Tiny Probabilities of Vast Value

By

Petra Kosonen

Worcester College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

in Philosophy

5 July 2022

Abstract

The topic of this thesis is how we should treat tiny probabilities of vast value. This thesis consists of six independent papers. Chapter 1 discusses the idea that utilities are bounded. It shows that bounded decision theories prescribe prospects that are better for no one and worse for some if combined with an additive axiology. Chapter 2, in turn, points out that standard axiomatizations of Expected Utility Theory violate dominance in cases that involve possible states of zero probability. Chapters 3–6 discuss the idea that we should ignore tiny probabilities in practical decision-making. Chapter 3 argues that discounting small probabilities solves the 'Intrapersonal Addition Paradox' and thus helps avoid the Repugnant Conclusion. Chapter 4 explores what the most plausible version of this view might look like and what problems the different versions have. Chapter 5 focuses on one type of problem, namely, money pumps. The Independence Money Pump, in particular, presents a difficult challenge for those who wish to discount small probabilities. Finally, Chapter 6 discusses the implications of discounting small probabilities for the value of the far future.

Word count: 70983

Acknowledgments

I am extremely grateful to my supervisors, Andreas Mogensen and Teru Thomas, for their invaluable feedback, insight and support.

I am also grateful to Tomi Francis, Gustav Alexandrie and Johan Gustafsson for their helpful comments and many discussions. In addition, I wish to thank the FTX Foundation and my friends in Nassau for making the last six months of writing this thesis a great experience.

I was financially supported by the Jenny and Antti Wihuri Foundation and the Forethought Foundation for Global Priorities Research, and then by Open Philanthropy and the FTX Foundation.

I also wish to thank the users of Manifold prediction market who gave me (as of 4th June 2022) an 89% chance of submitting this thesis on time.

There is a concept which corrupts and upsets all others. I refer not to Evil, whose limited realm is that of ethics; I refer to the infinite.

– Jorge Luis Borges

Contents

Introduction

Tiny Probabilities of Vast Value

ABSTRACT: This chapter explores different approaches to cases that involve tiny probabilities of huge payoffs. The main approaches discussed are Probability Fanaticism, Boundedness and Probability Discounting. First, the chapter discusses two arguments for maximizing expected utility: the long-run argument and representation theorems. Next, it investigates Probability Fanaticism, on which tiny probabilities of huge positive or negative payoffs can have enormous positive or negative expected utility (respectively). Various arguments for and against Probability Fanaticism are discussed. Then, the chapter considers Boundedness, namely, the idea that utilities are bounded. Finally, the chapter discusses Probability Discounting, on which tiny probabilities should be ignored in practical decision-making. Some other approaches are also discussed briefly. The chapter concludes that the paradoxes involving tiny probabilities of vast value show that some intuitively compelling principles of rationality must be given up.

1 Pascal's Hell

In the beginning, on a small planet in the Solar System, in the Milky Way galaxy…

Satan: I have an offer for you, Pascal, as I have heard that you might be interested in a small probability of a huge payoff.

Pascal: Anything that maximizes expected utility!

Satan: Great! And your utility function is unbounded, am I right?[1]

Pascal: Yes, and additive in terms of people's happy days of life.

Satan: Excellent. So, the offer is this: I will flip a coin, and if it lands on heads, I will help humanity settle on new planets in faraway galaxies and live in bliss until the heat death of the Universe.[2] Until the heat death happens, it will be like heaven. But if the coin lands on tails, then everyone on Earth will suffer excruciating pain for the next fifty years. That's the offer. If you decide to accept it, I will return to Earth every fifty years and give the same offer until either you (or your descendants) refuse the offer, the coin lands on heads, or the Sun expands and makes life on Earth impossible. If you decide not to accept the offer, humanity will live its earthly existence as mere mortals until life on this planet is no longer possible (humanity will not be able to expand out from Earth without my help!)

Pascal: Your offer sounds great—even odds of Utopia! And if we don't win this time, we'll almost certainly win eventually.

Satan: Oh, pardon me, I forgot to say that my coin is somewhat biased. If you

[1] The utility function does not necessarily have to be unbounded for this case to work—it is enough that the upper bound is very high and the lower bound very low. Chapter 1 of this thesis shows that standard axiomatization of Expected Utility Theory require a bounded utility function.

[2] This is the fate of the Universe in which Pascal and Satan live.

accept all the (20 million) offers, the probability of heads happening at least once is one-in-a-googolplex. I admit the odds aren't great. But if the coin lands on heads, I will create a thousand googolplex happy Earth-like planets.

Pascal: Not to worry, the offer is still amazing. The expected value of taking those gambles is clearly greater than the expected value of rejecting them. Actually, its expected value might even be greater than the expected value of the offer I initially thought you were making… So, I'm positively surprised.

Satan: Oops, I made a mistake. I read the wrong page. The instruction manual (*Creating Hell*) says that the probability of heads ever happening on Earth is one-in-Graham's-number. But it is in my power to create any finite number of happy Earth-like planets, so I believe I can still give you a good offer. If the coin lands on heads, I will create a million Graham's number of happy Earth-like planets.[3]

Pascal: Now your offer is even better! Although I dread the almost certain torture for everyone on Earth for the next billion years, the expected value of your offer is far greater than the expected value of not taking it. So, rationality compels me to accept it.

Pascal and Satan then agree on the deal, and Satan flips the coin. Unsurprisingly, it lands on tails.

Satan: You and everyone on Earth will now suffer excruciating pain for the next fifty years.

Pascal: Oh well. I made the right choice, given the information I had. And the future is still great in expectation. Thank you for your offer.

[3]The Universe Pascal and Satan live in is much larger than our Universe.

Satan: I'm always happy to help. See you again in fifty years!

Pascal: See you in fifty (long) years! You are always welcome here.

Satan: I never imagined persuading people to enter (finite) hell would be this easy…

* * *

So Satan traveled from one planet to another, and the inhabitants of those planets—also expected utility maximizers with unbounded utilities—always accepted his offer. And they all lived happily ever after (in expectation). But according to Satan's instruction manual, the probability of the coin ever landing on heads was merely one-in-a-googolplex, so the Universe was almost certain to be void of joy and laughter.[4]

2 Maximizing expected utility

The topic of this thesis is how we should treat tiny probabilities of vast value. This chapter goes over different possible approaches. I will start by considering the idea that rational agents maximize expected utility. Two arguments for maximizing expected utility will be discussed: the long-run argument and representation theorems. I will then present two puzzling cases that involve tiny probabilities of huge

[4]This dialogue is based on Pascal's Mugging by Bostrom (2009), which in turn is based on informal discussions by various people, including Yudkowsky (2007*b*). Pascal's Mugging is similar to Pascal's Wager, except that the former does not involve infinite utilities. Pascal (1958) famously argued that one should believe in God because of the possibility of gaining an infinitely good payoff in Heaven: "Let us weigh the gain and the loss in wagering that God is. Let us estimate these two chances. If you gain, you gain all; if you lose, you lose nothing. Wager, then, without hesitation that He is."

payoffs. The subsequent sections explore expected-utility maximization with unbounded and bounded utility functions, as well as alternatives to expected-utility maximization.

2.1 The long-run argument for maximizing expected utility

According to standard decision theory, a rational agent always maximizes expected utility. An act's expected utility is calculated by summing the utilities of its possible outcomes weighted by their probabilities of occurring, where 'utility' measures how preferable (or valuable) some outcome is compared to the alternatives. Let EU (X) denote the expected utility of prospect X, and let $X \succsim Y$ mean that X is at least as good as Y. Also, let O be the set of possible outcomes, $p_X(o)$ the probability of outcome o in prospect X and $u(o)$ the utility of o. Then, more formally, Expected Utility Theory states the following:

> **Expected Utility Theory:** For all prospects X and Y, $X \succsim Y$ if and only if EU(X) $\geq$ EU(Y), where
>
> $$\text{EU}(X) = \sum_{o \in O} p_X(o)u(o).$$

Why should one accept Expected Utility Theory? One argument for maximizing expected utility—the long-run argument—states that expected-utility maximization is the best policy in the long run. This is because, in the long run, the average amount of utility gained per trial is overwhelmingly likely to be close to

the expected value of an individual trial.[5] However, it is not certain that the average utility gain per trial would be close to the game's expected utility—it is merely highly likely.[6] Thus, the argument must be that expected-utility maximization is overwhelmingly likely to be the overall best policy. And, if something is overwhelmingly likely to be the overall best policy, then one should do that. So, one should maximize expected utility.

The long-run argument only works under certain further assumptions about what sorts of gambles will arise in the long run. For example, in Pascal's Hell, *not* maximizing expected utility is overwhelmingly likely to be the overall best policy. So, by the same argument, one should *not* maximize expected utility in this case. Thus, the principle from which the long-run argument gets its intuitive support recommends against expected-utility maximization in some cases. And, the true principles of rationality (if there are any) should apply even in hypothetical cases such as Pascal's Hell. Expected-utility maximization might be overwhelmingly likely to be the overall best policy for us. But it is not so always and for everyone. If one accepts the principle that one should choose whatever policy is overwhelmingly likely to be best overall (either for oneself or the group of all agents), then, under some circumstances, one should not maximize expected utility. So, some other argument is needed to establish that one should always do so.[7]

[5] Briggs (2019).

[6] The *Strong Law of Large Numbers* implies that, for any arbitrarily small real umber $\epsilon > 0$, the probability that the average payoff of a prospect falls within ϵ of its expected utility converges to 1 as the number of trials increases. In other words, as the sample size goes to infinity, the average gain per trial will become arbitrarily close to the prospect's expected utility with probability 1. So, in the long run, the average utility associated with a prospect is virtually certain to equal its expected utility. See Briggs (2019, §2.1).

[7] See Briggs (2019, §2.1) for more discussion of the long-run argument for expected-utility

2.2 Representation theorems

Another argument for maximizing expected utility relies on representation theorems, such as the von Neumann-Morgenstern axiomatization of Expected Utility Theory. This representation theorem shows that the following axioms together entail Expected Utility Theory: Completeness, Transitivity, Independence and Continuity.[8] Let $X \succ Y$ mean that X is strictly preferred (or simply 'preferred') to Y.[9] Also, let XpY be a risky prospect with a p chance of prospect X obtaining and a $1-p$ chance of prospect Y obtaining. These axioms then state the following:

Completeness: $X \succsim Y$ or $Y \succsim X$.

Transitivity: If $X \succsim Y \succsim Z$, then $X \succsim Z$.

Independence: If $X \succ Y$, then $XpZ \succ YpZ$ for all probabilities $p \in (0, 1]$.

Continuity: If $X \succ Y \succ Z$, then there are probabilities p and $q \in (0, 1)$ such that $XpZ \succ Y \succ XqZ$.

Agents who conform to said axioms can be represented as maximizing expected utility. The argument for expected-utility maximization from representation theorems states that these axioms are the axioms of rational preference.[10] Thus, rational agents can be represented as maximizing expected utility. But why

maximization.

[8] See von Neumann and Morgenstern (1947), Jensen (1967, pp. 172–182) and Hammond (1998, pp. 152–164).

[9] Some prospect X is strictly preferred to another prospect Y when X is weakly preferred to Y, but Y is not weakly preferred to X.

[10] Briggs (2019) and Zynda (2000).

should one think these are the axioms of rational preference? First, one might consider them intuitively plausible, so it might seem intuitively right that rational agents satisfy these axioms. Alternatively, they can be supported by money-pump arguments. A money-pump argument intends to show that agents who violate some alleged requirement of rationality are vulnerable to making a combination of choices that leads to a sure loss. There are money-pump arguments for Completeness, Transitivity, Independence and Continuity.[11] So, if vulnerability to this sort of exploitation is a sign of irrationality, then one ought to satisfy the axioms that together entail Expected Utility Theory. Thus, rational agents maximize expected utility.[12]

2.3 Tiny probabilities of vast utilities

However, maximizing expected utility seems to lead to counterintuitive choices in cases that involve tiny probabilities of huge payoffs (at least if utilities are unbounded or if the upper bound is very high or the lower bound is very low). One such case was presented earlier. It is based on the following case:[13]

> **Pascal's Mugging:** A stranger approaches you and promises to use magic that will give you a thousand quadrillion happy days in the Seventh Dimension if you pay him a small amount of money.

Should you pay the stranger? There is a very small but non-zero probability that

[11] See Gustafsson (forthcoming).

[12] As discussed later, these axioms imply a bounded utility function.

[13] Bostrom (2009). This case is based on informal discussions by various people, including Eliezer Yudkowsky (2007*b*).

the stranger is telling the truth. And if he is telling the truth, then the payoff is enormous. Provided the payoff is sufficiently great, the expected utility of paying the stranger is greater than that of keeping the money. Also, if you have a non-zero credence in him being able and willing to deliver any finite amount of utility, then he can always increase the payoff until the offer has positive expected utility, at least if your utilities are unbounded.[14] So, someone who maximizes expected utility with an unbounded utility function (or with a high enough upper bound) would pay the stranger—which seems counterintuitive.

Another case that involves tiny probabilities of huge payoffs is the St. Petersburg paradox introduced by Nicolaus Bernoulli:[15]

> **St. Petersburg Game:** A fair coin is flipped until it lands on heads.[16] The prize is then $\$2^n$, where n is the number of coin flips.

While Pascal's Mugging involves large finite payoffs and a finite number of possi-

[14] Contrary to this, Baumann (2009, p. 447) argues that the larger the payoff the mugger promises to deliver, the lower the probability you should assign to the proposition that he will stick with his promise. Moreover, Baumann (2009, p. 447) argues that your probabilities should go down faster than the stranger's offer's utilities go up. Relatedly, Robin Hanson has suggested that in a scenario in which many individuals exist, they cannot all have total control over each other's existence. So, your credence in being able to influence them all should be penalized in proportion to the number of individuals that exist. Thus, credences in the mugger telling the truth should decrease in proportion to the possible payoff. See Hanson (2007) and Yudkowsky (2007*a*). But, in the version of Pascal's Mugging presented in this chapter, the stranger promises to prolong *your* life rather than also help very many orphans (as in Bostrom's version). And, it is less surprising to be in a special position to have so much control over one's future self.

[15] Nicolaus Bernoulli originally proposed a version of this game in 1713. The game was simplified by Gabriel Cramer in 1728 and published by Daniel Bernoulli in 1738. See Pulskamp (2013) and Bernoulli (1954). There are variants of the St. Petersburg game that do not seem to make any sense by the lights of Expected Utility Theory because they have no unique expected utility. See, for example, Nover and Hájek (2004) on the Pasadena game.

[16] What happens if the coin never lands on heads? We may suppose that, in that case, the player wins nothing. As this is a zero-probability event, it does not affect the expected utility of the game. See Chapter 2 of this thesis on Expected Utility Theory and possible states of zero probability.

ble outcomes, the St. Petersburg game involves arbitrarily large finite payoffs and infinitely many possible outcomes. The St. Petersburg game has infinite expected monetary value, so an agent who maximizes expected monetary value would pay any finite amount to play it. But again, this seems counterintuitive. As Nicolaus Bernoulli (agreeing with his friend Gabriel Cramer) writes: "[T]here is no person of good sense who wished to give merely 20 coins."[17] Daniel Bernoulli (cousin of Nicolaus Bernoulli) argues that the expected utility of the game is finite because of the diminishing marginal utility of money.[18] However, one can change the game slightly to bypass this objection by changing the prize from money to something with no diminishing marginal utility, such as (possibly) days of life.[19]

To summarize, Expected Utility Theory states that rational agents maximize expected utility. Expected Utility Theory can be supported with the long-run argument, on which one should maximize expected utility because it is overwhelmingly likely to be the best policy in the long run. Alternatively, Expected Utility Theory can be supported by representation theorems. This argument states that the axioms of Expected Utility Theory are the axioms of rational preference. However, maximizing expected utility seems to lead to counterintuitive choices in cases that involve tiny probabilities of huge payoffs, such as Pascal's Hell, Pascal's Mugging and the St. Petersburg paradox. Expected-utility maximization gives counterin-

[17] Pulskamp (2013, p. 6).

[18] Bernoulli (1954).

[19] Monton (2019, p. 2). This is related to the *Super St-Petersburg Paradox* which Samuelson (1977, p. 32) attributes to Menger (1934) (see Menger [1967] for an English translation). Menger (1967, pp. 217–218) shows that if utilities are unbounded, one can always create a Super St-Petersburg game, in which the payoffs grow sufficiently fast so that the expected utility of the game is infinite.

tuitive recommendations in such cases if utilities are unbounded or if the upper bound is very high or the lower bound very low. The next section discusses an implication of expected-utility maximization with an unbounded utility function; the subsequent section explores expected-utility maximization with a bounded utility function. The idea that tiny probabilities should be ignored in practical decision-making is investigated in §5. Finally, §6 briefly discusses some other approaches.

3 Probability Fanaticism

This section discusses arguments for and against Probability Fanaticism, namely, the idea that we should let tiny probabilities of vast utilities dominate the expected utility calculations. As we will see, there are strong arguments for and against Probability Fanaticism, as some plausible principles support this idea while others undermine it.

3.1 The Continuum Argument for Probability Fanaticism

There seems to be something wrong with a theory that lets tiny probabilities of huge payoffs dictate one's course of action. It might even seem *fanatical*. Thus, we may call this view *Probability Fanaticism*. Probability Fanaticism is the idea that tiny probabilities of huge positive or negative payoffs can have enormous positive or negative expected utility (respectively). Formally, it states the following:[20]

Probability Fanaticism:

[20] Wilkinson (2022, p. 449). Beckstead and Thomas (2020) call this 'Recklessness'.

i *Positive Probability Fanaticism* For any probability $p > 0$, and for any finite utility u, there is some large enough utility U such that probability p of U (and otherwise nothing) is better than certainty of u.[21]

ii *Negative Probability Fanaticism* For any probability $p > 0$, and for any finite negative utility $-u$, there is some large enough negative utility $-U$ such that probability p of $-U$ (and otherwise nothing) is worse than certainty of $-u$.

Probability Fanaticism is supported by a *Continuum Argument*.[22] Consider for example the following case:[23]

> **Devil at Your Deathbed:** You have one year of life left. But the devil appears and offers you ten years of happy life instead, with probability 0.999. You accept the offer. But the devil then offers you 100 years of happy life instead, with probability 0.999^2—just 0.1% lower. After some 50,000 trades, you find yourself with a $0.999^{50,000}$ probability of $10^{50,000}$ years of happy life. Predictably, you die shortly thereafter.

In this case, each deal seems better than the one before. Accepting each deal massively increases the payoff while decreasing its probability by a tiny percentage. However, accepting all trades means trading a certain good payoff (one year of happy life) for an extremely tiny probability of a great payoff.

[21] In this context, 'otherwise nothing' means retaining the status quo or baseline outcome.
[22] This argument is from Beckstead (2013, §6) and Beckstead and Thomas (2020, §1).
[23] Beckstead and Thomas (2020, pp. 4–5)

If p is a probability and n is a number of happy lives, then let $p \cdot n$ be a prospect that gives probability p of n happy lives (and otherwise nothing). Then, the following principle supports accepting all the trades:[24]

> **Anti-Timidity:** For any probabilities $p \gg q$ and numbers of happy lives $N \gg n$, $p \cdot (n + N) \succ (p + q) \cdot n$.

Anti-Timidity says that one can always compensate for a tiny decrease in the probability of a good outcome by increasing the payoff sufficiently. Anti-Timidity is plausible. However, it implies (Positive) Probability Fanaticism; repeated applications of Anti-Timidity (together with transitivity) tell us that a tiny probability of a great payoff is better than certainty of a good payoff.[25] Whichever payoff one starts with, and for any tiny probability $p > 0$, there is some great enough payoff such that probability p of the great payoff (and otherwise nothing) is better than certainty of the original payoff. So, to deny Probability Fanaticism, one must reject Anti-Timidity or transitivity—yet both seem intuitively compelling.

[24] Russell (2021, p. 7) and Beckstead and Thomas (2020, p. 2).

[25] A similar argument can be given to support Negative Probability Fanaticism. Instead of Anti-Timidity, this argument uses the following principle:

> **Negative Anti-Timidity:** For any probabilities $p \gg q$ and numbers of unhappy lives $N \gg n$, $(p + q) \cdot n \succ p \cdot (n + N)$.

3.2 More is Better and Simple Separability imply Probability Fanaticism

Another argument for Probability Fanaticism is that it follows from two plausible principles, namely, More is Better and Simple Separability.[26] More is Better states the following:[27]

> **More is Better:** For probabilities $p \gg q$ and numbers $N \gg n$, $p \cdot N \succ q \cdot n$.

More is Better states that it is better to have a much higher probability of many more happy lives than a smaller probability of fewer happy lives.

Let X be a prospect that concerns what is going on in the part of the world we might make any difference to, and let Y be a prospect that concerns what happens somewhere far away, such as a distant galaxy. Also, let $X \oplus Y$ be the combined prospect of the near prospect X and the far prospect Y. Finally, let a 'simple prospect' be a prospect that has only a finite number of possible outcomes. Then, Simple Separability states the following:[28]

> **Simple Separability:** For all simple near prospects X and Y, and any simple far prospect Z, $X \succ Y$ if and only if $X \oplus Z \succ Y \oplus Z$.

