-agents depends on their choice of ϵ, but Smith's theory provides no guidance for making this choice. 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 specific features of a gamble, agent or situation, whereas Arrow seems to be criticising the idea of choosing a global cutoff point in advance, regardless of the specific problem.[21]

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. For example, if $\epsilon = 0.01$, then all outcomes over \$128 are treated as though they had probability zero, meaning the St. Petersburg game's new expected value is \$8. Similarly, if an agent's ϵ is greater than their P(God exists), then they can't be HULP exploited by Pascal's Wager, because the outcome of going to heaven is 'truncated'. If the probability of heaven is treated as 0, then the expected value of worship becomes $U(C)$ where C is the (finite) utility cost associated with religious service.

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 action with high expected utility. 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) in the truncated gamble, the HULP action may not be the action with maximum expected utility.

RNP seems superior to bounding utility, too. RNP offers protection from

[21] This discussion credited with gratitude to personal correspondence with Dr. Smith.

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. Bounding all utilities puts the cart before the horse in a way which RNP does not. 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 probability 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.

4 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.

In this chapter I will give a systematic account of how to choose ϵ, which should defend RNP from both charges of arbitrariness. A prima facie intuitive way to choose ϵ is to choose whichever value of ϵ will probably lead the decision-maker to the best outcome. This strategy has clear appeal — agents should generally do whatever leads them to the best outcome. Unfortunately, this means that to choose ϵ you must evaluate the expected utility of the problem under a certain ϵ-cutoff. This reintroduces the very problem that RNP is supposed to

solve.

The challenge, then, is to find a motivated way of choosing ϵ without expected utility reasoning. I propose a method for choosing epsilon on purely epistemic and empirical grounds without recourse to expected utility.

First, I will foreground 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 these assumptions 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 probability 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 ($p = e$) and some kind of decision theory failure has ensued.

I develop the analogy of a microscope or a "probability event horizon" to illustrate my point — just as the diffraction limit of an optical microscope limits the size of objects that can be examined, 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). This 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 an event E has equal likelihood under both the Cognitive Skepticism Hypothesis (CSH) and some other hypothesis H, but CSH has higher prior probability, then E cannot confirm H over CSH. If this is true for all events, then no evidence can confirm H over CSH, which I believe is good reason to stop considering H.

4.1 Ideal and real agents

As stated previously, this paper deals with normative decision theory and rational agents with idealised decision theory. This paper does not examine descriptive decision theory. However, in the rest of this chapter, I will assume these ideal decision theory agents are *physically real* and exist *within an environment.* To keep my analysis fully general, I will not make any claims about the nature of this environment — it could be our universe, some other physically-possible universe, a simulation, a toy environment, etc.

I will also assume that ideal rationality requires neither omniscience nor 100% error-free, reliable physical embodiment. Any agent with physical existence is subject to the laws of physics, which permit the sudden rearrangement of atoms in the agent's brain, or of the photons in the agent's visual field. Even an ideally rational artificial intelligence is subject to a minuscule chance of cosmic rays hitting their RAM and changing a 0 bit to a 1 bit. Even an ideally

rational human being may be misled by a sudden rearrangement of photons through quantum tunnelling. These events are, of course, overwhelmingly unlikely — but they are *possible*, and their possibility does not detract from the rationality of these ideal decision agents.

4.2 Presuppositions of decision theory

Decision theory is not the same thing as 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. Sometimes rational agents need to make a decision but can't consult decision theory. In these situations, it is entirely rational to make a decision without decision theory. For example:

- If a decision problem is underdetermined, e.g. how much should one pay for the following bet: "we flip a coin. On heads, you win ten dollars. On tails, something else happens."
- If an agent has no time to deliberate before making a decision, e.g. a child suddenly runs in front of a moving car, and the driver must decide whether to brake, swerve or handbrake turn.
- 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 psychedelic drugs.

In these situations, it is rational to make a choice without consulting decision theory, because decision theory simply cannot be applied in that situation. These examples demonstrate that just because an agent has both a perfect understanding of decision theory, and the full intention to consult it, doesn't mean they can. Sometimes, an agent will simply not meet the prerequisite conditions for actually using decision theory. *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).

Even ideal decision theory agents who make perfectly rational choices could find themselves unable to meet these requirements. Even ideal decision theory takes place *in the real world*, and it is no failure of rationality that sometimes the real world makes it difficult to think or act.

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 decision-making

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. Our ideal agent would be justified in believing they know and understand the full specification of, say, Simple Problem 1 above. But there is always a tiny 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 they meet our decision theory presuppositions, and therefore, cannot trust the output of their decision theory with absolute certainty.