Denying Simple Separability means that uncertainty over what happens in distant places can be relevant to what we ought to do, even when we cannot affect what

[26] This argument is also from Beckstead and Thomas (2020, §3.2). The presentation follows closely Russell (2021, §2).

[27] Russell (2021, p. 6).

[28] Russell (2021, p. 15).

happens in those distant places.

To see how More is Better and Simple Separability imply Probability Fanaticism, consider the following prospects:

> **More vs. Less:** Let $p \gg q$ and $N \gg n$. Also, let the probabilities of states 1, 2 and 3 be p, q and $1 - p - q$ (respectively).
>
> *More* Gives N happy lives in state 1 and nothing in states 2 and 3.
>
> *Less* Gives n happy lives in state 2 and nothing in states 1 and 3.

Suppose you face a choice between More and Less, while the inhabitants of a distant Earth-like planet face the following prospect:

> *Far* Gives n happy lives in state 1 and nothing in states 2 and 3.

Given that the Earth-like planet faces prospect Far, the choice you face is between More $\oplus$ Far and Less $\oplus$ Far (see table 1). And, given that More is better than Less (by More is Better), it follows by Simple Separability that More $\oplus$ Far is better than Less $\oplus$ Far. However, as seen in table 1, More $\oplus$ Far gives a slightly lower probability p of a much large number of happy people $n + N$. Thus, More is Better and Simple Separability imply Anti-Timidity: A slightly smaller probability of a much large number of happy lives is better than a slightly higher probability of many fewer happy lives. And, as we saw in the previous section, Anti-Timidity (together with transitivity) implies Probability Fanaticism. Therefore, More is Better and Simple Separability (together with transitivity) imply Probability Fanaticism. To deny Probability Fanaticism, one must reject More is Better, Simple Separability or transitivity—yet they all seem intuitively compelling.

TABLE 1

MORE ⊕ FAR VS. LESS ⊕ FAR

	State 1	**State 2**	**State 3**
Probability	p	q	$1-p-q$
More ⊕ *Far*	$n+N$	0	0
Less ⊕ *Far*	n	n	0

3.3 Stochastic Dominance and Simple Separability imply Probability Fanaticism

Another related argument for Probability Fanaticism is that it follows from Simple Separability and another very compelling principle, namely, Stochastic Dominance.[29] Let $X = \{x_1,\ p_1;\ x_2,\ p_2\ ;\dots\}$ stand for prospect X that gives non-zero probabilities p_1, p_2, and so on, of outcomes x_1, x_2, and so on. Stochastic Dominance then states the following:[30]

> **Stochastic Dominance:** For all prospects $X = \{x_1,\ p_1;\ x_2,\ p_2\ ;\dots\}$ and $Y = \{y_1,\ q_1;\ y_2,\ q_2\ ;\dots\}$, X is at least as good as Y if, for all out-

[29] The presentation follows closely Russell (2021, pp. 30–33). See Wilkinson (2022, §VI A) for a very similar argument. Also see Tarsney (2020), Beckstead and Thomas (2020) and Goodsell (2021).

[30] Buchak (2013, p. 42). More precisely, this is *first-order stochastic dominance*, an idea that was introduced to statistics by Mann and Whitney (1947) and Lehmann (1955), and to economics by Quirk and Saposnik (1962). The name 'first-degree stochastic dominance' is due to Hadar and Russell (1969, p. 27).

comes o,

$$\sum_{\{i \mid x_i \gtrsim o\}} p_i \geq \sum_{\{j \mid y_j \gtrsim o\}} q_j.$$

If in addition, for some outcome u,

$$\sum_{\{i \mid x_i \gtrsim u\}} p_i > \sum_{\{j \mid y_j \gtrsim u\}} q_j,$$

then X is better than Y.

One violates Stochastic Dominance if, for all outcomes, some prospect X gives an at least as high probability of an at least as great outcome as some other prospect Y does, but X is not judged at least as good as Y. One also violates Stochastic Dominance if, in addition, X gives a greater probability of an at least as great outcome as Y does for some outcome—yet X is not judged better than Y.

To see how Simple Separability and Stochastic Dominance imply Probability Fanaticism, consider the following prospects:

Safe vs. Risky:

Safe Certainly gives one happy life.

Risky Gives probability $p > 0$ of $n+1$ happy lives (a great outcome) and otherwise nothing.

Suppose p is tiny. Then, the comparison between Risky and Safe can be considered at a more abstract level whereby it simply corresponds to Positive Probability

Fanaticism. So, Probability Fanaticism is true if Risky is better than Safe. Also, suppose you face the choice between Safe and Risky, while the inhabitants of a distant Earth-like planet face the following prospect (see table 2):

> *Twin Earth* Gives p chance of nothing, q chance of one happy life, q chance of two happy lives, q chance of three happy lives, … , q chance of n happy lives, where $q < p$.

Table 2
Twin Earth

Probability	p	q	q	q	…	q
Safe	1	1	1	1	…	1
Risky	$n+1$	0	0	0	…	0
Twin Earth	0	1	2	3	…	n

When you take into account the prospect Twin Earth is facing, your options are as follows (see table 3):

> **Mixed Prospects:**
>
> *Safe* $\oplus$ *Twin Earth* Gives p chance of one happy life, q chance of two happy lives, q chance of three happy lives, q chance of four happy lives, … , q chance of $n+1$ happy lives.
>
> *Risky* $\oplus$ *Twin Earth* Gives p chance of $n+1$ happy lives, q chance of one happy life, q chance of two happy lives, q chance of three happy lives, … , q chance of n happy lives.

Table 3

Mixed Prospects

Probability	p	q	q	…	q
Safe $\oplus$ *Twin Earth*	1	2	3	…	$n+1$
Risky $\oplus$ *Twin Earth*	$n+1$	1	2	…	n

Next, given that p is greater than q, we may split the first column of table 3 into two columns that give probabilities $p-q$ and q (respectively), as shown in the following table:

Table 4

Mixed Prospects: Split

Probability	$p-q$	q	q	q	…	q
Safe $\oplus$ *Twin Earth*	1	1	2	3	…	$n+1$
Risky $\oplus$ *Twin Earth*	$n+1$	$n+1$	1	2	…	n

Next, we may reorder the outcomes of Risky $\oplus$ Twin Earth that are associated with probability q by moving each of them to the column on their left (see table 5). The leftmost outcome associated with probability q (i.e., $n+1$) is moved to the rightmost column (where n is in table 4).

Table 5

Mixed Prospects: Reorder

Probability	$p-q$	q	q	q	…	q
Safe $\oplus$ *Twin Earth*	1	1	2	3	…	$n+1$
Risky $\oplus$ *Twin Earth*	$n+1$	1	2	3	…	$n+1$

It is now evident from table 5 that the only difference between Safe $\oplus$ Twin Earth and Risky $\oplus$ Twin Earth is that the former gives probability $p - q$ of one happy life, while the latter gives the same probability of $n + 1$ happy lives. As it is better to obtain $n + 1$ happy lives than just one happy life, Risky $\oplus$ Twin Earth stochastically dominates Safe $\oplus$ Twin Earth: For all outcomes, it gives an at least as high probability of an at least as great outcome as Safe $\oplus$ Twin Earth does, and for one outcome, Risky $\oplus$ Twin Earth gives a greater probability of an at least as great outcome as Safe $\oplus$ Twin Earth does. So, by Stochastic Dominance, Risky $\oplus$ Twin Earth is better than Safe $\oplus$ Twin Earth.

Finally, given that Risky $\oplus$ Twin Earth is better than Safe $\oplus$ Twin Earth, it follows by Simple Separability that Risky is better than Safe: Probability Fanaticism is true. So, Probability Fanaticism follows from Simple Separability and Stochastic Dominance.[31] If one wishes to avoid Probability Fanaticism, one must reject Simple Separability or Stochastic Dominance, which are both intuitively compelling.

[31] This argument assumes that the number and not the location of happy lives is all that matters. More generally, Probability Fanaticism follows from Stochastic Dominance, Simple Separability and the following principle:

> **Positive Compensation:** For any near good x and far good y, there is a far good z such that $x \oplus y \sim 0 \oplus z$, and there is a near good w such that $x \oplus y \sim w \oplus 0$.

According to this principle, we can always compensate for making things worse nearby by making things sufficiently better far away (and vice versa). See Russell (2021) for the full argument.

3.4 Stochastic Dominance and Separability are jointly inconsistent

Above we saw how Probability Fanaticism follows from Simple Separability and Stochastic Dominance. However, Stochastic Dominance and a generalization of Simple Separability are jointly inconsistent.[32] This undermines the argument for Probability Fanaticism from Stochastic Dominance and Simple Separability.

Unlike Simple Separability, the generalization of Simple Separability applies to prospects that have an infinite number of possible outcomes. It states the following:[33]

Separability:

i For all near prospects X and Y, and any far prospect Z, $X \succ Y$ if and only if $X \oplus Z \succ Y \oplus Z$.

ii For all far prospects X and Y, and any near prospect Z, $X \succ Y$ if and only if $Z \oplus X \succ Z \oplus Y$.

To see why Stochastic Dominance and Separability are jointly inconsistent, consider the following versions of St. Petersburg games (see table 6):

St. Petersburg Games: A fair coin is flipped until it comes up heads.

St. Petersburg Gives 2^n happy lives, where n is the number of coin flips (and otherwise it gives nothing).

[32] This argument is from Russell (2021).
[33] Russell (2021, p. 15).

St. Petersburg$^-$ Gives $2^n - 1$ happy lives, where n is the number of coin flips (and otherwise it gives nothing).

St. Petersburg$^-$ gives the same probabilities as the St. Petersburg game but slightly worse outcomes. It seems clear that St. Petersburg is better than St. Petersburg$^-$; indeed, this is what Stochastic (and Statewise) Dominance tells us.[34]

TABLE 6
ST. PETERSBURG GAMES

No. of flips	1	2	3	…
Probability	1/2	1/4	1/8	…
St. Petersburg	2	4	8	…
St. Petersburg$^-$	1	3	7	…

Separability then tells us that two copies of St. Petersburg, one here and the other in a distant galaxy, are better than two copies of St. Petersburg$^-$, one here and the other in a distant galaxy: As St. Petersburg is better than St. Petersburg$^-$, by Separability, St. Petersburg $\oplus$ St. Petersburg is better than St. Petersburg$^-$ $\oplus$ St. Petersburg. Again, because St. Petersburg is better than St. Petersburg$^-$, St. Petersburg$^-$ $\oplus$ St. Petersburg is better than St. Petersburg$^-$ $\oplus$ St. Petersburg$^-$. Thus, by transitivity, St. Petersburg $\oplus$ St. Petersburg is better than St. Petersburg$^-$ $\oplus$ St. Petersburg$^-$.

However, we can arrange the mechanisms of these games so that St. Petersburg$^-$ $\oplus$ St. Petersburg$^-$ stochastically dominates St. Petersburg $\oplus$ St. Petersburg. In this

[34] According to Statewise Dominance, a prospect is better than another prospect if it gives a better outcome in all states.

case, the results of the two St. Petersburg games depend on the outcome of flipping a dime. And, the results of the two St. Petersburg^{-} games depend on the outcomes of flipping the dime and a penny (see table 7):

> **Correlated St. Petersburg Games:** A dime is flipped until it comes up heads, and a penny is flipped once.
>
> *Near and Far St. Petersburg* Give 2^n happy lives, where n is the number of coin flips with the dime.
>
> *Near St. Petersburg*$^{-}$ Gives one happy life if the penny comes up heads. Otherwise, it gives twice as much as St. Petersburg minus one.
>
> *Far St. Petersburg*$^{-}$ Gives one happy life if the penny comes up tails. Otherwise, it gives twice as much as St. Petersburg minus one.

Table 7
Correlated St. Petersburg Games

Outcome	$H, 1$	$H, 2$	$H, 3$	…	$T, 1$	$T, 2$	$T, 3$	…
Probability	$1/4$	$1/8$	$1/16$	…	$1/4$	$1/8$	$1/16$	…
Near St. Petersburg	2	4	8	…	2	4	8	…
Far St. Petersburg	2	4	8	…	2	4	8	…
Near St. Petersburg^{-}	1	1	1	…	3	7	15	…
Far St. Petersburg^{-}	3	7	15	…	1	1	1	…

'*H*' and '*T*' indicate the outcome of flipping the penny, and
'1', '2', … indicate the number of coin flips with the dime.

Note that both the Near and the Far St. Petersburg games give the same probabilities of the same outcomes as the St. Petersburg game in table 6. Similarly, both

the Near and the Far St. Petersburg$^-$ games give the same probabilities of the same outcomes as the St. Petersburg$^-$ game in table 6. Thus, Near St. Petersburg $\oplus$ Far St. Petersburg should be better than Near St. Petersburg$^-$ $\oplus$ Far St. Petersburg$^-$.

However, as seen in table 8, Near St. Petersburg $\oplus$ Far St. Petersburg gives the same probabilities of the same outcomes as Near St. Petersburg$^-$ $\oplus$ Far St. Petersburg$^-$. In every state, they result in the same number of happy lives. So, Near St. Petersburg $\oplus$ Far St. Petersburg is stochastically equivalent to Near St. Petersburg$^-$ $\oplus$ Far St. Petersburg$^-$. By Stochastic Dominance, each prospect is at least as good as the other. Therefore, they are equally good.

Table 8
Mixed St. Petersburg Games

Outcome	$H, 1$	$H, 2$	$H, 3$	…	$T, 1$	$T, 2$	$T, 3$	…
Probability	$1/4$	$1/8$	$1/16$	…	$1/4$	$1/8$	$1/16$	…
St. Petersburg $\oplus$ *St. Petersburg*	4	8	16	…	4	8	16	…
St. Petersburg$^-$ $\oplus$ *St. Petersburg*$^-$	4	8	16	…	4	8	16	…

'H' and 'T' indicate the outcome of flipping the penny, and
'1', '2', … indicate the number of coin flips with the dime.
'Near' and 'Far' have been omitted from the prospects' names.

Here is a recap of the argument: Stochastic Dominance tells us that St. Petersburg is better than St. Petersburg$^-$. Separability then tells us that Near St. Petersburg $\oplus$ Far St. Petersburg is better than Near St. Petersburg$^-$ $\oplus$ Far St. Petersburg$^-$. However, Near St. Petersburg $\oplus$ Far St. Petersburg is stochastically equivalent to Near St. Petersburg$^-$ $\oplus$ Far St. Petersburg$^-$. So, they must be equally good. But Near St. Petersburg $\oplus$ Far St. Petersburg cannot both be better than and equally as

good as Near St. Petersburg$^-$ $\oplus$ Far St. Petersburg$^-$. Thus, Stochastic Dominance and Separability are jointly inconsistent. Either St. Petersburg is not better than St. Petersburg$^-$, even though it stochastically dominates it. Alternatively, Near St. Petersburg $\oplus$ Far St. Petersburg is not better than Near St. Petersburg$^-$ $\oplus$ Far St. Petersburg$^-$, even though Separability tells us so.

To summarize, Stochastic Dominance and Simple Separability imply Probability Fanaticism. However, Stochastic Dominance and a generalization of Simple Separability are jointly inconsistent.[35] Moreover, whatever the justification of Simple Separability is should also apply to Separability. As Russell (2021, pp. 14–15) writes: "What would the motivation be for it [Simple Separability] that is not also motivation for the unrestricted principle [Separability]? It can't be simply the idea that if what is going on in distant space and time is the same for both of two options, then it is irrelevant to which is better. That idea supports full-fledged Separability. So is there something special about simple prospects that makes their value insensitive to what is going on in distant space and time?" Unless there is some unique justification for Simple Separability that does not also apply to the generalized version, one has no reason to accept Simple Separability if one rejects Separability. And, Stochastic Dominance tells us that Separability is wrong. Thus, for a lack of a unique justification for Simple Separability, the argument for Probability Fanaticism from Stochastic Dominance and Simple Separability does not go through.

[35] As before, this argument assumes that the number, and not the location, of happy lives is all that matters. More generally, Stochastic Dominance, Separability and Positive Compensation are jointly inconsistent. See Russell (2021) for the full argument.

3.5 Stochastic Dominance, Negative Reflection and Background Independence imply Probability Fanaticism

We have seen that Simple Separability and Stochastic Dominance imply Probability Fanaticism, but Separability and Stochastic Dominance are jointly inconsistent. This section shows that Stochastic Dominance, together with two other plausible principles, implies Probability Fanaticism.[36]

Consider the following prospects:

Safe* vs. Risky*:

Safe* Certainly gives a good outcome (utility v).

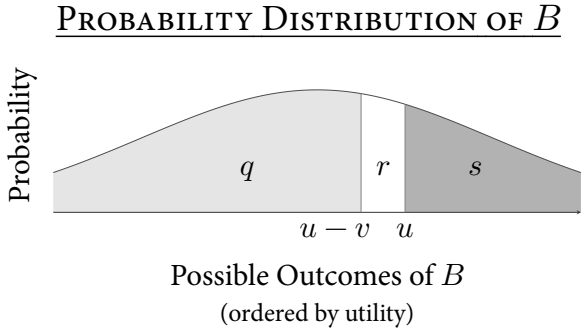
Risky* Gives a tiny probability $p > 0$ of a great outcome (utility V).

It can be shown that for some background prospect B, which is probabilistically independent of both Safe* and Risky*, Risky$^\star \oplus B$ stochastically dominates Safe$^\star \oplus B$.[37] For this to happen, we need Risky$^\star \oplus B$ to have at least as high a probability as Safe$^\star \oplus B$ of resulting in at least utility u, for all possible utilities u. Choose any utility $u < V$. Safe* certainly gives utility v, so the probability that Safe$^\star \oplus B$ gives at least utility u is the probability that B gives at least utility $u - v$ (area $r + s$ in

[36] This argument is from Wilkinson (2022), and the presentation follows closely Russell (2021).

[37] Wilkinson (2022, §VI). See also Tarsney (2020). Tarsney (2020) explores the idea that Stochastic Dominance is a sufficient principle of rationality. Prospects that give a higher expected utility but would not otherwise stochastically dominate their alternatives can become stochastically dominant given sufficient background uncertainty. However, background uncertainty generates stochastic dominance much less readily when the prospect involves tiny probabilities of huge payoffs. So, Tarsney (2020) argues that Stochastic Dominance as a sufficient principle of rationality can vindicate the intuition that we are often permitted to decline gambles like Pascal's Mugging or the St. Petersburg game. However, as the following argument shows, Stochastic Dominance sometimes demands that we make fanatical choices.

the graph below). Risky$^\star \oplus B$, in turn, gives at least utility u if either Risky* gives a great outcome (utility $V > u$) or if B gives at least utility u. Denote the probability that B gives at least utility u by s (see area s in the graph below). So, the probability that Risky$^\star \oplus B$ gives at least u is $p + (1-p)s$.



It can be shown that this (the probability that Risky$^\star \oplus B$ gives at least u) will be greater than the probability that Safe$^\star \oplus B$ gives at least u if the area r is less than or equal to the area q multiplied by p.[38] So, if the area r is small enough compared to area q, then Risky$^\star \oplus B$ gives an at least as high probability of u as Safe$^\star \oplus B$ does. For this to happen with all u, the interval between u and $u - v$ needs to be tiny enough. And for that to happen, the probabilities in B must go down slowly

[38]
$$
\begin{aligned}
& p + s(1-p) \geq r + s \\
\iff\ & s + p(1-s) \geq r + s \\
\iff\ & p(q+r) \geq r \\
\iff\ & r \leq pq + pr \\
\iff\ & r(1-p) \leq pq \\
\iff\ & r \leq \frac{pq}{1-p}, \text{ for which } r \leq pq \text{ is a sufficient condition.}
\end{aligned}
$$

enough as we approach $-\infty$ and rise and fall quickly enough as we pass the peak of the curve. There are some probability distributions with this property.[39] So, with some background prospect B, Risky$^\star \oplus B$ is better than Safe$^\star \oplus B$ by Stochastic Dominance.

Next, consider the following principle:[40]

> **Negative Reflection:** For prospects X and Y and a question Q, if X is not better than Y conditional on any possible answer to Q, then X is not better than Y unconditionally.

Given that Risky$^\star \oplus B$ is better than Safe$^\star \oplus B$ (by Stochastic Dominance), Negative Reflection tells that B must have some possible outcome b such that Risky$^\star \oplus b$ is better than Safe$^\star \oplus b$.

Finally, consider the following principle:[41]

> **Background Independence:** For any near prospects X and Y and any far outcome a, $X \succ Y$ if and only if $X \oplus a \succ Y \oplus a$.

[39] Wilkinson (2022, §VI) and Tarsney (2020).

[40] Russell (2021, p. 19). Compare *Negative Reflection* to the following related principle (Russell, 2021, p. 23):

> **Positive Reflection:** For prospects X and Y and a question Q, if X is at least as good as Y conditional on any possible answer to Q, then X is at least as good as Y unconditionally.