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 so defective that they fail to recognize their cognitive defects. But it is overwhelmingly unlikely that these events all occur in conjunction. Even if they do, 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 own mind, 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) estimate 3.48% of people experience some sort of psychotic 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 $p = 10^{-6}$. I imagine that among ideal decision-making agents the frequency would be far lower, because such agents should likely 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 separate events. Given this, it should be given a low prior probability by Occam's Razor, or by many formalizations of epistemology, such as Solomonoff Induction.[22]

[22]Solomonoff Induction (Solomonoff, 1964) is a mathematical formalization of prediction. It uses a mathematical representation of Occam's Razor, where simpler explanations are preferred to complex ones, to judge which hypothesis is the simplest explanation of some observed data. As such, it assigns low probability to complex hypotheses. Although Solomonoff Induction is uncomputable computable approximations to it exist, such as Minimum Message Length (Wallace & Dowe, 1999).

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

4.3 CSH as a lower bound on decision theory

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 use $p(CSH)$ as a lower bound on ϵ. I propose a lower bound 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?

4.4 Skepticism and dead hypotheses

DSH is supported by the phenomenon of *dead hypotheses* (Jaynes, 2003). A dead hypothesis is one which an agent can never confirm over a competitor. Jaynes uses the example of Soal & 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 experimenters. The observed data has equal likelihood under either H_d or H_f (i.e. given the observed data, both hypothesis are equally plausible — they 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 some hypothesis A has higher prior probability than some hypothesis B, and both hypotheses make the same predictions, then no observation could ever make an agent update to believe B more than A.

What kind of practical insights can agents draw from this? Imagine that one night you see an angel descend 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 anointed 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 lower bound for the value of ϵ.

Note that this bound on ϵ is motivated purely by epistemic and empirical concerns — *not* by utility. Therefore it avoids the circularity of choosing ϵ with reference to the stakes of the problem. This defends RNP from the charge of arbitrariness, making it an attractive candidate for a decision theory which protects agents from HULP exploitation.

4.5 Objections, microscopes and black holes

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 am sympathetic to 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 the fundamental norms or underpinnings of decision theory; rather I see it as a limit on its usefulness to decision makers. 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. Other particles like electrons can be used in situations where light is unusable.

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 conflict with the underlying theory of physics or probability. They merely draw out unintuitive 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 thesis. 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 defence 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.

5 Discussion and anticipated objections

In this section, I will respond to anticipated questions or concerns readers may have.

5.1 Overvaluing St. Petersburg

My thesis still overvalues the St. Petersburg game. Bearing in mind that $p(CSH)$ is a tiny number, perhaps less than 10^{-10}, agents who follow my proposal and use RNP with $\epsilon \geq p(CSH)$ will only truncate the St. Petersburg lottery after an extraordinary number of terms. Many would judge this value as still vastly higher than the genuine or rational value of the St. Petersburg game. It appears my proposal does not solve the problem of overvaluing the St. Petersburg game.

Response

Yes, under my theory the St. Petersburg game still has an absurdly high value. But my goal is not to resolve the overvaluation of the St. Petersburg game. My goal is to find a decision theory which cannot be HULP exploited. My proposed solution only requires that agents value the St. Petersburg game at some finite number, rather than an infinite or arbitrarily large value. Some values of ϵ will result in an RNP-agent overvaluing the St. Petersburg game, but provided $\epsilon > 0$ the St. Petersburg game cannot be used to HULP exploit them. Agents who overvalue the gamble are violating rationality in a different way to agents who can be HULP exploited by the gamble. Each specific HULP problem might be overvalued in a different way, and I do not aim to address their specific problems. I only aim to characterise and resolve issues of HULP exploitation.

5.2 HULP is an empty set

Readers may believe each HULP problem I've mentioned has its own problem-specific solution. If so, readers would likely see RNP as an unnecessary departure from standard decision theory — after all, we don't need to solve HULP exploitation with RNP if it can already be solved within the standard decision theory framework. Why look for a general solution to HULP exploitation instead of known arguments against each HULP problem?

After all, there are many standard arguments against Pascal's Wager: arguments against infinitely-valuable outcomes (McClennen, 1994), arguments for many gods (Saka, 2001), for the impossibility of god (Oppy, 1991) and the invalidity of its logic (Hájek, 2003). The St. Petersburg paradox has been attacked using risk-aversion (Weirich, 1984) or claiming an infinite series of coin-tosses is impossible (Jeffrey, 1990, pg. 154). Even the relatively obscure Pascal's Mugging has attracted problem-specific solutions (Baumann, 2009). Perhaps there are no HULP problems, only problems which *seem* HULP-like before they're "solved" with further philosophical analysis. If there are no actual HULP problems, then there is no need to burden our decision theory with the addition of RNP.