Both reflection principles are related to the Sure Thing Principle due to Savage (1972, pp. 21–22):

> **The Sure Thing Principle:** For all prospects X and Y, if an agent would not prefer X over Y if they learnt that some event E has happened, or if they learnt that E has not happened, then the agent does not prefer X over Y. Moreover, if the agent would prefer Y to X if they learnt that E has happened, and they would not prefer X to Y if they learnt that E has not happened, then the agent prefers Y to X.

[41] Wilkinson (2022, p. 467) and Russell (2021, p. 28).

Background Independence is similar to Separability. But unlike Separability, it requires that the 'background prospect' involves no uncertainty.[42] Given that Risky$^\star \oplus b$ is better than Safe$^\star \oplus b$ (by Stochastic Dominance and Negative Reflection), Background Independence implies that Risky* is better than Safe*. This means that Probability Fanaticism is true.

To conclude, three plausible principles, Stochastic Dominance, Negative Reflection and Background Independence, imply Probability Fanaticism. Thus, to deny Probability Fanaticism, one must reject Stochastic Dominance, Negative Reflection or Background Independence.

3.6 Stochastic Dominance and Negative Reflection imply Probability Fanaticism is false

We just saw how Stochastic Dominance, Negative Reflection and Background Independence imply Probability Fanaticism. However, if Probability Fanaticism is true, then two of the premises of the previous argument—Stochastic Dominance and Negative Reflection—are jointly inconsistent.[43] Thus, the argument cannot be sound.

To see why Stochastic Dominance and Negative Reflection are jointly inconsistent if Probability Fanaticism is true, consider the following versions of the St. Petersburg game:

[42] Background Independence is related to the Egyptology objection to the Average View in population ethics. See McMahan (1981, p. 115) and Parfit (1984, p. 420).

[43] This argument is from Russell (2021, §3) and Russell and Isaacs (2021). Also see Chalmers (2002) and Beckstead and Thomas (2020, §4).

St. Petersburg Games: A fair coin is flipped until it comes up heads.

St. Petersburg Gives 2^n happy lives, where n is the number of coin flips.

St. Petersburg$^+$ Gives $2^n + 1$ happy lives, where n is the number of coin flips.

The outcomes of St. Petersburg$^+$ are better than the outcomes of St. Petersburg. So, by Stochastic Dominance, St. Petersburg$^+$ is better than St. Petersburg.

However, if Probability Fanaticism is true, then none of the *outcomes* of St. Petersburg$^+$ are as good as the *prospect* St. Petersburg. This is because St. Petersburg and St. Petersburg$^+$ are better than any possible finite payoffs. So, any possible payoff of St. Petersburg$^+$ is worse than the *prospect* St. Petersburg. Negative Reflection, therefore, implies that St. Petersburg$^+$ is not better than St. Petersburg. Conditional on any way St. Petersburg$^+$ could turn out, St. Petersburg$^+$ is not better than St. Petersburg, so St. Petersburg$^+$ cannot be better than St. Petersburg. However, St. Petersburg$^+$ is better than St. Petersburg by Stochastic Dominance. So, if Probability Fanaticism is true, either Stochastic Dominance or Negative Reflection needs to go. They are jointly inconsistent. Thus, Stochastic Dominance and Negative Reflection imply that Probability Fanaticism is false.[44]

Suppose that Probability Fanaticism keeps Stochastic Dominance (and gives up Negative Reflection). In that case, it is dynamically inconsistent and vulnera-

[44]Stochastic Dominance and Negative Reflection imply that Probability Fanaticism is false. They cannot, therefore, be used in an argument for Probability Fanaticism. However, a principle that is related to Negative Reflection, together with Stochastic Dominance and Background Independence, implies that Positive Fanaticism or Negative Fanaticism is true. See Russell (2021, pp. 37–38).

ble to money pumps.[45] Consider for example the following case: You start with St. Petersburg$^+$. Once the result of St. Petersburg$^+$ is known, you can pay \$100 to exchange the outcome St. Petersburg$^+$ for the prospect St. Petersburg. Because any possible finite payoff of St. Petersburg$^+$ is worse than the prospect St. Petersburg, you would accept this trade. But before finding out the result of St. Petersburg$^+$, you can pay \$50 to simply keep the prospect St. Petersburg$^+$ and receive no further offers. You know that if you do not pay this \$50, you will end up with prospect St. Petersburg and with \$100 less in your wallet. But if you do pay this \$50, then you will end up with prospect St. Petersburg$^+$, and you will only have paid \$50. Therefore, by Stochastic Dominance, you should pay \$50 to avoid any further offers. But you have then been money pumped, as you have paid for something that you could have kept for free had you refused all the offers.[46] So, Probability Fanaticism combined with Stochastic Dominance is vulnerable to money pumps.

To conclude, this section has discussed arguments for and against Probability Fanaticism. §3.1 showed that Anti-Timidity and transitivity imply Probability Fanaticism. Then, §3.2 showed that More is Better and Simple Separability imply Probability Fanaticism. §3.3 showed that Stochastic Dominance and Simple Separability imply Probability Fanaticism. However, §3.4 showed that Stochastic

[45]This argument is from Russell and Isaacs (2021, p. 4 n. 5). Russell and Isaacs (2021) show that Probability Fanaticism violates Countable Independence, which is similar to Negative Reflection.

[46]If Probability Fanaticism rejects both Stochastic Dominance and Negative Reflection, then St. Petersburg$^+$ is not better than St. Petersburg. However, you are still permitted to pay to keep St. Petersburg$^+$ and receive no further offers. In fact, you are permitted to pay any finite amount to receive no further offers; whichever finite sum you pay, you will face a prospect with infinite expected utility. Thus, Probability Fanaticism still permits you to make a series of trades that results in a sure loss.

Dominance and (generalized) Separability are jointly inconsistent. §3.5 showed that Stochastic Dominance, Negative Reflection and Background Independence imply Probability Fanaticism. But §3.6 showed that Stochastic Dominance and Negative Reflection imply that Probability Fanaticism is false. §3.6 also showed that Probability Fanaticism is vulnerable to exploitation by money pumps. The debate between proponents and opponents of Probability Fanaticism is inconclusive, as there are strong arguments for and against it. However, as Russell (2021, p. 5) writes, "Whatever the truth of the matter, the ethics of huge numbers is deeply weird and full of surprises."

4 Bounded utilities

The rest of the sections discuss alternatives to Probability Fanaticism. This section explores the idea that utilities are bounded above and below.

Boundedness of utilities has been discussed as a possible alternative to Probability Fanaticism.[47] If utilities are real valued, then Boundedness means the following:

> **Boundedness:** There is some $M \in \mathbb{R}$ such that for all outcomes x, $|u(x)| < M$.

In other words, Boundedness rules out arbitrarily and infinitely good outcomes.

The following discussion focuses on Boundedness in the context of Expected Utility Theory. As discussed in Chapter 1 of this thesis, standard axiomatizations

[47] See Beckstead and Thomas (2020, §2.1) and Chapter 1 of this thesis.

of expected-utility maximization require utilities to be bounded.[48] Bounded utilities are, therefore, the standard in decision theory. However, bounded utilities seem troubling from the point of view of ethics. It seems odd that, for example, additional happy lives matter less the more happy lives there already are or that additional headaches matter less the more headaches (or other negative experiences) there already are. Also, bounded utilities imply that it is better to save some (very large) number n of lives for sure than to save *any number* of lives with a probability of almost one.[49] This happens when the value of n happy lives is close to the upper bound of utilities as then additional happy lives do not contribute much to expected utility.

Boundedness gives ethically even more untenable prescriptions. Consider for example the following prospects (see table 9):

> **Happy Lives vs. Headaches:** A fair coin is flipped.
>
> *Prospect A* Gives some large number m of happy lives with heads, and one person gets a headache with tails.
>
> *Prospect B* Gives some much larger number M of happy lives with heads, and two people get headaches with tails.

Suppose that the values of m and M happy lives are close to the upper bound of utilities. In that case, the additional happy lives in B may not contribute enough to B's expected utility to outweigh the disvalue of the possible additional headache.

[48] See Kreps (1988, pp. 63–64), Fishburn (1970, pp. 194, 206–207), Hammond (1998, pp. 186–191) and Russell and Isaacs (2021).

[49] More generally, Boundedness violates Anti-Timidity. See Beckstead and Thomas (2020, §2.1).

Then, Boundedness (from above) implies that A is better than B—which seems wrong from an ethical point of view.[50]

TABLE 9
HAPPY LIVES VS. HEADACHES

	Heads	**Tails**
A	Many happy lives	One headache
B	Very many happy lives	Two headaches

Next, consider the following prospects (see table 10):

> **Unhappy Lives vs. Lollipops:** A fair coin is flipped.
>
> *Prospect* C Gives some large number m of unhappy lives with heads, and one person gets a lollipop with tails.
>
> *Prospect* D Gives a much larger number M of unhappy lives with heads, and two people get lollipops with tails.

For similar reasons as explained above, D may be better than C if utilities are bounded below.[51] The implications of Boundedness is ethically untenable; the possibility of one additional lollipop should not compensate for an equally likely chance of many additional unhappy lives.

[50] This argument is from Beckstead and Thomas (2020, §3.3).
[51] See Beckstead and Thomas (2020, §3.4).

Table 10

Unhappy Lives vs. Lollipops

	Heads	Tails
C	Many unhappy lives	One lollipop
D	Very many unhappy lives	Two lollipops

Boundedness also implies that sometimes one should choose a small probability of a mediocre payoff instead of a high probability of a great payoff—which violates More is Better. To see how this violation happens, consider the following prospects (see table 11):[52]

Great vs. Mediocre Past:

Great Gives some great payoff (such as very many happy lives) if humanity's past was great (high probability p); otherwise, nothing happens.

Mediocre Gives some mediocre payoff (such as a few happy lives) if humanity's past was mediocre (small probability $1 - p$); otherwise, nothing happens.

In this case, Boundedness implies that Mediocre might be better than Great. If humanity's past was great (in which case the value of the world is near the upper bound of utilities), then the great payoff does not contribute much to utility. However, if humanity's past was mediocre, then the mediocre payoff makes a large contribution to utility. Thus, Boundedness implies that one should choose a small

[52] This argument is from Beckstead and Thomas (2020, §3.5).

probability of a mediocre payoff (and otherwise nothing) instead of a high probability of a great payoff (and otherwise nothing). This is a violation of More is Better.

Table 11
Great vs. Mediocre

	Great past	**Mediocre past**
Probability	p	$1-p<p$
Great	Many happy lives	Nothing
Mediocre	Nothing	A few happy lives

We have seen that Boundedness has ethically worrying implications. Chapter 1 of this thesis shows another troubling feature of Boundedness. It shows that decision theories on which utilities are bounded, such as Expected Utility Theory, violate Ex Ante Pareto if combined with an additive axiology, such as Total Utilitarianism. According to Total Utilitarianism, a population is better than another just in case the total quantity of well-being it contains is greater. Ex Ante Pareto, in turn, states the following:

> **Ex Ante Pareto:** For all prospects X and Y, if X is at least as good as Y for everyone, and X is better than Y for some, then X is better than Y.

The combination of Expected Utility Theory and Total Utilitarianism violates Ex Ante Pareto because the total quantity of well-being might be infinite or arbitrarily large. Thus, there must be a non-linear transformation from the total quantity of well-being into utilities used in decision-making. This non-linear transformation

is required if one has a non-zero credence in the possibility that an infinite or arbitrarily large number of individuals exist. But it is also required if one wishes to avoid Probability Fanaticism. However, such a transformation leads to violations of Ex Ante Pareto. So, the reconciliation of Expected Utility Theory and Total Utilitarianism prescribes prospects that are better for none and worse for some. Chapter 1 also discusses how this relates to a well-known result in this area, namely, Harsanyi's social aggregation theorem.

Chapter 2 of this thesis is somewhat related to the discussion of Boundedness. It points out that standard axiomatizations of Expected Utility Theory violate Statewise Dominance with prospects that involve possible states of zero probability. Statewise Dominance says the following:

> **Statewise Dominance:** If the outcome of prospect X is at least as good as the outcome of prospect Y in all states, and the outcome of X is better than the outcome of Y in some possible state, then X is better than Y.

At least at first glance, Expected Utility Theory tells us to be indifferent between two prospects when they are otherwise the same, except that one gives a better outcome than the other in a possible state of zero probability. But as some have suggested, Expected Utility Theory might be supplemented with dominance reasoning to get the verdict that the dominating prospect is better than the dominated one. However, Chapter 2 shows that if Expected Utility Theory is supplemented with dominance reasoning in this way, it will violate the Continuity axiom of Expected Utility

Theory. So, if an expected-utility maximizer wishes to retain Statewise Dominance even in cases that involve possible states of zero probability, they must adopt some axiomatization of Expected Utility Theory that does not have Continuity as one of the axioms.

To conclude, bounded utilities have been proposed as an alternative to Probability Fanaticism. Boundedness follows from standard axiomatizations of Expected Utility Theory, so it is the orthodox view in decision theory. However, Boundedness is troubling from an ethical point of view. For example, if utilities are bounded, it is better to save some (very large) number n of lives for sure than to save *any number* of lives with a probability of almost one. Also, it sometimes implies that the possibility of a very large number of additional happy lives cannot compensate for the disvalue of an equally likely additional headache. Similarly, it sometimes implies that the possibility of a single additional lollipop can compensate for the disvalue of an equally likely possibility of a very large number of unhappy lives. Also, Boundedness sometimes implies that one should choose a small probability of a mediocre outcome over a high probability of a great outcome. Furthermore, Chapter 1 of this thesis shows that decision theories on which utilities are bounded violate Ex Ante Pareto if combined with an additive axiology. Also, as shown in Chapter 2, standard axiomatizations of Expected Utility Theory violate Statewise Dominance in cases that involve possible states of zero probability.

5 Probability Discounting

This section discusses another alternative to Probability Fanaticism: discounting small probabilities.[53]

5.1 Discounting small probabilities

In response to cases that involve very small probabilities of huge payoffs, some have argued that we should discount very small probabilities down to zero—let's call this *Probability Discounting*. For example, Monton (2019) argues that small probabilities should be discounted down to zero, while Smith (2014*b*) argues that one is rationally permitted, but not required, to do so.[54] Probability Discounting avoids the counterintuitive implication that you should pay the stranger in Pascal's Mugging because it tells you to discount the tiny probability of the mugger telling the truth. Similarly, Probability Discounting allows one to value the St. Petersburg game at a reasonable price. In fact, Probability Discounting was originally proposed by Nicolaus Bernoulli as a solution to the St. Petersburg paradox.[55] He

[53]Note that, unlike here, 'discounting' typically does not mean ignoring altogether or bringing all the way down to zero. For example, 'temporal discounting' does not typically mean disvaluing positive outcomes in the future altogether, but instead, holding them less valuable than similar outcomes in the present.

[54]Smith argues that discounting small probabilities allows one to get a reasonable expected utility for the Pasadena game (see [Nover and Hájek 2004]). See Hájek (2014), Isaacs (2016) and Lundgren and Stefánsson (2020) for criticisms of discounting small probabilities. There is a related discussion on *de minimis* principles, on which a risk can be ignored or treated very differently from other risks if the risk is sufficiently small. See for example Peterson (2002) and Lundgren and Stefánsson (2020).

[55]Monton (2019) calls discounting small probabilities 'Nicolausian discounting' after Nicolaus Bernoulli. Other proponents of Probability Discounting include, for example, Buffon and Condorcet. See Hey et al. (2010) and Monton (2019, pp. 16–17).

writes: "[T]he cases which have a very small probability must be neglected and counted for nulls, although they can give a very great expectation."[56]

There are many ways of cashing out Probability Discounting. On one of the simplest versions of this view (i.e., *Naive Discounting*), one should conditionalize on very-small-probability outcomes not occurring and then maximize expected utility. On this view, there is some threshold probability such that outcomes whose probabilities are below this threshold are ignored. A slightly more complicated version (i.e., *Lexical Discounting*) uses very-small-probability outcomes as tiebreakers in cases where the prospects would otherwise be equally good. Both of these versions ignore *outcomes* associated with tiny probabilities. Instead, one could ignore *states* of the world that have tiny probabilities of occurring (as *State Discounting* does). Chapter 4 of this thesis discusses these and other versions of Probability Discounting in more detail. It explores what the most plausible version of Probability Discounting might look like and what are some problems such theories face.

5.2 Implications of Probability Discounting

Two chapters of this thesis examine the implications of Probability Discounting for population ethics and the value of the far future. These implications are briefly outlined below.

Population Ethics. The Repugnant Conclusion, introduced by Parfit, states:[57]

> "For any possible population of at least ten billion people, all with a

[56] Pulskamp (2013, p. 2).
[57] Parfit (1984, p. 388).

> very high quality of life, there must be some much larger imaginable population whose existence, if other things are equal, would be better even though its members have lives that are barely worth living."

The Repugnant Conclusion is a consequence of standard Total Utilitarianism. The Repugnant Conclusion strikes many as an unacceptable consequence, and various attempts at constructing an alternative population axiology to Total Utilitarianism have been made.[58] Nebel (2019) argues for the Repugnant Conclusion via the "Intrapersonal Repugnant Conclusion", on which certainty of a mediocre life is better for individuals than a sufficiently small chance of an excellent life. In Chapter 3 of this thesis, I deny that acceptance of the Intrapersonal Repugnant Conclusion leads us to the Repugnant Conclusion. I point out that on many views which avoid the Repugnant Conclusion, we should discount small probabilities down to zero to avoid an implausibly reckless decision theory. If we do, then Nebel's crucial premise of Ex Ante Pareto fails because discounting at the individual level can fail to match up with discounting at the population level. Thus, Probability Discounting helps us avoid the Repugnant Conclusion.

Value of the Far Future. Chapter 6 of this thesis discusses the implication of Probability Discounting for

> **Longtermism:** In the most important decision situations, our acts' expected influence on the value of the world is mainly determined by their possible consequences in the far future.[59]

[58] For an overview, see Greaves (2017).
[59] MacAskill (2019) and Greaves and MacAskill (2021).

According to Longtermism, morally speaking what matters the most is the far future. The case for Longtermism is straightforward: Given the enormous number of people who might exist in the far future, even a tiny probability of affecting how the far future goes outweighs the importance of our acts' consequences in the near term. But if we discount very small probabilities down to zero, we may have an objection to Longtermism provided that its truth depends on tiny probabilities of vast value. Contrary to this, Chapter 6 argues that discounting small probabilities does not undermine Longtermism. However, Probability Discounting might have implications for what longtermists should focus on.[60]

5.3 Problems with Probability Discounting

Probability Discounting might allow us to reject Probability Fanaticism and escape the Repugnant Conclusion. But it also faces some serious problems, as outlined below.

Threshold. One obvious problem with Probability Discounting is where the 'discounting threshold' is located. When are probabilities small enough to be discounted? Some have proposed possible thresholds. For example, Buffon suggested

[60] For example, Probability Fanaticism might imply that 'effective altruists' should accept Pascal's Wager. See footnote 4. They would have then made a full circle: Donate 10% of your income to your local church, mosque or synagogue. Or Probability Fanaticism might imply something weirder (see for example Wilkinson [2022, pp. 445–446]). In contrast, Probability Discounting allows one to escape this implication, provided that one's credence in heaven is low enough. Beckstead and Thomas (2020, §5) show that Probability Fanaticism leads to *Infinity Obsession*:

> **Infinity Obsession:** Any non-zero probability, no matter how small, of an infinite payoff is better than any finite payoff for sure.

that the threshold should be one-in-ten-thousand. And Condorcet gave an amusingly specific threshold: 1 in 144,768. Buffon chose his threshold because it was the probability of a 56-year-old man dying in one day—an outcome reasonable people usually ignore.[61] Condorcet had a similar justification.[62] More recently, Monton (2019, p. 17) has suggested a threshold of 1 in 2 quadrillion—significantly lower than the thresholds given by the historical thinkers. Monton (2019, §6.1) thinks that the threshold is subjective within reason: There is no single objective threshold for everybody.

However, there seems to be no way of choosing the discounting threshold such that Probability Discounting rules out all and only the objectionable choices.[63] For example, suppose the discounting threshold is just below 1 in 2 quadrillion. In that case, a prospect that gives any finite payoff for sure, no matter how good, is worse than a 1 in 2 quadrillion probability of some other finite payoff (assuming unbounded utilities). But a prospect with a 1 in 2 quadrillion probability does not seem less objectionable than a prospect with a slightly lower probability. So, Probability Discounting does not solve the problem it was meant to solve, as it still implies objectionably fanatical choices. However, this problem might be somewhat mitigated by letting the discounting threshold be vague.[64]

Individuation Problem. Another problem with Probability Discounting comes

[61] Hey et al. (2010, p. 257). See Monton (2019, pp. 8–9) for a discussion of Buffon's view.

[62] Condorcet's justification for his threshold is that 1 in 144,768 was the difference between the probability that a 47-year-old man would die within 24 hours and the probability that a 37-year-old man would, and that difference would not keep anyone awake at night. See Monton (2019, pp. 16–17).

[63] This point is raised by Beckstead and Thomas (2020, §3.5).

[64] Beckstead and Thomas (2020, p. 20). See also Lundgren and Stefánsson (2020, p. 911).

from individuating outcomes and states. The problem is that if we individuate outcomes/states very finely by giving a great deal of information about them, then all outcomes/states will have probabilities below the threshold. As discussed later in this thesis, one possible solution is to individuate outcomes by utilities. The idea is that outcomes/states are considered the "same" outcome/state if their associated utilities are the same.

Dominance Violations. One problem for some versions of Probability Discounting is that they violate dominance.[65] Imagine a lottery that gives you a tiny probability of some prize (and otherwise nothing), and compare this to a lottery that surely gives you nothing. The former lottery dominates the latter, but some versions of Probability Discounting say they are equally good. One can solve this dominance violation by considering very-small-probability outcomes/states as tiebreakers in cases where the prospects are otherwise equally good. However, this is not enough to avoid violating dominance because the resulting views still violate dominance in more complicated cases (as discussed in Chapter 4).

Money Pumps. Some versions of Probability Discounting, such as *Tail Discounting*, avoid the above mentioned dominance violations. According to Tail Discounting, one should first order all the possible outcomes of a prospect in terms of betterness. Then one should ignore the 'tails', that is, the very best and the very worst outcomes. Tail Discounting solves the problems with individuating outcomes and dominance violations. But it also has one big problem: It can be money pumped

[65] Isaacs (2016), Smith (2016), Monton (2019, pp. 20–21), Lundgren and Stefánsson (2020, pp. 912–914) and Beckstead and Thomas (2020, §2.3) also discuss Probability Discounting and dominance violations.

(as discussed in Chapter 4). So, someone with this view would end up paying for something they could have kept for free, which makes Tail Discounting less plausible as a theory of instrumental rationality.

In fact, vulnerability to exploitation by money pumps may be one of the most challenging problem for all versions of Probability Discounting. One money pump, in particular, presents a difficult challenge. Probability discounters are vulnerable to this money pump as a result of violating the Independence axiom of Expected Utility Theory. The basic problem for Probability Discounting is that by mixing gambles, one can arbitrarily reduce the probabilities of different states or outcomes within the compound lottery until these probabilities end up below the discounting threshold. Therefore, mixtures of gambles can end up being valued differently than the gambles that are mixed together. How probability discounters can avoid exploitation by money pumps is discussed in Chapter 4 and, in more detail, in Chapter 5.

Ex Ante Pareto. As discussed in Chapter 3, accepting Ex Ante Pareto and engaging in Probability Discounting gets one in trouble. Consider, for example, the following case:

> **Celebratory Gunfire:** Someone shoots into the air in an area full of people during a celebration, which causes people to feel excitement for a few seconds. The probability of any particular individual being hit by the bullet when it falls is negligibly small, but there is a high probability that someone is hit by it.

We may suppose that the value of everyone feeling excitement is not enough to outweigh the badness of the likely injury. However, the prospect of shooting into the air is *ex ante* better than not shooting for everyone; each individual feels excitement, and the probability of being hit by the bullet is rationally negligible. Thus, Ex Ante Pareto tells us that shooting into the air is right, even though the bullet will almost certainly hit someone. So, if one accepts Probability Discounting, one should reject Ex Ante Pareto, or one would permit the infliction of arbitrarily severe harms for little or no benefits.

Each-We Dilemmas. Another problem Probability Discounting faces is Each-We Dilemmas, which will be discussed in Chapter 6. According to Parfit (1984, p. 91), a theory faces Each-We Dilemmas if "there might be cases where, if each does better in this theory's terms, we do worse, and vice versa." Each-We Dilemmas arise for Probability Discounting for the same reason as violations of Ex Ante Pareto arise: Probabilities can accumulate. If many individuals discount a tiny probability of some event happening, and the probabilities are sufficiently independent for the different agents, then the total discounted probability can be high. This can result in catastrophic outcomes. Consider, for example, the following case:

> **Asteroid:** An asteroid is heading toward the Earth and will almost certainly hit unless stopped. There are multiple asteroid defense systems, and (unrealistically) each has a tiny probability of hitting the asteroid and preventing a catastrophe. However, the probability that one of them succeeds is high if enough of them try. Attempting to stop

the asteroid involves some small cost ϵ.

If agents discount the probability of them successfully stopping the asteroid and consequently do nothing, then the asteroid will almost certainly hit the Earth. But this outcome could be prevented almost certainly if enough agents attempt to do so. To solve these kind of cases, probability-discounting agents would need to somehow take into account the choices other people face and consider whether the collective has a non-negligible chance of making a difference. However, this solution leads us to another problem: violations of Separability.

Separability. Probability Discounting violates Separability if the choices other people face can affect what you ought to do, even when the other agents are far away and you cannot influence what goes on near them. The solution to Each-We Dilemmas asks us to change our actions depending on what choices other agents face. For example, if there was only a single asteroid defense system, then Probability Discounting would recommend that the agent operating it not attempt to stop the asteroid. However, if there are multiple asteroid defense systems, then this approach would recommend attempting to stop the asteroid because the probability that someone successfully stops it is non-negligible.

Earlier it was shown that Stochastic Dominance and Separability are jointly inconsistent. In Russell's (2021, pp. 13–14) words: "This looks like very bad news for Separability." Since violating Separability is a problem for all theories (on pain of violating Stochastic Dominance), violating Separability may not seem especially worrying for Probability Discounting. However, it was only shown that Stochastic Dominance and Separability are inconsistent in a case where the outcomes (of

the near and far prospects) are correlated. In contrast, Probability Discounting violates Separability even when the outcomes are probabilistically independent for the different agents. We might think that probabilistic independence makes violating Separability even worse.

To summarize, I have discussed some problems Probability Discounting faces. These include choosing the discounting threshold, individuating outcomes/states, violating dominance, vulnerability to money pumps, violating Ex Ante Pareto, facing Each-We Dilemmas and violating Separability. These problems will be discussed in more detail in the following chapters of this thesis. The next section discusses some alternative approaches to tiny probabilities of vast value.

6 Alternatives

This section discusses other approaches suggested in response to cases that involve tiny probabilities of huge payoffs.

6.1 Conditionalizing on knowledge

One possibility is to conditionalize on one's knowledge before maximizing expected utility—let's call this *Knowledge-Based Discounting*.[66] It might be argued that, in Pascal's Mugging, you *know* that the mugger will not deliver a thousand quadrillion happy days in the Seventh Dimension. And, possibly, you also know that you will

[66] See Hong (n.d.) and Francis and Kosonen (n.d.) on Knowledge-Based Discounting.

not gain a great payoff with the St. Petersburg game.[67] Thus, conditionalizing on knowledge before maximizing expected utility could solve at least some cases with tiny probabilities of huge payoffs.

But Knowledge-Based Discounting is vulnerable to some of the same problems Probability Discounting faces, such as money pumps.[68] Consider, for example, the following lotteries:

> *Ticket A* Gives a great payoff if you guess all seven lottery numbers correctly (and otherwise it gives nothing).
>
> *Ticket B* Gives a modest positive payoff if you guess at least five lottery numbers correctly (and otherwise it gives nothing).

Suppose you know that ticket A wins nothing, but you do not know that ticket B wins nothing. If it is possible to have knowledge in lottery cases, then there must be some (possibly vague and context-dependent) threshold probability for when a probability is high enough to count as knowledge. We may suppose that the probability of not winning with A is above this threshold, but the probability of not winning with B is below this threshold. Consequently, B is worth some positive amount, while A is worthless (or at most better than nothing). The setup is as follows: You currently have B. If you guess at least five lottery numbers right,

[67] This claim is more contested. It is often argued that one cannot have knowledge in lottery cases, as it seems that one does not know that one's lottery ticket will not win, even though it is very unlikely to win. For a discussion of lottery cases, see for example Smith (2014*a*). See Hong (n.d.) for a defense of Knowledge-Based Discounting in the context of the St. Petersburg paradox. If Knowledge-Based Discounting is to avoid Probability Fanaticism in all cases, then it must be possible to have knowledge in lottery cases, such as the St. Petersburg paradox.

[68] See Francis and Kosonen (n.d.).

then you will be offered A in exchange for B. But if you learn that you guessed five lottery numbers right, you no longer know that you did not guess all seven numbers right. In that case, you would only need to have guessed two more numbers right, and for all you know, you might have. So, you then would prefer A to B and happily accept the trade.

This is unfortunate. Right now, you know that you will not win anything with A. So it would be better to keep B. However, you also know that if you win anything with B, you will accept the trade and end up with A. Luckily, you are offered a chance to avoid this situation: If you pay some amount, you will not be offered A in exchange for B in case you guess at least five numbers right. And, given that B is worth some positive amount while A is worth nothing, you accept this offer. But you have then paid for something you could have kept for free.

More generally, Knowledge-Based Discounting gets you in trouble if there can be cases where you know that P, but some evidence would make you lose the knowledge that P and you do not know that such evidence will not arise.[69] In this case, although you know that A wins nothing, this belief loses the status of knowledge if you guess at least five lottery numbers right. And you do not know that you will not guess at least five lottery numbers right.

To summarize, Knowledge-Based Discounting advises one to conditionalize

[69] Knowledge-Based Discounting might escape this problem if one accepts the *KK Principle*: If one knows that P, then one also knows that one knows it. If the KK principle is true, then either you do not know that you will not guess seven numbers correctly (so A is worth some positive amount), or you know that you will not guess at least five numbers correctly (so neither A nor B is worth any positive amount). But you cannot know that you will not guess seven numbers correctly and be uncertain about whether you might lose this knowledge.

on one's knowledge before maximizing expected utility. Similarly to Probability Fanaticism and Probability Discounting, Knowledge-Based Discounting is diachronically inconsistent and thus vulnerable to money pumps.

6.2 Assigning zero probability

It seems that every approach to tiny probabilities of huge payoffs has serious shortcomings. In order to escape the paradoxes with non-simple lotteries, one might be tempted to assign a zero probability to the possibility of the St. Petersburg game (and its variants).[70] The idea is that one can accept Expected Utility Theory and escape the paradoxical results discussed earlier in this chapter, such as money pumps.

However, this solution seems *ad hoc*. Assignments of probability should only respond to *epistemic reasons*. They should not respond to instrumental reasons, such as getting money pumped.[71] Yet, arguments for Bayesianism often rely on such instrumental reasons: Unless one uses conditionalization to update credences, one will get Dutch Booked. However, something else might be going on in these arguments. Succumbing to a Dutch Book is an indication that one's beliefs about the world are inconsistent. So, the argument for conditionalization is not that failing to conditionalize gets one Dutch Booked, and that is instrumentally bad. Instead,

[70] Various people have suggested this (personal correspondence). Note that this proposal does not avoid Probability Fanaticism—its only purpose is to make Expected Utility Theory behave well with non-simple lotteries. But it says nothing about cases such as Pascal's Mugging.

[71] This is controversial. For example, in epistemology, there is a view that rejects the claim that only epistemic reasons should influence beliefs:

> **Pragmatic Encroachment:** A difference in pragmatic circumstances can constitute a difference in knowledge.

See Ichikawa and Steup (2018, §12).

the argument is that getting Dutch Booked is a symptom of having inconsistent beliefs.[72] So, Dutch Book arguments need not rely on the idea that beliefs (or credence assignments) should respond to instrumental reasons. However, similarly, one might insist that getting money pumped because one accepts Probability Fanaticism is a symptom of having inconsistent beliefs. The money pump shows that there is something wrong with St. Petersburg-style probability and utility assignments. But it is hard to see why that would be the case.

7 Conclusion

Cases that involve tiny probabilities of vast value present a puzzle as it seems that all approaches have implausible implications. The main approaches discussed were Probability Fanaticism, Boundedness and Probability Discounting. First, the chapter discussed two arguments for maximizing expected utility: the long-run argument and representation theorems. Next, it explored Probability Fanaticism, on which tiny probabilities of large positive or negative payoffs can have enormous positive or negative expected utility (respectively). We saw that there are strong arguments for and against Probability Fanaticism. Then, the chapter discussed the possibility that utilities are bounded. Boundedness will be discussed in more de-

[72] This is, in fact, what Lewis (1999, pp. 404–405) argues: "Note also that the point of any Dutch book argument is not that it would be imprudent to run the risk that some sneaky Dutchman will come and drain your pockets. After all, there aren't so many sneaky Dutchmen around; and anyway, if ever you see one coming, you can refuse to do business with him. Rather, the point is that if you are vulnerable to a Dutch book, whether synchronic or diachronic, that means that you have two contradictory opinions about the expected value of the very same transaction. To hold contradictory opinions may or may not be risky, but it is in any case irrational."

tail in Chapters 1 and 2 of this thesis. Finally, the chapter investigated Probability Discounting, on which tiny probabilities should be ignored in practical decision-making. Probability Discounting will be the focus of Chapters 3–6. Some other approaches were also discussed briefly. To conclude, paradoxes concerning tiny probabilities of vast value show that some intuitively compelling principles of rationality must be given up.

References

Baumann, P. (2009), 'Counting on numbers', *Analysis* **69**(3), 446–448.

Beckstead, N. (2013), On the overwhelming importance of shaping the far future, PhD thesis, Rutgers, the State University of New Jersey.

Beckstead, N. and Thomas, T. (2020), 'A paradox for tiny probabilities and enormous values'. Global Priorities Institute Working Paper No. 10–2020.
URL: *https://globalprioritiesinstitute.org/nick-beckstead-and-teruji-thomas-a-paradox-for-tiny-probabilities-and-enormous-values/*

Bernoulli, D. (1954), 'Exposition of a new theory on the measurement of risk', *Econometrica* **22**(1), 23–36.

Bostrom, N. (2009), 'Pascal's Mugging', *Analysis* **69**(3), 443–445.

Briggs, R. A. (2019), Normative Theories of Rational Choice: Expected Utility, *in*

E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', Fall 2019 edn, Metaphysics Research Lab, Stanford University.

Buchak, L. (2013), *Risk and Rationality*, Oxford University Press, Oxford.

Chalmers, D. J. (2002), 'The St. Petersburg two-envelope paradox', *Analysis* **62**(2), 155–157.

Fishburn, P. C. (1970), *Utility Theory for Decision Making*, Wiley, New York.

Francis, T. and Kosonen, P. (n.d.), 'Ignore outlandish possibilities'. Unpublished manuscript.

Goodsell, Z. (2021), 'A St Petersburg paradox for risky welfare aggregation', *Analysis* **81**(3), 420–426.

Greaves, H. (2017), 'Population axiology', *Philosophy Compass* **12**(11), e12442.

Greaves, H. and MacAskill, W. (2021), 'The case for strong longtermism'. Global Priorities Institute Working Paper No. 5–2021.
URL: *https://globalprioritiesinstitute.org/hilary-greaves-william-macaskill-the-case-for-strong-longtermism-2/*

Gustafsson, J. E. (forthcoming), *Money-Pump Arguments*, Cambridge University Press, Cambridge.

Hadar, J. and Russell, W. R. (1969), 'Rules for ordering uncertain prospects', *The American Economic Review* **59**(1), 25–34.

Hájek, A. (2014), 'Unexpected expectations', *Mind* **123**(490), 533–567.

Hammond, P. J. (1998), Objective expected utility: A consequentialist perspective, *in* S. Barberà, P. J. Hammond and C. Seidl, eds, 'Handbook of Utility Theory Volume 1: Principles', Kluwer, Dordrecht, pp. 143–211.

Hanson, R. (2007), 'Pascal's Mugging: Tiny probabilities of vast utilities'.
URL: *https://www.lesswrong.com/posts/a5JAiTdytou3Jg749/pascal-s-mugging-tiny-probabilities-of-vast-utilities?commentId=Q4ACkdYFEThA6EE9P*

Hey, J. D., Neugebauer, T. M. and Pasca, C. M. (2010), Georges-Louis Leclerc de Buffon's 'Essays on moral arithmetic', *in* A. Sadrieh and A. Ockenfels, eds, 'The Selten School of Behavioral Economics: A Collection of Essays in Honor of Reinhard Selten', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 245–282.

Hong, F. (n.d.), 'What do you know in St. Petersburg? An exploration of "knowledge-first" decision theory'. Unpublished manuscript.

Ichikawa, J. J. and Steup, M. (2018), The Analysis of Knowledge, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', Summer 2018 edn, Metaphysics Research Lab, Stanford University.

Isaacs, Y. (2016), 'Probabilities cannot be rationally neglected', *Mind* **125**(499), 759–762.

Jensen, N. E. (1967), 'An introduction to Bernoullian utility theory: I. Utility functions', *The Swedish Journal of Economics* **69**(3), 163–183.

Kreps, D. M. (1988), *Notes on the Theory of Choice*, Westview Press, Boulder.

Lehmann, E. L. (1955), 'Ordered families of distributions', *The Annals of Mathematical Statistics* **26**(3), 399–419.

Lewis, D. (1999), Why conditionalize?, *in* 'Papers in Metaphysics and Epistemology', Vol. 2 of *Cambridge Studies in Philosophy*, Cambridge University Press, Cambridge, pp. 403–407.

Lundgren, B. and Stefánsson, H. O. (2020), 'Against the *De Minimis* principle', *Risk Analysis* **40**(5), 908–914.

MacAskill, W. (2019), 'Longtermism', Effective Altruism Forum.
URL: *https://forum.effectivealtruism.org/posts/qZyshHCNkjs3TvSem/longtermism*

Mann, H. B. and Whitney, D. R. (1947), 'On a test of whether one of two random variables is stochastically larger than the other', *The Annals of Mathematical Statistics* **18**(1), 50–60.

McMahan, J. (1981), 'Problems of population theory', *Ethics* **92**(1), 96–127.

Menger, K. (1934), 'Das unsicherheitsmoment in der wertlehre: Betrachtungen im anschliß an das sogenannte petersburger spiel', *Zeitschrift für Nationalökonomie / Journal of Economics* **5**(4), 459–485.

Menger, K. (1967), The role of uncertainty in economics, *in* M. Shubik, ed., 'Es-

says in Mathematical Economics: In Honor of Oskar Morgenstern', Princeton University Press, Princeton, pp. 211–231.

Monton, B. (2019), 'How to avoid maximizing expected utility', *Philosophers' Imprint* **19**(18), 1–24.

Nebel, J. M. (2019), 'An intrapersonal addition paradox', *Ethics* **129**(1), 309–343.

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

Parfit, D. (1984), *Reasons and Persons*, Clarendon Press, Oxford.

Pascal, B. (1958), *Pascal's Pensées*, E. P. Dutton & Co., New York.
URL: *https://www.gutenberg.org/files/18269/18269-0.txt*

Peterson, M. (2002), 'What is a *de minimis* risk?', *Risk Management* **4**(2), 47–55.

Pulskamp, R. J. (2013), 'Correspondence of Nicolas Bernoulli concerning the St. Petersburg Game'. Unpublished manuscript. Accessed through: https://web.archive.org/.
URL: *http://cerebro.xu.edu/math/Sources/NBernoulli/correspondence_petersburg_game.pdf*

Quirk, J. P. and Saposnik, R. (1962), 'Admissibility and measurable utility functions', *The Review of Economic Studies* **29**(2), 140–146.

Russell, J. S. (2021), 'On two arguments for fanaticism'. Global Priorities Institute Working Paper No. 17–2021.

URL: *https://globalprioritiesinstitute.org/on-two-arguments-for-fanaticism-jeff-sanford-russell-university-of-southern-california/*

Russell, J. S. and Isaacs, Y. (2021), 'Infinite prospects', *Philosophy and Phenomenological Research* **103**(1), 178–198.

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

Savage, L. J. (1972), *The Foundations of Statistics*, 2 edn, Dover Publications, New York.

Smith, M. (2014*a*), 'What else justification could be', *Noûs* **44**(1), 10–31.

Smith, N. J. J. (2014*b*), 'Is evaluative compositionality a requirement of rationality?', *Mind* **123**(490), 457–502.

Smith, N. J. J. (2016), 'Infinite decisions and rationally negligible probabilities', *Mind* **125**(500), 1199–1212.

Tarsney, C. (2020), 'Exceeding expectations: Stochastic dominance as a general decision theory'. Global Priorities Institute Working Paper No. 3–2020.
URL: *https://globalprioritiesinstitute.org/christian-tarsney-exceeding-expectations-stochastic-dominance-as-a-general-decision-theory/*

von Neumann, J. and Morgenstern, O. (1947), *Theory of Games and Economic Behavior*, 2 edn, Princeton University Press, Princeton.

Wilkinson, H. (2022), 'In defence of fanaticism', *Ethics* **132**(2), 445–477.

Yudkowsky, E. (2007*a*), 'A comment on Pascal's Mugging: Tiny probabilities of vast utilities'.