Response

Yes, many HULP problems have their own specific "solution" which purports to explain why standard decision theory assigns finite value to the gamble. However, each solution is only satisfying as a *specific solution* to *one specific problem*. They do not aim to solve HULP exploitation generally. The Principle of Uniform Solution states "it is natural to expect all the paradoxes of a single family to have a single kind of solution," (Priest, 1994, pg. 32). In section 1.3 I gave the necessary and sufficient conditions to classify a decision problem as HULP: having a walk away action (whose sole outcome has zero utility) and

HULP action (which has at least one low-probability outcome with arbitrarily high utility). It would be strange and unsatisfying if these problems were all to have wildly different solutions. In section 4, I analysed several general solutions (bounded utilities, dominance-based decision theories and RNP), each of which was a uniform solution to all HULP problems. I champion RNP only because it is the best of these general solutions.

Another way to think about this is that none of these problem-specific solutions deal with HULP exploitation. Yes, Pascal's Wager leads to agents serving God, but serving God is not HULP exploitation. It is merely one way of achieving HULP exploitation. Similarly, the St. Petersburg paradox can be explained away, but the explanation dispels only this specific HULP problem. It does nothing to explain why HULP exploitation itself is impossible, or to help agents facing HULP exploitation. Even if we grant that the extension of HULP is empty, the intension demands analysis. Why should all these problems just *happen* to be non-issues in different ways? A general solution is required.

Finally, no philosophical consensus exists regarding which of these problem-specific solutions are valid (if any). If you grant that even one valid HULP problem exists, you will require a decision theory which avoids HULP exploitation. As research into decision theory grows, we are likely to discover new HULP problems — or new variants on old HULP problems — which evade the solutions of older problems. We have no guarantee that every single HULP problem has an iron-clad argument against it. So again: a more general HULP solution is needed. If not RNP, then some other solution must be formulated.

5.3 Dead hypotheses contradict Bayesianism

The dead hypothesis theory appears to contradict a core Bayesian idea: that if an agent updates their belief according to Bayes' theorem, then as they increasingly

observe data, their beliefs will become increasingly accurate. As Resnik (1987, pg. 56) puts it, "large amounts of data bearing on the truth of p can "wash out" poor initial probability estimates." For example, suppose ESP (extra-sensory perception, psychic power) exists. Any agent for whom H_{ESP} is a dead hypothesis would never believe in ESP. No matter how much supporting evidence was presented to them, the agent would continue to believe they were hallucinating or being deceived. This contradicts the Bayesian idea that we should converge towards the truth if we apply Bayes' rule. Contradicting Bayesianism is a good reason to reject a line of statistical reasoning, and therefore my explanation of how to choose ϵ does not stand.

Response

Jaynes anticipates this confusion and explains why the dead hypothesis phenomena is actually a prediction of Bayesian statistics, if not Bayesianism. According to Jaynes, the idea "whatever the new information D, it should tend to bring different people into closer agreement with each other" is actually a confused misunderstanding of Bayesian statistics (Jaynes, 2003, pg. 127). This is because it fails to take agents' priors into account.

For example, consider two perfect Bayesians, Mr. A and Ms. B. They both observe that *Green Left Weekly* reports a common pharmaceutical has recently been declared unsafe. Call this observation D, and the hypothesis "the drug is genuinely unsafe" H. Mr. A thinks *Green Left Weekly* is very reliable. For him, $P(D|H) = 0.9$ and $P(D|\bar{H}) = 0.1$. Ms. B distrusts *Green Left Weekly*, and is inclined to believe the opposite of whatever they print. For her $P(D|H) = 0.4$ and $P(D|\bar{H}) = 0.6$.

Before reading the report, neither had any idea whether the drug was safe or not: $P(H) = 0.5$. After reading the report and applying Bayes' Theorem, Mr. A revises his estimation to $P(H|D) = 0.9$, and Ms. B revises hers to $P(H|D) =$

0.4. Their beliefs have therefore diverged, even though they have observed the same evidence. This demonstrates how a straightforward application of Bayes' Theorem contradicts the supposed "Bayesian" notion that observations always lead to increased accuracy of belief. After all, the drug is either safe or unsafe, so either Mr. A's or Ms. B's posterior beliefs updated in the wrong direction.