URL: *https://www.lesswrong.com/posts/a5JAiTdytou3Jg749/pascal-s-mugging-tiny-probabilities-of-vast-utilities?commentId=kqAKXskjohx4SSyp4*

Yudkowsky, E. (2007*b*), 'Pascal's Mugging: Tiny probabilities of vast utilities'.

URL: *http://www.overcomingbias.com/2007/10/pascals-mugging.html*

Zynda, L. (2000), 'Representation theorems and realism about degrees of belief', *Philosophy of Science* **67**(1), 45–69.

Chapter 1

*Bounded Utilities and Ex Ante Pareto**

ABSTRACT: This chapter shows that decision theories on which utilities are bounded, such as standard axiomatizations of Expected Utility Theory, violate Ex Ante Pareto if combined with an additive axiology, such as Total Utilitarianism. A series of impossibility theorems point toward Total Utilitarianism as the right account of axiology, while money-pump arguments put Expected Utility Theory in a favorable light. However, it is not clear how these two views can be reconciled. This question is particularly puzzling if utilities are bounded (as standard axiomatizations of Expected Utility Theory imply) because the total quantity of well-being might be infinite or arbitrarily large. Thus, there must be a non-linear transformation from the total quantity of well-being into utilities used in decision-making. This non-linear transformation is also required if one wishes to avoid Probability Fanaticism. However, such a transformation leads to violations of Ex Ante Pareto. So, the reconciliation of Expected Utility Theory and Total Utilitarianism prescribes prospects that are better for none and worse for some.

*I wish to thank Gustav Alexandrie, Tomi Francis, Andreas Mogensen and Teruji Thomas for valuable feedback on this chapter.

This chapter investigates the compatibility of two standard theories: Total Utilitarianism and Expected Utility Theory with a bounded utility function. Let's call the combination of these views *Bounded Expected Totalism*. Unfortunately, this chapter argues that Bounded Expected Totalism violates *Ex Ante Pareto*, the principle that what is ex ante better for everyone is better overall.[1,2] This principle is often used by utilitarians to justify their theory in opposition to others, such as prioritarianism and egalitarianism.

Insofar as Expected Utility Theory is the dominant theory of choice under uncertainty, this argument could be seen as undermining Total Utilitarianism. However, I ultimately take the lesson to be different. As I explained in the introductory chapter, considerations of Probability Fanaticism motivate either the use of a bounded utility function or some form of Probability Discounting (or perhaps some third option). So, I take the arguments in this chapter to speak differentially in favor of Probability Discounting. Actually, as I will explain in Chapter 3, Probability Discounting also leads to violations of Ex Ante Pareto. So, to put the point more accurately, the plausibility of Ex Ante Pareto does not favor Bounded Expected Utility Theory over Probability Discounting.

The chapter proceeds as follows. I will first define Bounded Expected Total-

[1]There is no inconsistency with Harsanyi's social aggregation theorem. As will be explained later, a bounded expected totalist must reject Harsanyi's conclusion, so they cannot accept all his premises.

[2]This chapter focuses on the compatibility of Expected Utility Theory and Total Utilitarianism, but the problem with Ex Ante Pareto arises for, for example, Critical-Level Utilitarianism in exactly the same way. The problem also arises for Average Utilitarianism and many other theories if individual utilities are unbounded. See for example the argument in Goodsell (2021), which applies to any axiology that is utilitarian in same-number cases. This chapter shows that, even if utilities are bounded, Total Utilitarianism combined with Expected Utility Theory violates Ex Ante Pareto.

ism more formally and explain why this is a prima facie attractive view. I will then proceed to illustrate why this view must violate Ex Ante Pareto. A background issue, which is laid out in §2.2, is how the well-being of a single individual can be traded off between different states of nature. The question is essentially whether the personal value of prospects is risk-averse with respect to well-being. I will give separate examples of Ex Ante Pareto violations that involve risk-neutrality (§3) and risk-aversion (§4). §5.2 gives a general argument for why Bounded Expected Totalism must violate Ex Ante Pareto. I conclude in §6 by sketching how my examples relate to the classic result in this area, namely, Harsanyi's social aggregation theorem.

1 Background

This section introduces some background. First, it explains Total Utilitarianism and Expected Utility Theory. Then, it discusses the idea that utilities are bounded and why this follows from standard axiomatizations of Expected Utility Theory. Lastly, it discusses bounded utilities as a possible way of getting intuitively right recommendations in cases that involve tiny probabilities of huge payoffs.

1.1 Total Utilitarianism and Expected Utility Theory

A series of impossibility theorems point toward Total Utilitarianism as the right account of axiology, while money-pump arguments put Expected Utility Theory

in a favorable light.[3] The former view states that a population is better than another just in case the total quantity of well-being it contains is greater, while the latter states that a prospect is better than another just in case its expected utility is greater.[4] Let $X \succsim Y$ mean that X is at least as good as Y. Also, let W(A) denote the total quantity of well-being in the state of affairs A and let w(S_i) denote the well-being of individual S_i. Then, more formally, Total Utilitarianism states the following:

> **Total Utilitarianism:** For all states of affairs A and B (in which n and m individuals exist, respectively), $A \succsim B$ if and only if W $(A) \geq$ W (B), where
>
> $$\text{W}(A) = \sum_{i=1}^{n} w(S_i) \text{ and W}(B) = \sum_{i=1}^{m} w(S_i).^{5}$$

Next, let EU (X) denote the expected utility of prospect X, $p(E_i)$ the probability of event E_i and $u(x_i)$ the utility of outcome x_i (which results from event E_i). Then, Expected Utility Theory states the following:[6]

---

[3] See for example Arrhenius (2000) and Gustafsson (forthcoming). The impossibility theorems point toward Total Utilitarianism because they show that we cannot escape the Repugnant Conclusion without being forced to accept even more unpalatable conclusions. See also Zuber et al. (2021).

[4] In the case of Total Utilitarianism, 'better' is used in an axiological sense; in the case of Expected Utility Theory, 'better' is concerned with instrumental rationality.

[5] To cover cases in which an infinite number of individuals exist in state of affairs A, we may extend Total Utilitarianism as follows:

$$\text{W}(A) = \sum_{i=1}^{\infty} w(S_i).$$

[6] I am assuming that prospects can be countably infinite, that is, assign a non-zero probability

Expected Utility Theory: For all prospects X and Y, $X \succsim Y$ if and only if $\text{EU}(X) \geq \text{EU}(Y)$, where

$$\text{EU}(X) = \sum_{i=1}^{\infty} p(E_i)\, u(x_i).$$

Combining Total Utilitarianism and Expected Utility Theory with a bounded utility function, we get *Bounded Expected Totalism*:

Bounded Expected Totalism: Total Utilitarianism and Expected Utility Theory with a bounded utility function are both true.

1.2 Boundedness

What does it mean for utilities to be bounded? If utilities are real-valued, then boundedness means the following:

Boundedness: There is some $M \in \mathbb{R}$ such that for all outcomes x, $|u(x)| < M$.

In other words, Boundedness rules out arbitrarily and infinitely good outcomes.

Standard axiomatizations of expected utility maximization require utilities to be bounded.[7] Consider, for example, the von Neumann-Morgenstern axiomatization of Expected Utility Theory.[8] Let $X \succ Y$ mean that X is better than Y. Also, let

to countably infinite number of outcomes. This assumption is needed because some of the cases discussed in this chapter involve such prospects.

[7] See for example Kreps (1988, pp. 63–64), Fishburn (1970, pp. 194, 206–207), Hammond (1998, pp. 186–191) and Russell & Isaacs (2021).

[8] The following axioms together entail Expected Utility Theory: Completeness, Transitivity, In-

XpY be a risky prospect with a p chance of prospect X obtaining and a $1-p$ chance of prospect Y obtaining. Then, if prospects are compared by their expected utilities, Boundedness follows from the following von Neumann-Morgenstern axiom:

> **Continuity:** If $X \succ Y \succ Z$, then there are probabilities p and $q \in (0, 1)$ such that $XpZ \succ Y \succ XqZ$.

To see why Continuity implies Boundedness (assuming that prospects are compared for their expected utilities), let's consider the two ways in which Boundedness might be false.[9] First, Boundedness might be false because there is an infinite sequence of prospects A_1, A_2, A_3, … such that A_2 is at least twice as good as A_1, A_3 is at least twice as good as A_2, and so on, with respect to some baseline. Let A be a mixed prospect that assigns probability $1/2^k$ to prospect A_k. Then, we have that

$$\mathrm{EU}(A) = \sum_{i=1}^{\infty} p(A_i)\, u(A_i) = \infty.$$

---

dependence and Continuity. See von Neumann & Morgenstern (1947), Jensen (1967, pp. 172–182) and Hammond (1998, pp. 152–164).

[9] Boundedness is false if *Limitedness* or *Finiteness* is false:

> **Limitedness:** There is no infinite sequence of prospects X_1, X_2, X_3, … such that X_2 is at least twice as good (bad) as X_1, X_3 is at least twice as good (bad) as X_2, and so on, with respect to some baseline Z.
>
> **Finiteness:** No prospect is infinitely better (worse) than another good (bad) prospect.

Limitedness is from Russell & Isaacs (2021, p. 12). Limitedness, unlike Finiteness, allows infinite utilities (as long as there are no series of at least twice as good prospects with infinite expected utility). Russell & Isaacs (2021) show that Countable Independence rules out violations of Limitedness (via St. Petersburg-style cases). However, Countable Independence does not rule out unbounded utilities, as some prospects might still be infinitely better than other prospects. Given St. Petersburg-style cases, Finiteness implies Limitedness.

Next, choose some prospects B and C such that $\infty > \text{EU}(B) > \text{EU}(C) > -\infty$. Then, we have that A is better than B, which is better than C. However, for all probabilities $q \in (0, 1)$, $\text{EU}(AqC) = \infty$. Therefore, AqC is better than B for all probabilities $q \in (0, 1)$. This is a violation of Continuity.[10]

Secondly, and more generally, Boundedness is false if some prospect A is infinitely better than another (good) prospect B. This leads to a violation of Continuity because the mixed prospect ApC (where C certainly gives nothing) is better than B for all probabilities $p \in (0, 1)$. So, the supposition that Boundedness is false leads to violations of Continuity. Thus, it follows from Continuity that Boundedness is true.[11]

1.3 Probability Fanaticism

Boundedness has been discussed as a possible alternative to Probability Fanaticism.[12] Probability Fanaticism is the idea that tiny probabilities of large positive or negative payoffs can have enormous positive or negative expected utility (respectively):[13]

Probability Fanaticism:

i *Positive Probability Fanaticism* For any probability $p > 0$, and

[10] This is a modified argument from Kreps (1988, pp. 63–64).

[11] These arguments show that Continuity implies an upper bound on utilities. One can give similar arguments to show that Continuity implies a lower bound on utilities.

[12] See for example Beckstead & Thomas (2020).

[13] Wilkinson (2022, p. 449). For discussions related to Probability Fanaticism, see Beckstead (2013, ch. 6), Beckstead & Thomas (2020), Goodsell (2021), Russell & Isaacs (2021), Russell (2021) and Wilkinson (2022).

for any finite utility u, there is some large enough utility U such that probability p of U (and otherwise nothing) is better than certainty of u.[14]

ii *Negative Probability Fanaticism* For any probability $p > 0$, and for any finite negative utility $-u$, there is some large enough negative utility $-U$ such that probability p of $-U$ (and otherwise nothing) is worse than certainty of $-u$.

If utilities are bounded, then sufficiently small probabilities of even very good (or very bad) outcomes do not contribute much to the expected utility of a prospect. For a given probability, there is an upper/lower bound on the contribution to expected utility from outcomes associated with that probability. If the probability gets smaller, this bound also shrinks proportionally so that small enough probabilities cannot help but contribute only a small amount of expected (positive or negative) utility.

For any tiny probability of a great outcome, there is still some certain modest positive outcome that is worse. However, it is not the case that for any certain modest positive outcome, an *arbitrarily* small probability of a sufficiently great outcome is better. If the probability of the great outcome is small enough, increases in the payoff can no longer compensate for decreases in its probability. So, Boundedness prevents such outcomes from dominating the expected utility calculations, and thus, it escapes Probability Fanaticism (assuming fixed upper and lower bounds

[14] In this context, 'otherwise nothing' means retaining the status quo or baseline outcome.

on utilities).

Let's call a case *fanatical* if tiny probabilities of enormous positive or negative outcomes dominate the expected utility calculations in that case. One example of a fanatical case is *Pascal's Mugging*:[15]

> **Pascal's Mugging:** A stranger approaches Pascal and claims to be an Operator from the Seventh Dimension. The stranger promises to perform magic that will help quadrillions of orphans in the Seventh Dimension if Pascal pays the mugger ten livres.

Pascal thinks that the mugger is almost certainly lying. However, if utilities are unbounded, the mugger can always increase the payoff until the offer has positive expected utility—at least if Pascal assigns some non-zero probability to the mugger being able and willing to deliver any finite quantity of utility for Pascal.[16] Then, with some number of orphans, the expected-utility-maximizing act is to pay the mugger ten livres. Moreover, the mugger can also ask for more money and increase the payoff accordingly. So, someone who maximizes expected utility with an unbounded utility function would be willing to pay any sum, provided that the payoff is sufficiently large.

In contrast, Bounded Expected Totalism has upper and lower bounds on utilities. Consequently, there is an upper limit to how much a bounded expected to-

[15] Bostrom (2009). The case presented here is a slightly modified version of Bostrom's case. In Bostrom's case, the mugger promises to give Pascal an extra thousand quadrillion happy days and help many orphans in the Seventh Dimension. The case is based on informal discussions by various people, including Eliezer Yudkowsky (2007*b*).

[16] Contrary to this, see Hanson (2007), Yudkowsky (2007*a*) and Baumann (2009).

talist would be willing to pay the mugger (assuming fixed upper and lower bounds on utilities). Bounded Expected Totalism does not escape the mugging entirely because, for any payoff offered by the mugger, there is *some* amount a bounded expected totalist would pay. After all, a tiny chance of obtaining the upper or avoiding the lower limit of utilities is worth something. But at least a bounded expected totalist would not lose all their money.[17] So, Bounded Expected Totalism helps avoid the worst instances of Probability Fanaticism (again assuming fixed upper and lower bounds on utilities).

However, this chapter shows that, under some circumstances, Bounded Expected Totalism violates Ex Ante Pareto: It prescribes prospects that are better for no one and worse for some.

> **Ex Ante Pareto:** For all prospects X and Y, if X is at least as good as Y for everyone, and X is better than Y for some, then X is better than Y.

Bounded Expected Totalism violates Ex Ante Pareto if there is a non-zero probability that an infinite or arbitrarily large number of individuals exist. But it also violates Ex Ante Pareto if it avoids Probability Fanaticism (as I will explain shortly).

2 Bounded Expected Totalism

This section presents Bounded Expected Totalism in more detail and discusses the cardinal structure of well-being.

[17] This may not be true if the mugger repeatedly returns with the same offer.

2.1 The social transformation function

Let *well-being* refer to how good some outcome is for an individual. And, let *social utility* refer to how good some outcome is overall, from an axiological point of view. Also, let *expected individual utility* represent how good some prospect is for an individual, and let *expected social utility* represent how good some prospect is overall. In the context of Expected Utility Theory, I will denote these by $\mathrm{EU}_{\mathrm{Ind}}$ and $\mathrm{EU}_{\mathrm{Soc}}$, respectively. In general, I will use *individual betterness* to refer to betterness from an individual's point of view. Similarly, I will use *overall/impersonal betterness* to refer to betterness from a moral point of view.

To combine Total Utilitarianism and Expected Utility Theory, we need a *social transformation function* that takes the total quantity of well-being as input and gives social utilities as output. This transformation function must be non-linear if an infinite or arbitrarily large number of happy individuals might exist, as then the total sum of individuals' well-being might be infinite or arbitrarily large (and similarly for negative well-being).[18] But Bounded Expected Totalism requires expected social utilities to be bounded. So, the expected social utilities assigned to prospects that might result in an infinite or arbitrarily large number of happy individuals must be bounded.[19]

[18]Note that the total quantity of well-being is not necessarily infinite if an infinite number of individuals exist. For example, suppose that for each individual $k \in \{1, 2, \ldots\}$, k's well-being measure takes a value in the interval $(0, 2^{-k})$. Then, an infinite number of individuals exist but the total quantity of well-being is bounded. However, this can be ruled out by requiring the individual well-being measures to have the same range.

[19]Beckstead & Thomas (2020, p. 9) write that Boundedness conflicts with the most natural understanding of utilitarianism as an evaluative theory on which improving n lives by a given amount improves the world by n times as much as improving one life. Similarly, they point out that Total

One might object that the total quantity of well-being cannot be infinite or arbitrarily large because there is an upper limit to how many individuals might exist. This upper limit might be due to, for example, the Universe being finite. However, this may not be true, so we need a decision theory that can also handle these possibilities.[20] If there is even a tiny probability that an infinite or arbitrarily large number of individuals exist, then the transformation function must be non-linear for utilities to be bounded. Consider for example the following versions of Pascal's Mugging:

> **Pascal's Mugging (infinite orphans):** The mugger promises to perform magic that will help an *infinite* number of orphans in the Seventh Dimension if Pascal pays the mugger ten livres.

> **Pascal's Mugging (St. Petersburg case):** The mugger promises to perform magic that gives a $1/2$ probability of helping two orphans, a $1/4$ probability of helping four orphans, a $1/8$ probability of helping eight orphans, and so on if Pascal pays the mugger ten livres.

Suppose Pascal has a non-zero credence in the mugger telling the truth. In that case, he needs to assign some expected social utility to the possibility of helping an infinite or arbitrarily large number of orphans. And, if utilities are bounded,

---

Utilitarianism and its variants put unbounded value on creating good lives.

[20] As Branwen (2009) put it: "Scientists have suggested infinite universes on multiple occasions, and we cannot rule the idea out on any logical ground. Should our theory of rationality stand or fall on what the cosmologists currently think?" Also, Bostrom (2011, p. 10) writes that recent cosmological evidence suggests that the world is probably infinite, which means that it contains an infinite number of galaxies, stars and planets. And, Bostrom writes, if there are an infinite number of planets, then there is, with probability one, an infinite number of people.

then the utility assigned cannot be infinite. Thus, the social transformation function must be non-linear. Moreover, anyone could be confronted with these kind of offers. So, we need a theory that can handle cases such as these.

In the previous two cases, the mugger promises to help an infinite number of orphans in expectation, which forces the social transformation function to be non-linear.[21] However, even if the mugger does not promise to help an infinite number of individuals in expectation, Bounded Expected Totalism does not avoid Probability Fanaticism if the social transformation function is linear and there is no upper limit to how many individuals might exist. For example, the mugger can always promise to help a greater number of orphans and thus increase the payoff arbitrarily high:

> **Pascal's Mugging (any number of orphans):** The mugger promises to perform magic that will help n number of orphans, where n is finite but arbitrarily large.

If social utilities are linear with the total quantity of well-being, then Bounded Expected Totalism recommends paying the mugger any sum of money, provided that the number of orphans is sufficiently high. That is, for any tiny probability p of the mugger telling the truth, and for any sum of money x, there is some finite number of orphans n, such that Pascal ought to pay the mugger x if the mugger promises to help n orphans. Thus, Bounded Expected Totalism does not avoid

[21] We might object that Total Utilitarianism is not intended to apply in infinite cases. After all, in infinite cases, the total quantity of well-being is not well-defined. So, we might think that Total Utilitarianism does not make sense if there might be an infinite number of individuals.

Probability Fanaticism if there is no upper limit to how many individuals might exist and the social transformation function is linear.

Lastly, even if we were certain that there is an upper limit to how many individuals might exist, the total quantity of well-being might still be very large. In that case, Bounded Expected Totalism could do with a linear social transformation function, as the requirement for utilities to be bounded would already be satisfied. However, if Bounded Expected Totalism is to avoid fanatical prescriptions in cases that involve tiny probabilities of huge payoffs, then the upper and lower bounds cannot be very high or very low (respectively). So, if a very large number of individuals exist, then the transformation function must be non-linear—or Bounded Expected Totalism does not avoid Probability Fanaticism in an intuitively adequate way.

Bounded Expected Totalism would, technically, avoid Probability Fanaticism if there is an upper limit to how many individuals might exist (and individual utilities are bounded). This is because then it would not be true that, for any certain modest outcome, an arbitrarily small probability of a sufficiently great outcome is better (and similarly for negative outcomes). However, Bounded Expected Totalism would still prescribe what might be considered fanatical choices in cases that involve tiny probabilities of huge outcomes, even if there is an upper limit to how many individuals might exist. This happens because the values of those outcomes can be very high (or very low) and, thus, dominate the expected utility calculations. For example, Bounded Expected Totalism might advise Pascal to pay a too high a price to the mugger.

So, there are three reasons to adopt a non-linear social transformation function: First, in expectation, an infinite number of individuals might exist, and these possibilities must be assigned a bounded expected social utility. Secondly, arbitrarily many individuals might still exist, in which case Bounded Expected Totalism does not avoid Probability Fanaticism if the social transformation function is linear. Lastly, even if there is an upper limit to how many individuals might exist, the number of possible individuals might still be very large. In that case, Bounded Expected Totalism would prescribe fanatical choices in cases that involve tiny probabilities of huge outcomes.