The confusion arises because of this fallacy:

> "If a piece of information D supports S, the idea goes, then all agents who learn D should increase their belief in S. We committed a subtle form of the mind projection fallacy by supposing that the relation D supports S is an absolute property of the propositions D and S ... whether D does or does not support S depends on our prior information. The same D that supports S for one person may refute it for another" (Jaynes, 2003, pg. 131)

This shows the idea quoted in Resnik is not a mathematical prediction of Bayesianism. If an agent's priors are particularly malformed, evidence which seems to support a proposition may actually support its negation. If an agent could be convinced that no other hypothesis could explain the observed ESP — if you could reduce their belief in all other deception or hallucination hypotheses — then they would eventually come around to ESP. It may not be contingently possible to present them with strong enough empirical evidence to overcome their skepticism, but this is no fault of the statistical reasoning.

5.4 Lower-bounded ϵ ignores grave risks

A lower bound on ϵ means we can no longer reason decision-theoretically about a range of improbable events. Consider a new nuclear reactor design, and let X be the event of its catastrophic meltdown. If $P(X) < \epsilon$ then, on my theory, we

should ignore the risk of meltdown. This seems unwise — surely it's rational and prudent to consider the risk of meltdown.

Response

Firstly, I believe that most plausible hypotheses are *far* more probable than CSH, and therefore likely above ϵ probability. CSH is an incredibly bold proposition. It is almost equivalent to asserting the existence of Descartes' malicious demon. Almost all risks we encounter in daily life will be far above $p(CSH)$.

Secondly, most important real-world decisions are made by multiple people. When multiple agents make a decision the chance of CSH is greatly reduced, because the odds of CSH being true for each agent are independent. I believe there is a possibility for group agents or group decision-making structures to therefore set a lower value of ϵ.

Imagine several agents who agree on a decision theory and problem specification each consult decision theory separately. If they each suffer CSH, their answers are likely to vary (because there are many different ways to fail the presuppositions of decision theory, and many different ways to get the wrong answer from decision theory under CSH). On the other hand, if only a tiny fraction of them suffer CSH, then the vast majority of their answers will be identical, restoring confidence in the decision-making process. This account remains speculative, however future work could extend my work on ϵ to group decision processes.

5.5 RNP disallows strict ϵ values

In (Smith, 2014, pg. 476) Smith explicitly denies the idea of a universal ϵ cutoff, writing that there cannot be "some probability threshold such that no-one need ever — in any decision problem — consider outcomes — of any lottery — whose

probability lies below this threshold." Has my proposal done exactly this?

Response

I agree that a universal cutoff would be undesirable, but my account of ϵ is not a universal cutoff. I define $\epsilon = p(CSH)$, which will depend on both the agent and their environment. Some agents will be more likely to undergo this complete cognitive failure. Some environments will be more reliable and less likely to completely mislead an agent or damage their cognitive capacity. I believe pinning ϵ to an agent's prior for CSH aligns with Smith's idea that "for another decision problem and/or another lottery and/or another decision maker, it might be a different ϵ" (Smith, 2014, pg. 472).

6 Evaluation

This paper makes four contributions to decision theory:

1. Identifying the HULP class of decision problems and showing Pascal's Wager, the St. Petersburg Paradox and Pascal's Mugging are all HULP problems.

2. Showing that malicious agents can arbitrarily reorder the preferences of an expected utility maximiser by presenting them with HULP decision problems (a process I call HULP exploitation).

3. Analysing solutions to HULP, both successful (Rationally Negligible Probabilities) and unsuccessful (dominance-based decision theory, bounded utility functions).

4. Exploring RNP: finding justified values of ϵ without recourse to expected utility; analysing what this can tell us about the limits of decision theory.

I believe the first and second contribution are the most valuable. Readers may disagree with my account of the choice of ϵ, they may dislike RNP, or they may propose their own solutions to HULP exploitation. However, I feel the value of HULP and HULP exploitation themselves stand on their own as serious decision theory problems worthy of further research. Readers might not think that RNP is the solution to HULP problems, but I hope most of them will agree that HULP is a genuine problem crying out for a genuine and uniform solution.

Decision theory is one of 20th century academia's great successes, and one of the few ways for humans to reason about minds greater than our own. Ideal decision theory is more than just an inspiration goal for human reasoners to aspire to. It is also our best guess about how to design and instruct other minds. When we design robots, artificial intelligences, corporate structures and

governance rules, we design them with reference to our best models of decision making. Ideal decision theory does more than just inspire humans: it guides and instructs our mind children, the non-human entities we dream and create.