Suppose that the social transformation function is non-linear. It will also have the following qualities: First, more well-being is always better, so the social transformation function must be strictly increasing with the total quantity of well-being; it must assign greater utilities to outcomes that contain more well-being. Secondly, because utilities are bounded above, similar increases in well-being must (after some point at least) matter less and less. Consequently, the social transformation function must be strictly concave on some subset of its domain. Furthermore, because utilities are also bounded below, similar increases in negative well-being must (after some point at least) matter less and less. Thus, the social transformation function must be strictly convex on some subset of its domain. Lastly, for utilities to be bounded, the social transformation function must be sufficiently concave with positive total well-being and sufficiently convex with negative total well-being; the contribution of additional (positive or negative) well-being to social utility must tend to zero.

Let f be this transformation function. Also, let $p(E_i)$ denote the probability of event E_i, $\mathrm{W}(A_i)$ the total quantity of well-being in state of affairs A_i (which results from event E_i) and $\mathrm{w}(S_{ij})$ the well-being of individual S_j in state of affairs A_i. Then, we can state Bounded Expected Totalism formally as follows:[22]

> **Bounded Expected Totalism:** For all prospects X and Y, $X \succsim Y$ if and only if $\mathrm{EU}_{\mathrm{Soc}}(X) \geq \mathrm{EU}_{\mathrm{Soc}}(Y)$, where
>
> $$\mathrm{EU}_{\mathrm{Soc}}(X) = \sum_{i=1}^{n} p(E_i) f(\mathrm{W}(A_i)) = \sum_{i=1}^{n} p(E_i) f\left(\sum_{j=1}^{m} w(S_{ij})\right).$$

Bounded Expected Totalism is the view that outcomes are ranked by their total quantity of well-being, and prospects are ranked by expected social utility, where social utility is some bounded function of the total quantity of well-being. On Bounded Expected Totalism, when calculating the value of a prospect, one first

[22]This chapter discusses what might be called *Ex-Post Bounded Expected Totalism*. However, there is another way Bounded Expected Totalism can deal with risk. This view—let's call it *Ex-Ante Bounded Expected Totalism*—first calculates the total quantity of well-being in every possible state of the world. Then, it multiplies the total quantity of well-being of each state with the probability of that state and sums these up. Finally, it transforms the expected well-being of a prospect into its expected social utility. Formally, *Ex-Ante* Bounded Expected Totalism states the following:

> ***Ex-Ante* Bounded Expected Totalism:** For all prospects X and Y, $X \succsim Y$ if and only if $\mathrm{EU}_{\mathrm{Soc}}(X) \geq \mathrm{EU}_{\mathrm{Soc}}(Y)$, where
>
> $$\mathrm{EU}_{\mathrm{Soc}}(X) = f\left(\sum_{i=1}^{n} p(E_i)\mathrm{W}(A_i)\right) = f\left(\sum_{i=1}^{m} p(E_i) w(S_i)\right).$$

Ex-Ante Bounded Expected Totalism violates Continuity. For example, let A be a St. Petersburg-style lottery (with the outcomes being total quantities of well-being), B a prospect that certainly gives a modest good outcome and C a prospect that certainly gives nothing. The expected total well-being of the mixed prospect ApC is infinite for all $p \in (0, 1)$. Thus, the expected social utility of ApC equals the upper bound of utilities, which is greater than the expected social utility of B. So, A is better than B, which is better than C, but ApC is better than B for all $p \in (0, 1)$—which is a violation of Continuity.

calculates the total quantity of well-being in every possible state of the world. Then, one transforms each state's total quantity of well-being into social utilities. Finally, to get the expected social utility of a prospect, one multiplies the social utility of each state with that state's probability and sums these up.

2.2 The cardinal structure of well-being

As mentioned above, the social transformation function takes the total quantity of well-being as input. To make sense of 'total quantity of well-being', we need well-being to have a 'cardinal structure', which allows us to make statements about *how much* more well-being an individual has in some outcome compared to another outcome.[23]

Where does this structure come from? There are two ways of deriving the cardinal structure of well-being. First, the cardinal structure of well-being might be understood in a 'primitivist' sense, according to which it can be defined independently of the individual betterness relation on gambles.[24] Alternatively, the cardinal structure of well-being might be understood in a technical sense as, for example, von Neumann-Morgenstern utilities. On the technical understanding, if the individual betterness relation satisfies a set of axioms, it can be represented by an expectational utility function.

Broome suggests that the meaning of our quantitative notion of good (i.e., well-

[23]Note that in order to talk of 'negative utilities', a cardinal structure is not sufficient; for that, well-being must have a ratio structure—which the von Neumann-Morgenstern axioms cannot deliver. Total Utilitarianism requires a meaningful zero level of well-being, which a merely interval/cardinal scale does not provide.

[24]Greaves (2015).

being) must be determined in this way. He proposes that 'utility' embodies the results of weighing good across states of nature.[25] Broome (1991, p. 147) writes: "To say that two differences in good are the same may mean nothing more than that they count the same when weighed against each other; they are evenly balanced in determining overall good. This would mean that two differences in good are the same whenever the corresponding differences in utility are the same. And that would be enough to ensure that utility is an increasing linear transform of good. Utility, then, would measure good cardinally. […] In brief, the suggestion is that our metric of good may be determined by weighing across states of nature."[26]

If von Neumann-Morgenstern utilities represent the cardinal structure of well-being, then individual betterness is, by definition, risk-neutral with respect to well-being. It might still be risk-averse with respect to money or happy years of life. But it cannot be risk-averse with respect to well-being because well-being just is the quantity whose expectation the betterness relation can be represented as maximizing. This view satisfies the following principle:[27]

> **Bernoulli's hypothesis:** One alternative is at least as good for a person as another if and only if it gives the person at least as great an

[25] Broome (1991, p. 146). Note that we need not equate utility with how much the agent values those gambles (i.e., their preferences). Utilities tell us which gambles are better and worse for a person relative to a given probability assignment, and—especially since the probability assignment at issue need not be the agent's own—this need not coincide with what the agent prefers.

[26] Broome (1991, p. 148) also concedes that we might find a metric of well-being in some other way. For example, instead of weighing up across the dimension of states of nature, he writes that this metric might be found by weighing up across a different dimension, such as the dimension of time.

[27] Broome (1991, p. 142). I have replaced 'good' with 'well-being'.

expectation of their well-being.

Bernoulli's hypothesis implies risk-neutrality about well-being.[28] It also tells us that utility represents well-being cardinally.

This chapter focuses mostly on lifetime well-being. But many of the same issues arise when we aggregate intrapersonal well-being over time.[29] Let *momentary well-being* mean how good things are for a person at some time. At least in theory, an agent can live infinitely or arbitrarily long at a given level of bliss. Therefore, for well-being/utilities to be bounded, momentary well-being must have diminishing marginal well-being/utility. Additional happy years of life must contribute less the more happy years the agent already has (and similarly for unhappy years of life).

If Bernoulli's hypothesis is false, then individual betterness might be risk-averse with respect to well-being. For example, agents might be represented as maximizing risk-weighted expected utility.[30] Alternatively, well-being could be understood

[28] Broome (1991, pp. 124 and 203).

[29] See Broome (1991, p. 226) on the *Intertemporal Addition Theorem*:

> **Intertemporal Addition Theorem:** If a person's overall betterness relation and their momentary betterness relations obey the axioms of Expected Utility Theory, and the overall betterness relation satisfies a temporal version of Pareto, then the person's overall betterness relation can be represented by an expectational utility function that is the sum of expectational utility functions representing their momentary betterness relations.

The temporal version of Pareto says that if two alternatives are equally good for a person at every time, they are equally good for them. And, if one alternative is at least as good as another for the person at every time and definitely better for them at some time, it is better for them. See Broome (1991, p. 225). The Intertemporal Addition Theorem is a variation of Harsanyi's social aggregation theorem discussed in §6 of this chapter.

[30] See for example Quiggin (1982), Buchak (2013) and Buchak (2017). Risk-weighted expected utility theory (a member of rank-dependent theories) does not avoid fanatical prescriptions in the prudential case unless individual utilities are bounded. Similarly, (impersonal) risk-weighted expected utility theory does not avoid Probability Fanaticism unless social utilities are bounded. See

in a primitivist sense. The primitivist view requires that quantities of well-being have meaning independently of how much they count when evaluating uncertain prospects.[31] But if such a metric of well-being is available, then individual betterness might be risk-averse with respect to this (non-technical) well-being. Note that this view is compatible with Expected Utility Theory (but not with Bernoulli's hypothesis).

Let an *agent's transformation function* be a function that takes that person's well-being levels as input and outputs their individual utilities (to be used in decision-making under risk). If individual betterness over prospects is sufficiently risk-averse with respect to well-being, such that the agent's transformation function approaches asymptotically some upper bound with more well-being, then well-being itself can be unbounded without leading to unbounded utilities.

Finally, individual betterness might be risk-neutral with respect to well-being. And, happy days of life might not contribute less to well-being the more happy days the agent already has (and similarly for unhappy days). Given that individuals might live arbitrarily long at a constant positive well-being level, this view implies that both well-being and utilities are unbounded. This leads to a prudential analogue of Probability Fanaticism:

Prudential Fanaticism:

i *Positive Prudential Fanaticism* For any probability $p > 0$, and for any finite individual utility u, there is some large enough in-

Monton (2019, §5.7) and Beckstead & Thomas (2020, p. 12).

[31] Broome (1991, p. 217).

dividual utility U such that probability p of U (and otherwise nothing) is prudentially better than the certainty of u for some individual S.

ii *Negative Prudential Fanaticism* For any probability $p > 0$, and for any finite negative individual utility $-u$, there is some large enough negative individual utility $-U$ such that probability p of $-U$ (and otherwise nothing) is prudentially worse than the certainty of $-u$ for some individual S.

To summarize, social utilities might be bounded if the total quantity of well-being is itself necessarily bounded. However, this is not true; therefore, Bounded Expected Totalism requires a social transformation function that takes the total quantity of well-being as input and outputs social utilities. To recap, this social transformation function must be non-linear for three reasons: First, in expectation, an infinite number of individuals might exist, so the total quantity of well-being might be infinite in expectation. But Bounded Expected Totalism requires expected social utilities to be bounded. Secondly, arbitrarily many individuals might exist. In that case, the social transformation function must be non-linear or Bounded Expected Totalism does not avoid Probability Fanaticism. Lastly, even if there is an upper limit to how many individuals might exist, the number of individuals might still be very large. In that case, the social transformation function must be non-linear or Bounded Expected Totalism prescribes fanatical choices in cases that involve tiny probabilities of huge outcomes.

The social transformation function uses the 'total quantity of well-being' as input. To make sense of this notion, well-being must have a cardinal structure. This structure could be primitive, that is, given independently of individual betterness relation on gambles. Alternatively, it could be defined using Bernoulli's hypothesis. If the cardinal structure is defined using Bernoulli's hypothesis, then individual betterness is risk-neutral. But if it is primitive, or defined in some other way, then it is at least initially an open question whether individual betterness is risk-neutral, risk-averse, or what, with respect to well-being. Next, I will show that Bounded Expected Totalism violates Ex Ante Pareto if individual betterness is risk-neutral with respect to well-being. §4 shows that Bounded Expected Totalism violates Ex Ante Pareto if individual betterness is risk-averse with respect to well-being.

3 The risk-neutral case

This section shows that Bounded Expected Totalism violates Ex Ante Pareto if individual betterness is risk-neutral with respect to well-being.

Let well-being levels be represented by real numbers. As argued above, the social transformation function f must be strictly concave on some subset of its domain. For the sake of argument, let's suppose it is strictly concave at 1. Then, there must be some positive constants δ and ϵ such that $f(1) - f(1-\delta) > f(1+\delta+\epsilon) - f(1)$. This is because the smaller benefit (δ) contributes more when added to a population at a lower well-being level than the greater benefit ($\delta + \epsilon$) when added to a population at a higher well-being level.

Next, consider the following prospects:

The Risk-Neutral Case:

Risky Gives a 0.5 probability of a well-being level of $1 + \delta + \epsilon$; otherwise, it gives a well-being level of $1 - \delta$.

Safe Surely gives a well-being level of 1.

Suppose that the betterness relation of some agent, Alice, is risk-neutral with respect to her well-being. Then, Risky is better than Safe for Alice, as Risky gives a higher expectation of well-being than Safe does.

But is Risky also better than Safe impersonally? The answer is no. Given that the constants δ and ϵ are such that $f(1) - f(1 - \delta) > f(1 + \delta + \epsilon) - f(1)$, Safe is impersonally better than Risky (even though Risky is still better than Safe for Alice, given that it gives a higher expectation of her well-being for all positive values of δ and ϵ). The situation is illustrated by the following graph:[32]

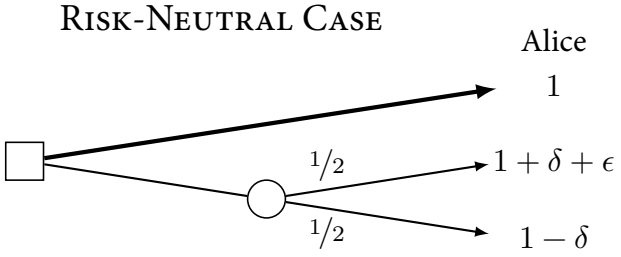
[32]Gustafsson (2022) presents this case to illustrate that *Ex-Post* Prioritarianism violates Ex Ante Pareto, a fact that goes back at least to Rabinowicz (2002). For an overview of this topic, see for example Fleurbaey (2018). See also Broome (1991, Ch. 9). Bounded Expected Totalism coincides with *Ex-Post* Prioritarianism in one-person cases. So, we can appeal to the standard fact that *Ex-Post* Prioritarianism violates Ex Ante Pareto. *Ex-Post* Prioritarianism states the following:

> ***Ex-Post* Prioritarianism:** For all prospects X and Y, $X \succsim Y$ if and only if $\mathrm{EU}_{\mathrm{Soc}}(X) \geq \mathrm{EU}_{\mathrm{Soc}}(Y)$, where
>
> $$\mathrm{EU}_{\mathrm{Soc}}(X) = \sum_{i=1}^{n} p(E_i) \left(\sum_{j=1}^{m} f\left(w(S_{ij})\right) \right).$$

Ex-Post Bounded Expected Totalism differs from *Ex-Post* Prioritarianism because it first sums up everyone's well-being and then converts this sum into social utilities. In contrast, the latter view first converts individuals' well-being levels and then sums up the converted well-being levels. *Ex-Post* Bounded Expected Totalism applies the transformation function to the total quantity of well-being; *Ex-Post* Prioritarianism applies it to the well-being of individuals. On *Ex-Post* Prioritarianism, so-

A Violation of Ex Ante Pareto:

Risk-Neutral Case



Here, the square represents a choice node, while the circle represents a chance node. Going up at the choice node means accepting Safe, and going down at the choice node means accepting Risky. Thus, if we go up, Alice gets a well-being level of 1. On the other hand, if we go down, there are two possible states of the world, each with a 0.5 probability. In state 1, Alice gets a well-being level of $1 + \delta + \epsilon$. And, in state 2, Alice gets a well-being level of $1 - \delta$.

The expected social utility of going up is $\text{EU}_{\text{Soc}}(\text{Safe}) = f(1)$. And, the expected social utility of going down is $\text{EU}_{\text{Soc}}(\text{Risky}) = \frac{1}{2} \cdot f(1+\delta+\epsilon) + \frac{1}{2} \cdot f(1-\delta)$. Given that $f(1) - f(1-\delta) > f(1+\delta+\epsilon) - f(1)$, $\text{EU}_{\text{Soc}}(\text{Risky})$ is less than $\text{EU}_{\text{Soc}}(\text{Safe})$.[33] Thus, going up is impersonally better than going down, according to Bounded Expected Totalism. However, going down is better than going up for Alice (and equally good for everybody else). Thus, we have a violation of Ex Ante

cial utilities are unbounded because the sum of converted well-being levels can be arbitrarily high, given that arbitrarily many individuals might exist. On *Ex-Post* Bounded Expected Totalism, social utilities are bounded because, although the sum of everyone's well-being can be arbitrarily high, the total quantity of well-being has diminishing marginal utility. Similarly, *Ex-Ante* Bounded Expected Totalism differs from *Ex-Ante* Prioritarianism because the former applies the transformation function to the total expected well-being of a prospect, while the latter applies it to the expected well-being of individuals.

[33] By rearranging $f(1) - f(1-\delta) > f(1+\delta+\epsilon) - f(1)$, we get $f(1) + f(1) > f(1+\delta+\epsilon) + f(1-\delta)$. Next, by dividing both sides by 2, we get $f(1) > \frac{1}{2} \cdot f(1+\delta+\epsilon) + \frac{1}{2} \cdot f(1-\delta)$.

Pareto.[34]

To summarize, Bounded Expected Totalism violates Ex Ante Pareto if individual betterness is risk-neutral with respect to well-being. This happens because the social transformation function is concave on some subset of its domain.[35] Consequently, Bounded Expected Totalism is at least sometimes risk-averse with respect to (positive) well-being.

4 The risk-averse case

This section shows that Bounded Expected Totalism violates Ex Ante Pareto even if individual betterness is risk-averse with respect to well-being.[36]

[34]If individual utilities are unbounded above while social utilities are bounded above, then Bounded Expected Totalism violates Ex Ante Pareto in the following case as well:

> **Unbounded individual utilities:**
>
> *Risky** Gives a tiny probability p of a very high positive well-being level w_1 (and otherwise nothing).
>
> *Safe** Surely gives a modest positive well-being level w_2.

Suppose individuals maximize unbounded expected utility, but social utilities are bounded. Then, with some values of p, w_1 and w_2, Risky* is better than Safe* for individuals, but Safe* is impersonally better than Risky*. This is a violation of Ex Ante Pareto. This happens because, in the impersonal case, the additional well-being in w_1 is insufficient to compensate for the tiny probability of obtaining it; however, for individual agents, it is sufficient. Bounded Expected Totalism violates Ex Ante Pareto in a similar case (changing what needs to be changed) if individual utilities are unbounded below while social utilities are bounded below.

[35]The same argument can be applied, changing what needs to be changed, as long as the social transformation function is concave on some subset of its domain—it need not be concave specifically at 1.

[36]It is already known that individual risk attitudes incompatible with Expected Utility Theory can cause tensions with Ex Ante Pareto. See for example Nebel (2020) and Mongin & Pivato (2015). However, the violation of Ex Ante Pareto discussed in this section happens even if the risk aversion is of the kind that is compatible with Expected Utility Theory.

If individual betterness is risk-averse with respect to well-being, then it may no longer be true that Risky is better than Safe for Alice. So, Bounded Expected Totalism might not violate Ex Ante Pareto in the way discussed earlier. If Alice's transformation function corresponds to the social transformation function when Alice is the only person who exists, then Risky is at least as good as Safe for Alice if and only if Risky is at least as good as Safe impersonally (and vice versa). So, Bounded Expected Totalism avoids violating Ex Ante Pareto in the earlier case.

However, how much Alice's well-being contributes to social utility depends on how many individuals exist and what their well-being levels are. The greater the total quantity of well-being, the smaller the contribution of additional well-being is. Suppose that, when Alice is the only person who exists, Alice's loss of δ would reduce social utility by x units, and her gain of $\delta + \epsilon$ would increase it by more than x units. Then, in the one-person case, Risky is better than Safe (both impersonally and, by Ex Ante Pareto, for Alice).[37]

Now change the case; suppose that, besides Alice, there is a large number N of other, unaffected people.

> **Alice and Others:** A large number N of other people have very good lives in state 1 ($p = 0.5$) and neutral lives in state 2 ($p = 0.5$).

[37] Note that this step requires the following version of Ex Ante Pareto:

> **Weak Ex Ante Pareto:** For all prospects X and Y, if X is at least as good as Y for everyone, then X is at least as good as Y.

Also, this step assumes Completeness. Without Completeness, Weak Ex Ante Pareto does not entail that Risky must be better than Safe for Alice if Risky is better than Safe impersonally—they could be incomparable for her.

Risky Gives Alice a well-being level of $1 + \delta + \epsilon$ in state 1 and a well-being level of $1 - \delta$ in state 2.

Safe Gives Alice a well-being level of 1 in states 1 and 2.

In the state where Alice would lose δ (state 2), the other people have neutral lives (i.e., lives whose addition does not increase or decrease the total quantity of well-being). It follows that, no matter how large N is, her loss of δ would still reduce social utility in that state by x units. On the other hand, in the state where Alice would win $\delta + \epsilon$ (state 1), the N people have very good lives. Let α denote the total quantity of well-being of the N people with very good lives. As we increase N, the social utility in state 1 approaches the upper limit of utilities until it comes within x units of the upper limit. Then, increasing Alice's well-being by $\delta + \epsilon$ contributes less than x to social utility in that state. So, the $\delta + \epsilon$ increase in Alice's well-being in state 1 is no longer sufficient to compensate for the possible loss of δ well-being (and x units of utility) in state 2. It follows that, with a sufficiently large N, Safe is impersonally better than Risky. This contradicts Ex Ante Pareto since Risky is better than Safe for Alice, and Safe and Risky are equally good for each of the N additional people.

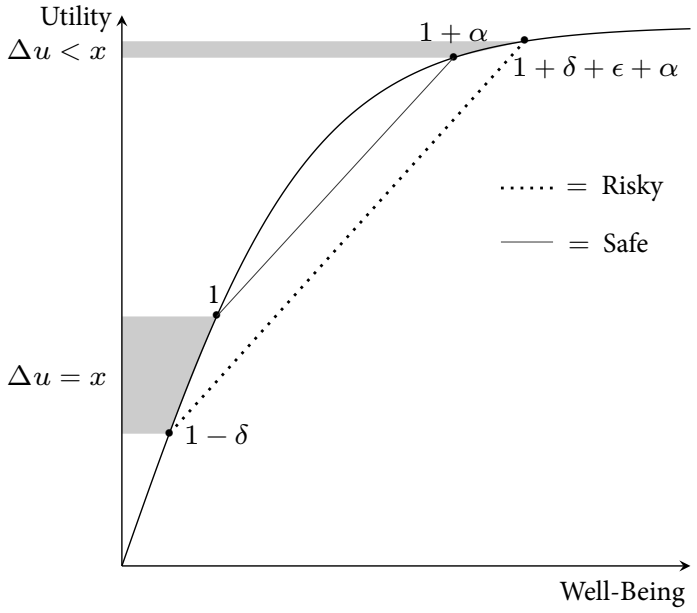
TABLE 1

ALICE AND OTHERS

	State 1	**State 2**
p	0.5	0.5
Risky	Alice: $1 + \delta + \epsilon$ Others: α	Alice: $1 - \delta$ Others: 0
Safe	Alice: 1 Others: α	Alice: 1 Others: 0

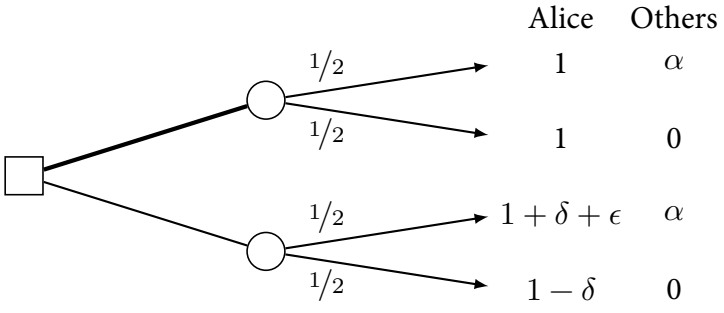
As can be seen from the graph below, Alice's loss of δ (with Risky) reduces social utility by x units, from $u(1)$ to $u(1 - \delta)$. However, her gain of $\delta + \epsilon$ (with Risky) increases social utility by less than x units, from $u(1 + \alpha)$ to $u(1 + \delta + \epsilon + \alpha)$. Thus, Safe is impersonally better than Risky. But we have assumed that Alice's gain of $\delta + \epsilon$ increases her own utility by more than x units. So, Risky is better than Safe for Alice, and we have a violation of Ex Ante Pareto.

Alice and Others



A Violation of Ex Ante Pareto:

Risk-Averse Case



To summarize, Bounded Expected Totalism violates Ex Ante Pareto even if individual betterness is risk-averse with respect to well-being.

5 Bounded above and below

This section gives a general argument for why Bounded Expected Totalism must violate Ex Ante Pareto if social utilities are bounded above and below. This argument shows that a violation of Ex Ante Pareto must happen regardless of whether individual utilities are bounded or unbounded and whether individual betterness is risk-neutral, risk-averse or risk-seeking. But first, to introduce some background, I will discuss a case that shows how Bounded Expected Totalism violates Ex Ante Pareto if individual well-being is unbounded and both individual and social utilities are bounded above and below.

5.1 Unbounded individual well-being

Assuming that overall betterness can be represented with an expectational utility function, social utilities must be bounded above and below in order to avoid both Positive and Negative Probability Fanaticism. Similarly, to avoid Positive and Negative Probability Fanaticism in the prudential case, individual utilities must be bounded above and below. This will lead to a violation of Ex Ante Pareto if individual well-being is unbounded. Consider the following prospects:

Risky vs. Safe**:**

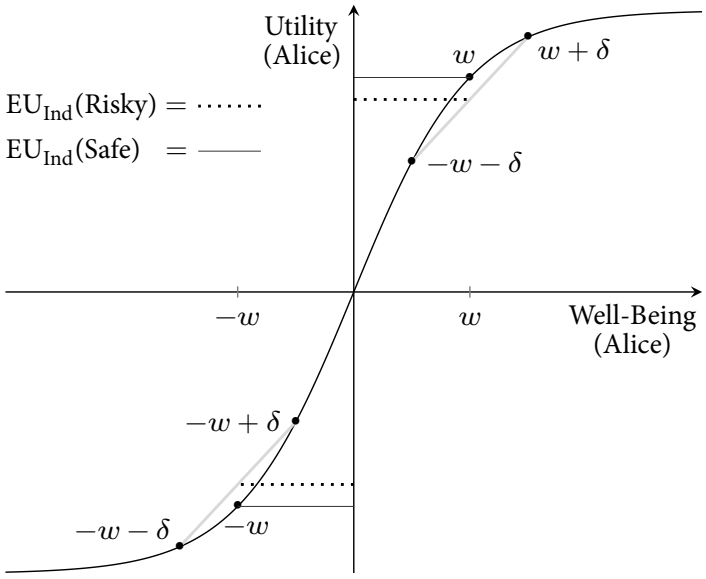
*Risky*** Gives a 0.5 probability of δ additional well-being; otherwise, it decreases well-being by $-\delta$.

*Safe*** Does not increase or decrease well-being.

As explained before, if utilities are bounded above, then (at least at some point) Alice's transformation function is concave with positive well-being; additional well-being matters less the happier Alice already is. This means that, at least sometimes, Alice's betterness relation is risk-averse with respect to her well-being. On the other hand, if utilities are bounded below, then (at least at some point) Alice's transformation function is convex with negative well-being; additional unhappiness matters less the unhappier Alice already is. This means that, at least sometimes, Alice's betterness relation is risk-seeking with respect to her ill-being.

In expectation, neither Risky** nor Safe** affects Alice's well-being. So, which of Risky** and Safe** is better for Alice can depend on whether Alice is overall happy or unhappy (see the graph below). With some positive background well-being level w, Safe** is better than Risky** for Alice. In contrast, with some negative background well-being level $-w$, Risky** is better than Safe** for Alice.

Risky** vs. Safe**



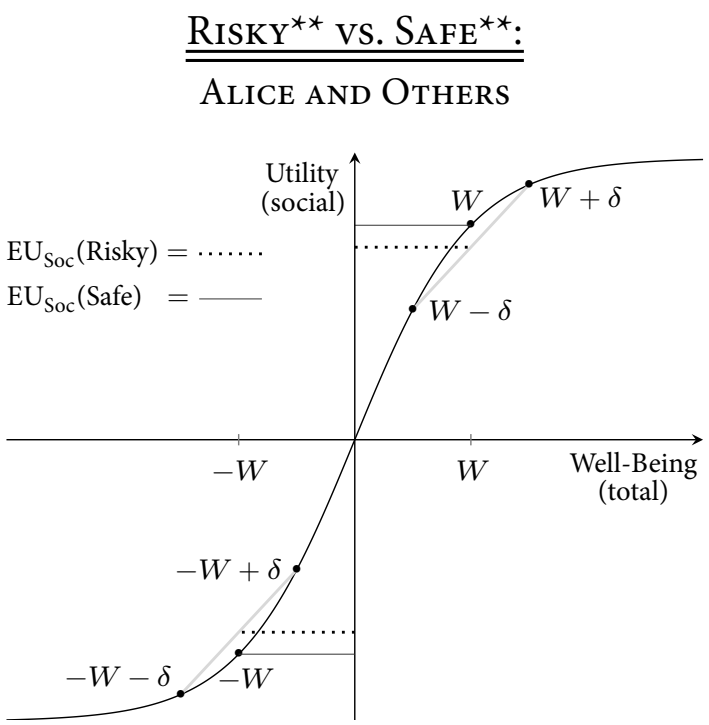
Next, to avoid Positive and Negative Probability Fanaticism, social utilities must also be bounded above and below. If social utilities are bounded above, then (at least at some point) the social transformation function is concave with a positive total quantity of well-being. This means that, at least sometimes, the overall betterness relation is risk-averse with respect to well-being. On the other hand, if social utilities are bounded below, then (at least at some point) the social transformation function is convex with a negative total quantity of well-being. This means that, at least sometimes, the overall betterness relation is risk-seeking with respect to well-being. So, with some positive total quantity of well-being W, Safe** is impersonally better than Risky**. On the other hand, with some negative total quantity of well-being $-W$, Risky** is impersonally better than Safe**.

But now consider the following cases:

Alice and Others*: Alice's well-being will increase, decrease or stay the same depending on the choice and result of Risky** and Safe** (and nobody else is affected).

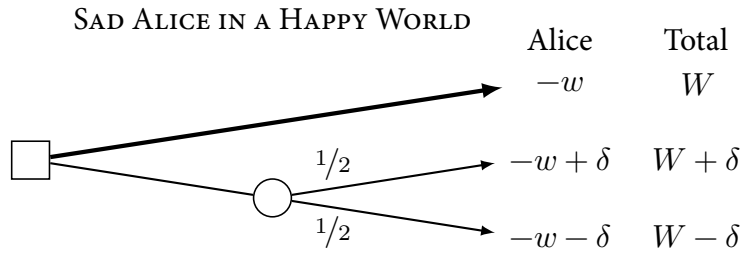
Sad Alice in a happy world: Alice has a baseline well-being level of $-w$. The total quantity of well-being in the world is W (includes Alice's well-being).

Happy Alice in a sad world: Alice has a baseline well-being level of w. The total quantity of well-being in the world is $-W$.

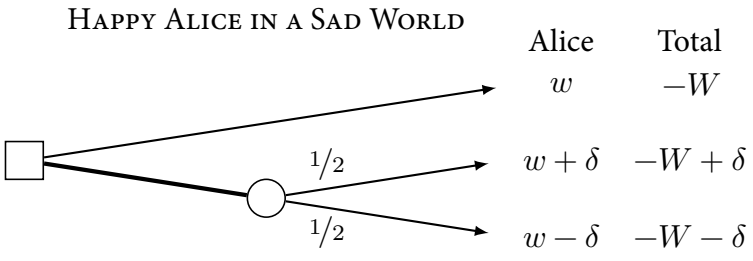


When Alice's baseline well-being level is $-w$, but the total quantity of well-being in the world is W, Safe** is impersonally better than Risky**, but Risky** is better than Safe** for Alice. On the other hand, when Alice's baseline well-being level is w but the total quantity of well-being in the world is $-W$, Risky** is impersonally better than Safe**, but the reverse is true for Alice. Consequently, Bounded Expected Totalism violates Ex Ante Pareto if both individual and social utilities are bounded above and below but individual well-being is unbounded.

A Violation of Ex Ante Pareto:



A Violation of Ex Ante Pareto:



5.2 The general argument

Next, I will give a general argument which shows that Bounded Expected Totalism must violate Ex Ante Pareto, regardless of whether individual utilities are bounded

or unbounded and whether individual betterness is risk-averse, risk-neutral or risk-seeking. The general argument goes as follows: Fix any $\delta > 0$. The following two claims are true:

(1) If Risky** is impersonally at least as good as Safe** no matter how much total well-being there is in the background population, then social utility is unbounded above.

(2) If Safe** is impersonally at least as good as Risky** no matter how much total well-being there is in the background population, then social utility is unbounded below.

If social utility is bounded above and below, there must be a counterexample to Ex Ante Pareto. Suppose, for example, that Risky** is at least as good as Safe** for Alice. This could be because Alice's betterness relation is risk-neutral with respect to her well-being and Risky** is therefore equally as good as Safe** for Alice. Alternatively, Alice's betterness relation might be risk-seeking. Either way, if social utilities are bounded above, then (1) shows that Risky** cannot be impersonally at least as good as Safe** no matter how much total well-being there is in the background population. So, with some total quantity of well-being, Safe** is impersonally better than Risky**—which contradicts Ex Ante Pareto.

Similarly, suppose that Safe** is at least as good as Risky** for Alice. Again, this might be because Alice's betterness relation is risk-neutral with respect to her well-being. Alternatively, it could be because her betterness relation is risk-averse. However, given that social utilities are bounded below, (2) shows that Safe** cannot

be impersonally at least as good as Risky** no matter how much total well-being there is in the background population. So, with some total quantity of well-being, Risky** is impersonally better than Safe**, contrary to Ex Ante Pareto.

Proof of (1) goes as follows: Consider background populations with total well-being levels of 0, δ, 2δ, 3δ, and so on. Let $x = f(\delta) - f(0)$. If Risky** is impersonally at least as good as Safe** with respect to all these background populations, then the difference between $f(n\delta)$ and $f((n-1)\delta)$ is at least as great as the difference between $f((n-1)\delta)$ and $f((n-2)\delta)$, for each $n > 2$. It follows that $f(n\delta)$ is at least as great as nx. Thus, f is unbounded above. One can give a similar proof for (2). So, if social utilities are bounded above and below, there must be a counterexample to Ex Ante Pareto, regardless of whether individual utilities are bounded or unbounded and whether individual betterness is risk-averse, risk-neutral or risk-seeking.

To summarize, this section first showed that Bounded Expected Totalism violates Ex Ante Pareto if both individual and social utilities are bounded above and below. Next, this section presented a general proof to the effect that Bounded Expected Totalism must violate Ex Ante Pareto regardless of whether individual betterness is risk-neutral, risk-averse or risk-seeking and whether individual utilities are bounded or unbounded.

6 Harsanyi's social aggregation theorem

This section discusses how the earlier examples relate to a famous result in this area, namely, Harsanyi's social aggregation theorem.

Harsanyi's social aggregation theorem shows that if both individual and social betterness relations can be given an expected utility representation, and the overall betterness relation satisfies Ex Ante Pareto, then social utilities are weighted sums of individual utilities.[38] Let me explain Harsanyi's premises in more detail. The first premise says that each individual's betterness relation obeys the von Neumann-Morgenstern axioms.[39] So, the individual betterness relation can be represented by an expectational utility function. The second premise says that the overall betterness relation obeys the von Neumann-Morgenstern axioms. So, overall betterness can also be represented by an expectational utility function. The third premise is Ex Ante Pareto.[40] The conclusion of Harsanyi's theorem is that social utilities are

[38] Harsanyi (1955). Harsanyi (1955) uses individual utilities to describe individual preferences. But we may reinterpret them as describing individual betterness instead of individual preferences. See Broome (1991).

[39] Harsanyi (1955) uses Marschak's (1950) versions of the von Neumann & Morgenstern (1947) axioms. Marschak's (1950, p. 117) Postulate II states:

> **Postulate II (Continuity):** If $X \succ Y \succ Z$, then there is a probability $p \in (0, 1)$ such that $Y \sim XpZ$.

This postulate implies, in a similar way as shown before, that utilities must be bounded.

[40] Harsanyi (1955) uses Pareto Indifference in the original formulation of the theorem, while Harsanyi (1977, p. 65) uses Weak Ex Ante Pareto:

> **Pareto Indifference:** For all prospects X and Y, if X and Y are equally good for everyone, then X and Y are overall equally good.
>
> **Weak Ex Ante Pareto:** For all prospects X and Y, if X is at least as good as Y for everyone, then X is overall at least as good as Y.

weighted sums of individual utilities. Thus, overall betterness can be represented as maximizing the expectation of a weighted sum of individual utilities. If, in addition, we assume equal weighting for all individuals, then this theorem shows that the social utility function must be a sum of individual utilities.[41]

Harsanyi's theorem shows, in other words, that if individual and overall betterness relations are represented by expectational utility functions, then in order to satisfy Ex Ante Pareto, the social utility function must be a linear combination of individual utilities. Earlier in this chapter, I showed that Total Utilitarianism combined with Bounded Expected Utility Theory violates Ex Ante Pareto.[42] There-

---

Using Weak Ex Ante Pareto instead of Pareto Indifference guarantees that positive individual well-being contributes *non-negatively* to social utilities. Using Ex Ante Pareto instead of Weak Ex Ante Pareto guarantees that positive individual well-being contributes *positively* to social utilities. See Weymark (1994) on Harsanyi's theorem with different Pareto principles.

[41] Broome (1991, §10) argues that Harsanyi's social aggregation theorem, together with Bernoulli's hypothesis, leads to utilitarianism.

[42] As mentioned in footnote 40, the original argument by Harsanyi (1955) uses Pareto Indifference instead of Ex Ante Pareto. Bounded Expected Totalism also violates this condition if individual betterness satisfies the von Neumann-Morgenstern axioms. Consider the following prospects:

> **Alice and Bob:**
>
> *Risky* Gives Alice and Bob a well-being level of 1 in state 1 ($p = 0.5$) and a well-being level of 0 in state 2 ($p = 0.5$).
>
> *Safe* In state 1 ($p = 0.5$), Alice gets a well-being level of 1 and Bob a well-being level of 0. In state 2 ($p = 0.5$), Alice gets a well-being level of 0 and Bob a well-being level of 1.

Risky and Safe are equally good for both Alice and Bob. So, by Pareto Indifference, Risky and Safe are impersonally equally good. Next, recall that the social transformation function must be strictly concave on some subset of its domain if social utilities are bounded above (for the reasons discussed in §2.1). We may suppose it is strictly concave on the interval [0, 2]. Consequently, Safe is impersonally better than Risky. This violation of Pareto Indifference happens because when the social transformation function is strictly concave, it is impersonally better to spread the total quantity of well-being across different states than to have it all in one state. But individual betterness is indifferent to how the well-being of different individuals is spread across states.

fore, if one accepts Bounded Expected Totalism, that premise of Harsanyi's theorem fails. The reason that led to its failure was that a non-linear social transformation function is needed because the number of individuals might be infinite or arbitrarily large. In fact, it is unsurprising that one of Harsanyi's premises must be rejected; if the number of individuals might be infinite or arbitrarily large, then social utilities cannot be weighted sums of individual utilities because this could lead to unbounded social utilities.[43,44] So, given that a bounded expected totalist rejects Harsanyi's conclusion, they cannot accept all his premises.

This is worrying because Harsanyi's theorem is often considered one of the best arguments for utilitarianism. The conclusion of Harsanyi's theorem is that, for any fixed and finite population, social utility is an affine (or linear) function of total individual utility. However, once we consider the possibility of an infinite or arbitrarily large population, we find that social utility must be non-linear if social utilities are bounded and additive with individual utilities.[45] And this leads to violations of Ex Ante Pareto.

All this can be taken to support *Average Utilitarianism*, namely, the view that one population is better than another if and only if the average well-being it contains is greater.[46] Alternatively, these cases might be taken to undermine Bounded-

[43] See Blackorby et al. (2007) for an extension of Harsanyi's social aggregation theorem to variable populations.

[44] As mentioned earlier, this need not be true. See footnote 18 on p. 70.

[45] Harsanyi (1977, p. 60) himself discusses what he calls the 'boundary problem for the society', namely, whose utility functions ought to be included in our social-welfare function. He considers whether to include, for example, higher animals, distant future generations, robots or the inhabitants of other planets. However, he does not mention the possibility that doing so might lead to infinite or arbitrarily large populations.

[46] Average Utilitarianism does not require a non-linear social transformation function; if indi-

ness (and Continuity). One might accept, for example, *Unbounded Expected Totalism*, namely, the view that combines Total Utilitarianism and Expected Utility Theory with an unbounded utility function. However, this view cannot be supported by a version of Harsanyi's theorem that relies on the von Neumann-Morgenstern axiomatization of Expected Utility Theory, as this axiomatization has Continuity as one of its axioms. But one might attempt to justify Unbounded Expected Totalism with a Harsanyi-style argument that does not rely on Continuity.[47] Finally, as mentioned earlier, the arguments in this chapter might be taken to support Probability Discounting indirectly. As I will explain in Chapter 3 of this thesis, Probability Discounting also leads to violations of Ex Ante Pareto.[48] But given that both theories violate Ex Ante Pareto, the plausibility of Ex Ante Pareto does not favor Bounded Expected Totalism over Probability Discounting.

7 Conclusion

This chapter has shown that Bounded Expected Totalism violates Ex Ante Pareto. Separate examples of Ex Ante Pareto violations were given for risk-neutrality and

vidual utilities are bounded, then the average of those must also be bounded. So, Average Utilitarianism avoids violating Ex Ante Pareto. However, Average Utilitarianism has other implausible implications, such as the *Sadistic Conclusion* (Arrhenius 2000, p. 251):

> **The Sadistic Conclusion:** When adding people without affecting the original people's welfare, it can be better to add people with negative well-being rather than positive well-being.