Viewed from this perspective, HULP exploitation is a serious problem which urgently requires solving. If robots, artificial intelligences or suitably elaborate corporations can be exploited by sham priests, shady gamblers or muggers in wizard costumes, we are in serious trouble. We entrust these systems with large amounts of power, and if that power can be hijacked not through violence or threat but the mere offer of a gamble, we risk huge harms. RNP is my best shot at protecting agents from HULP exploitation.

The most unorthodox work in this paper revolves around my case for choosing ϵ based on skepticism. I did not expect to reach this conclusion when I started my work, but I stand by it. Even the most ideally rational entity must concede there is a tiny chance of skepticism being correct. In most cases, this tiny probability can be safely disregarded. However, when we deal with hypotheses at the very edge of credibility, skepticism cannot be ignored. Simple particle physics calculations often ignore gravity, because at particle scales its force is so weak that it hardly affects one's calculations. However, at extreme scales gravity starts to take effect again, and ignoring it will ruin one's calculations. I suggest ideal agents should the skeptical hypothesis the same way: discount it when its effect on their decisions is negligible, but at extreme scales, include it because it can and should affect their decisions.

I find HULP exploitation fascinating because our human intuition often clashes with the maths of decision theory. Rarely, however, do human intuitions win. Usually we amend or ignore our intuitions in favour of the maths. HULP, I believe, is one of the rare cases where human intuition scores a clear win. The possibility of HULP exploitation gives us a clear guide for future work.

It is a sign, like vulnerability to Dutch Books, that a decision theory requires amending. As humans, we are free to ignore decision theory and follow our gut if we believe it's the right thing to do. But there exists a growing number of agents who share our planet and do not have this capacity. We are now in the business of creating systems which are shackled to the decision theory they are born with. As we trust these systems with larger and larger decisions, we must ensure we can trust their decision making.

References

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

Baumann, P. (2009). Counting on numbers. *Analysis*.

Bernoulli, D. (1738, reprinted 1954). 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*, (pp. 23–36).

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

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

Brito, D. L. (1975). Becker's theory of the allocation of time and the St. Petersburg paradox. *Journal of Economic Theory*, *10*(1), 123–126.

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

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

Cowen, T., & High, J. (1988). Time, bounded utility, and the St. Petersburg paradox. *Theory and Decision*, *25*(3), 219–223.

Diderot, D., & de Bottens, G. P. (1746). *Pensées philosophiques*. Librairie philosophique.

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

Hacking, I. (1980). Strange expectations. *Philosophy of Science*, (pp. 562–567).

Hájek, A. (2003). Waging war on Pascal's Wager. *The Philosophical Review*, *112*(1), 27–56.

Hájek, A., et al. (2008). Dutch book arguments. *The Oxford handbook of rational and social choice*, (pp. 173–195).

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

Jeffrey, R. C. (1990). *The logic of decision*. University of Chicago Press.

Mackie, J. L. (1990). *Miracle of Theism*. Oxford University Press.

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

Menger, K. (1934). Das Unsicherheitsmoment in der Wertlehr. *Zeitschrift fr Nationalkonomie*, *51*, 459–485.

Nover, H., & Hájek, A. (2004). Vexing expectations. *Mind*, *113*(450), 237–249.

Omohundro, S. M. (2008). The basic AI drives. In *AGI*, vol. 171, (pp. 483–492).

Oppy, G. (1991). On Rescher on Pascal's Wager. *International Journal for Philosophy of Religion*, *30*(3), 159–168.

Pascal, B., & Havet, E. (1852). *Pensées*. Dezobry et E. Magdeleine.

Perälä, J., Suvisaari, J., Saarni, S. I., Kuoppasalmi, K., Isometsä, E., Pirkola, S., Partonen, T., Tuulio-Henriksson, A., Hintikka, J., Kieseppä, T., et al. (2007). Lifetime prevalence of psychotic and bipolar I disorders in a general population. *Archives of general psychiatry*, *64*(1), 19–28.

Priest, G. (1994). The structure of the paradoxes of self-reference. *Mind*, *103*(409), 25–34.

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

Saka, P. (2001). Pascal's Wager and the many Gods objection. *Religious Studies*, *37*(03), 321–341.

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

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

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

Solomonoff, R. J. (1964). A formal theory of inductive inference. Part I. *Information and control*, *7*(1), 1–22.

Von Neumann, J., & Morgenstern, O. (1944). Games and economic behavior. *Princeton, N.J.*.

Wallace, C. S., & Dowe, D. L. (1999). Minimum message length and Kolmogorov complexity. *The Computer Journal*, *42*(4), 270–283.

Weirich, P. (1984). The St. Petersburg gamble and risk. *Theory and Decision*, *17*(2), 193–202.