[47] Fleurbaey (2009) gives such an argument using statewise dominance and anonymity instead of the von Neumann-Morgenstern axioms. Relatedly, McCarthy et al. (2020) show that one can argue for Expected Utility Theory with an unbounded utility function from Pareto and anonymity.

[48] See also Kosonen (2021).

risk-aversion. A general argument to the effect that Bounded Expected Totalism must violate Ex Ante Pareto was also given. Lastly, the implications of these cases for Harsanyi's social aggregation theorem were discussed.

The violations of Ex Ante Pareto happen because there is a non-zero probability that an infinite or arbitrarily large number of individuals exist. These Ex Ante Pareto violations also happen if one wishes to avoid Probability Fanaticism. Since Bounded Expected Totalism cannot avoid Probability Fanaticism without violating Ex Ante Pareto, these violations of Ex Ante Pareto undermine the plausibility of Bounded Expected Totalism as an alternative to Probability Fanaticism.

To conclude, combining Total Utilitarianism and Expected Utility Theory with a bounded utility function results in violations of Ex Ante Pareto: The combination of these views implies that a prospect can be impersonally better than another prospect even though it is worse for everyone who is affected by the choice.

References

Arrhenius, G. (2000), 'An impossibility theorem for welfarist axiologies', *Economics and Philosophy* **16**(2), 247–266.

Baumann, P. (2009), 'Counting on numbers', *Analysis* **69**(3), 446–448.

Beckstead, N. (2013), On the overwhelming importance of shaping the far future, PhD thesis, Rutgers, the State University of New Jersey.

Beckstead, N. & Thomas, T. (2020), 'A paradox for tiny probabilities and enormous

values'. Global Priorities Institute Working Paper No.10.

URL: *https://globalprioritiesinstitute.org/nick-beckstead-and-teruji-thomas-a-paradox-for-tiny-probabilities-and-enormous-values/*

Blackorby, C., Bossert, W. & Donaldson, D. (2007), 'Variable-population extensions of social aggregation theorems', *Social Choice and Welfare* **28**(4), 567–589.

Bostrom, N. (2009), 'Pascal's Mugging', *Analysis* **69**(3), 443–445.

Bostrom, N. (2011), 'Infinite ethics', *Analysis and Metaphysics* **10**, 9–59.

Branwen, G. (2009), 'Notes on Pascal's Mugging'.

URL: *https://www.gwern.net/mugging*

Broome, J. (1991), *Weighing Goods: Equality, Uncertainty and Time*, Blackwell, Oxford.

Buchak, L. (2013), *Risk and Rationality*, Oxford University Press, Oxford.

Buchak, L. (2017), 'Precis of *Risk and Rationality*', *Philosophical Studies* **174**(9), 2363–2368.

Fishburn, P. C. (1970), *Utility Theory for Decision Making*, Wiley, New York.

Fleurbaey, M. (2009), 'Two variants of Harsanyi's aggregation theorem', *Economics Letters* **105**(3), 300–302.

Fleurbaey, M. (2018), 'Welfare economics, risk and uncertainty', *Canadian Journal of Economics* **51**(1), 5–40.

Goodsell, Z. (2021), 'A St Petersburg Paradox for risky welfare aggregation', *Analysis* **81**(3), 420–426.

Greaves, H. (2015), 'Antiprioritarianism', *Utilitas* **27**(1), 1–42.

Gustafsson, J. E. (2022), 'Ex-ante prioritarianism violates sequential ex-ante Pareto', *Utilitas* **34**(2), 167–177.

Gustafsson, J. E. (forthcoming), *Money-Pump Arguments*, Cambridge University Press, Cambridge.

Hammond, P. J. (1998), Objective expected utility: A consequentialist perspective, *in* S. Barberà, P. J. Hammond & C. Seidl, eds, 'Handbook of Utility Theory Volume 1: Principles', Kluwer, Dordrecht, pp. 143–211.

Hanson, R. (2007), 'Pascal's Mugging: Tiny probabilities of vast utilities'.
URL: *https://www.lesswrong.com/posts/a5JAiTdytou3Jg749/pascal-s-mugging-tiny-probabilities-of-vast-utilities?commentId=Q4ACkdYFEThA6EE9P*

Harsanyi, J. C. (1955), 'Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility', *Journal of Political Economy* **63**(4), 309–321.

Harsanyi, J. C. (1977), *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*, Cambridge University Press, Cambridge.

Jensen, N. E. (1967), 'An introduction to Bernoullian utility theory: I. Utility functions', *The Swedish Journal of Economics* **69**(3), 163–183.

Kosonen, P. (2021), 'Discounting small probabilities solves the Intrapersonal Addition Paradox', *Ethics* **132**(1), 204–217.

Kreps, D. M. (1988), *Notes on the Theory of Choice*, Westview Press, Boulder.

Marschak, J. (1950), 'Rational behavior, uncertain prospects, and measurable utility', *Econometrica* **18**(2), 111–141.

McCarthy, D., Mikkola, K. & Thomas, T. (2020), 'Utilitarianism with and without expected utility', *Journal of Mathematical Economics* **87**, 77–113.

Mongin, P. & Pivato, M. (2015), 'Ranking multidimensional alternatives and uncertain prospects', *Journal of Economic Theory* **157**, 146–171.

Monton, B. (2019), 'How to avoid maximizing expected utility', *Philosophers' Imprint* **19**(18), 1–24.

Nebel, J. M. (2020), 'Rank-weighted utilitarianism and the veil of ignorance', *Ethics* **131**(1), 87–106.

Quiggin, J. (1982), 'A theory of anticipated utility', *Journal of Economic Behavior and Organization* **3**(4), 323–343.

Rabinowicz, W. (2002), 'Prioritarianism for prospects', *Utilitas* **14**(1), 2–21.

Russell, J. S. (2021), 'On two arguments for fanaticism'. Global Priorities Institute Working Paper 17–2021.

URL: *https://globalprioritiesinstitute.org/on-two-arguments-for-fanaticism-jeff-sanford-russell-university-of-southern-california/*

Russell, J. S. & Isaacs, Y. (2021), 'Infinite prospects', *Philosophy and Phenomenological Research* **103**(1), 178–198.

von Neumann, J. & Morgenstern, O. (1947), *Theory of Games and Economic Behavior*, 2 edn, Princeton University Press, Princeton.

Weymark, J. A. (1994), Harsanyi's social aggregation theorem with alternative Pareto principles, *in* W. Eichhorn, ed., 'Models and Measurement of Welfare and Inequality', Springer, Berlin, Heidelberg, pp. 869–887.

Wilkinson, H. (2022), 'In defence of fanaticism', *Ethics* **132**(2), 445–477.

Yudkowsky, E. (2007*a*), 'A comment on Pascal's Mugging: Tiny probabilities of vast utilities'.

URL: *https://www.lesswrong.com/posts/a5JAiTdytou3Jg749/pascal-s-mugging-tiny-probabilities-of-vast-utilities?commentId=kqAKXskjohx4SSyp4*

Yudkowsky, E. (2007*b*), 'Pascal's Mugging: Tiny probabilities of vast utilities'.

URL: *http://www.overcomingbias.com/2007/10/pascals-mugging.html*

Zuber, S., Venkatesh, N., Tännsjö, T., Tarsney, C., Stefánsson, H. O., Steele, K., Spears, D., Sebo, J., Pivato, M., Ord, T., Ng, Y.-K., Masny, M., MacAskill, W., Kuruc, K., Hutchinson, M., Gustafsson, J. E., Greaves, H., Forsberg, L., Fleurbaey, M., Coffey, D., Cato, S., Castro, C., Campbell, T., Budolfson, M., Broome, J., Berger, A., Beckstead, N. & Asheim, G. B. (2021), 'What should we agree on about the Repugnant Conclusion?', *Utilitas* **33**(4), 379–383.

Chapter 2

*Expected Utility Theory and Possible States of Zero Probability**

ABSTRACT: At least at first glance, Expected Utility Theory tells us to be indifferent between two prospects when they are otherwise the same, except that one gives a better outcome than the other in a possible state of zero probability. But as some have suggested, Expected Utility Theory might be supplemented with dominance to get the verdict that the dominating prospect is better than the dominated one. However, I will show that if Expected Utility Theory is supplemented with dominance in this way, it will violate the Continuity axiom of Expected Utility Theory.

Consider the following principle of rationality:

> **Statewise Dominance:** If the outcome of prospect X is at least as preferred as the outcome of prospect Y in all states, then X is at least as good as Y. Furthermore, if in addition the outcome of X is strictly

*I wish to thank Tomi Francis, Andreas Mogensen and Teruji Thomas for valuable feedback and discussions.

preferred to the outcome of Y in some possible state, then X is strictly better than Y.

Hájek (2014, pp. 556–558) presents a case in which Expected Utility Theory violates Statewise Dominance when the principle is formulated in this way.[1] This dominance violation happens because, although we tend to think of probability zero as meaning impossible, this is not strictly true. Consider the following prospects:

Prospect A A fair coin is tossed an infinite number of times. If the coin lands heads on every toss, then the agent goes to heaven; otherwise, nothing happens.

Prospect B As above, but the agent goes to hell if the coin lands heads on every toss; otherwise, nothing happens.

In this case, A statewise dominates B. However, Expected Utility Theory assigns the same expected utility to both prospects because the probability that the coin lands heads on every toss is zero. Consequently, Expected Utility Theory permits the choice of a statewise-dominated prospect.

To avoid violating Statewise Dominance in this way, Hájek (2014, p. 556) argues that decision-makers should sometimes consider states of probability zero. He argues that there is more to the machinery of decision theory than just Expected

[1] Let EU (X) denote the expected utility of prospect X and let $X \succsim Y$ mean that X is at least as good as Y. Also, let O be the set of possible outcomes, $p_X(o)$ the probability of outcome o in prospect X and $u(o)$ the utility of o. Then, Expected Utility Theory states the following:

> **Expected Utility Theory:** For all prospects X and Y, $X \succsim Y$ if and only if EU(X) $\geq$ EU(Y), where
>
> $$\text{EU}(X) = \sum_{o \in O} p_X(o)u(o).$$

Utility Theory, and he goes on to suggest that Expected Utility Theory can be supplemented with dominance.[2] He argues, following Easwaran (2014), that there is no conflict between dominance and Expected Utility Theory in this case because Expected Utility Theory should not be interpreted as telling us that prospects with tied expected utilities must be treated with indifference. Instead, he argues that it should be interpreted as failing to tell us anything in such cases. So, without conflicting with what Expected Utility Theory tells us, we may choose on some other basis. As Easwaran (2014, p. 14) writes, in cases where the expected utility of a bet is the same as the status quo, some non-numerical feature may serve as a tiebreaker. Hájek (2014, p. 557) suggests that dominance may serve as a tiebreaker between A and B. However, I will show that expected utility theorists cannot use Statewise Dominance to argue that A is better than B, at least if they wish to keep standard axiomatizations of Expected Utility Theory.

1 A violation of Continuity

Let $X \succ Y$ mean that X is better than Y. Also, let XpY be a risky prospect with a p chance of prospect X obtaining and a $1 - p$ chance of prospect Y obtaining. Then, using Statewise Dominance to argue that A is better than B would result in

[2] Hájek (2014, p. 557). Russell (2021, p. 12–13 n. 9) also suggests that a prospect that spares a child from malaria if an ideally sharp dart hits a particular point (and otherwise nothing happens) may be better than the prospect of certainly getting nothing, even though it gives a probability zero of a positive outcome. Russell suggests that what is best may depend on what features of its outcomes are sure, which can come apart from what is almost sure.

a violation of the following axiom of Expected Utility Theory:[3]

> **Continuity:** If $X \succ Y \succ Z$, then there are probabilities p and $q \in (0, 1)$ such that $XpZ \succ Y \succ XqZ$.

To see why using Statewise Dominance would result in a violation of Continuity, consider the following case (see table 1).

> *Prospect* A^* A fair coin is tossed an infinite number of times. The agent gets \$10 if the coin lands heads on every toss; otherwise, nothing happens.
>
> *Prospect* B^* As above, but the agent gets \$1 if the coin lands heads on every toss; otherwise, nothing happens.
>
> *Prospect* C Certainly gives $-$\$10 (the agent loses \$10).

By Statewise Dominance, A^* is better than B^*, which is better than C.

Next, consider the following mixed prospect:

> *Prospect* A^*pC Gives A^* with probability p and C with probability $1 - p$.

In this case, A^*pC is worse than B^* for all probabilities $p \in (0, 1)$. This is so because A^*pC gives a probability p of nothing and a (non-zero) probability $1 - p$ of losing \$10, while B^* gives a probability one of nothing. Suppose the utility of money equals the monetary amount. Consequently, the expected utility of A^*pC

[3] Jensen (1967, p. 174). Note that strictly speaking Statewise Dominance is undefined in the framework of decision theory under risk, as this notion pertains to decision theory under uncertainty, where there is an explicit underlying state space.

is $EU(A^*pC) < 0$, and the expected utility of B^* is $EU(B^*) = 0$.[4] So, now we have that A^* is better than B^*, which is better than C, but A^*pC is worse than B^* for all probabilities $p \in (0, 1)$—which is a violation of Continuity.[5]

TABLE 1
A VIOLATION OF CONTINUITY

Probability	0	p	$1 - p$
A^*pC	\$10	\$0	−\$10
B^*	\$1	\$0	\$0

2 Conclusion

To conclude, standard axiomatizations of Expected Utility Theory are incompatible with using Statewise Dominance in cases that involve possible states of probability zero because doing so would result in a violation of the Continuity axiom.[6]

[4] $EU(A^*pC) = (1 - p) \times (-10) = -10 + 10p < 0$ for all probabilities $p \in (0, 1)$.

[5] Some might insist that the probability that the coin lands heads on every toss is not zero but infinitesimal. See for example Lewis (1980, p. 270) and Hájek (2014, p. 556 n. 19). It is unclear how infinitesimal probabilities figure in the decision-making process. If they can only serve as tiebreakers in cases where the prospects are otherwise equally preferable, then A^*pC is still worse than B^* for all probabilities $p \in (0, 1)$. So, using infinitesimal probabilities does not help avoid violating Continuity. On the other hand, utilities associated with infinitesimal probabilities might do something more than merely serve as tiebreakers. But it is unclear what their role would be. Any positive or negative contributions to utility would depart from Expected Utility Theory. Note that those who appeal to infinitesimal probabilities might weaken Continuity to accommodate non-Archimedean probabilities. See for example Hammond (1994). See Williamson (2007) for an argument against the appeal to infinitesimals.

[6] Note that if we define Statewise Dominance not in terms of possible states but in terms of states that have non-zero and non-infinitesimal probabilities (as is typically done), then Expected Utility Theory does not violate Statewise Dominance:

> **Statewise Dominance (Non-Zero and Non-Infinitesimal Probabilities:)** If the outcome of prospect X is at least as preferred as the outcome of prospect Y in all

References

Easwaran, K. (2014), 'Regularity and hyperreal credences', *Philosophical Review* **123**(1), 1–41.

Hájek, A. (2014), 'Unexpected expectations', *Mind* **123**(490), 533–567.

Hammond, P. J. (1994), Elementary non-Archimedean representations of probability for decision theory and games, *in* P. Humphreys, ed., 'Patrick Suppes: Scientific Philosopher', volume i: Probability and probabilistic causality edn, Kluwer, pp. 25–59.

Jensen, N. E. (1967), 'An introduction to Bernoullian utility theory: I. Utility functions', *The Swedish Journal of Economics* **69**(3), 163–183.

Lewis, D. (1980), A subjectivist's guide to objective chance, *in* R. C. Jeffrey, ed., 'Studies in Inductive Logic and Probability', volume ii edn, University of California Press, Berkeley, pp. 263–293.

Russell, J. S. (2021), 'On two arguments for fanaticism'. Global Priorities Institute Working Paper 17–2021.

URL: *https://globalprioritiesinstitute.org/on-two-arguments-for-fanaticism-jeff-sanford-russell-university-of-southern-california/*

states with non-zero and non-infinitesimal probability, then X is at least as good as Y. Furthermore, if in addition the outcome of X is strictly preferred to the outcome of Y in some state with a non-zero and non-infinitesimal probability, then X is strictly better than Y.

Williamson, T. (2007), 'How probable is an infinite sequence of heads?', *Analysis* **67**(3), 173–180.

Chapter 3

*Probability Discounting Solves the Intrapersonal Addition Paradox**

ABSTRACT: Nebel (2019) argues for the Repugnant Conclusion via the "Intrapersonal Repugnant Conclusion", on which certainty of a mediocre life is better for individuals than a sufficiently small chance of an excellent life. In this chapter, I deny that accepting the Intrapersonal Repugnant Conclusion leads us to the Repugnant Conclusion. I point out that on many views which avoid the Repugnant Conclusion, we should discount small probabilities down to zero to avoid an implausibly reckless decision theory. But if we do, then Nebel's crucial premise of Ex Ante Pareto fails because discounting at the individual level can fail to match up with discounting at the population level.

*I wish to thank Tomi Francis, Johan Gustafsson, Andreas Mogensen, Teruji Thomas, two anonymous reviewers of Ethics and the audience of the Slippery Slope Normativity Summit 2020 for valuable feedback and discussions. A version of this chapter was published in *Ethics*. See Kosonen (2021).

The Repugnant Conclusion, introduced by Parfit, states:[1]

> "For any possible population of at least ten billion people, all with a very high quality of life, there must be some much larger imaginable population whose existence, if other things are equal, would be better even though its members have lives that are barely worth living."

More generally, the Repugnant Conclusion is that for any population wherein each individual has very high positive welfare, there is some much larger population wherein each individual has very low positive welfare, which is better.[2] The Repugnant Conclusion is a consequence of Total Utilitarianism, which states that one population is better than another just in case the total quantity of welfare it contains is greater. The Repugnant Conclusion strikes many as an unacceptable consequence, and various attempts at constructing an alternative population axiology to Total Utilitarianism have been made.[3] However, a series of impossibility theorems have shown that no axiology can satisfy simultaneously all intuitively compelling principles that have been identified.[4]

In *An Intrapersonal Addition Paradox*, Nebel argues for the Repugnant Conclusion via an intrapersonal analogue of it.[5] In this chapter, I deny that accepting the Intrapersonal Repugnant Conclusion leads us to the Repugnant Conclusion. My argument is that on many views which avoid the Repugnant Conclusion, we should discount small probabilities down to zero to avoid an implausibly reckless decision theory. But if we do, then Nebel's

[1] Parfit (1984, p. 388).

[2] Arrhenius (2000, p. 248).

[3] For an overview, see Greaves (2017).

[4] Arrhenius (2000).

[5] Nebel (2019, pp. 309–343).

crucial premise of Ex Ante Pareto fails because discounting at the individual level can fail to match up with discounting at the population level. The structure of this chapter is as follows: §1 presents Nebel's argument. §2 explores the idea that we should discount small probabilities down to zero. §3 responds to Nebel's argument. §4 argues that those who discount small probabilities must reject Ex Ante Pareto. §5 concludes.

1 Nebel's argument

This section presents Nebel's argument. His argument proceeds in two stages. The first stage is an argument for the Intrapersonal Repugnant Conclusion, and the second stage is an argument from the Intrapersonal Repugnant Conclusion to the Repugnant Conclusion. I will begin by discussing the first stage of the argument.

1.1 The Intrapersonal Repugnant Conclusion

Nebel's *Intrapersonal Repugnant Conclusion* states:[6]

> **Intrapersonal Repugnant Conclusion:** For any person S, there exists some probability p such that any prospect in which S would have a wonderful life with probability p or less, and would otherwise never exist, is worse for S than certainty of a life that is barely worth living.[7]

[6] Nebel (2019, p. 314). There is another—distinct—claim that could also be naturally described as the Intrapersonal Repugnant Conclusion: For any life lived by S at a very high welfare level for n years, there is some much longer life that would be better for S in which her welfare is barely above the zero level at each point in time. Temkin (2012, p. 119) calls this the Single Life Repugnant Conclusion. See also McTaggart (1927, pp. 452-453).

[7] The sense of 'worse' at issue here is '*ex ante* worse'. And, the sense of 'betterness' that is now used belongs to *ex ante* axiology, as opposed to *ex post* axiology, which is at issue in standard discussions of population

When arguing for the Intrapersonal Repugnant Conclusion, Nebel considers a couple that is planning to conceive a child by injecting a single sperm into a single egg.[8] Suppose that only one person could possibly originate from this pair of gametes—call her Sally. Sally's parents have three different ways they can do this injection: $\mathcal{A}$, $\mathcal{Z}$ and $\mathcal{A}+$ (see table 1). $\mathcal{A}$ will give Sally a very happy life at welfare level a if state 1 obtains, but she will not exist if state 2 obtains. $\mathcal{Z}$ will give Sally a low positive life at welfare level z in both outcomes. Lastly, $\mathcal{A}+$ will give Sally welfare level $a+$ (slightly above a) if state 1 obtains and welfare level $z-$ (slightly below z) if state 2 obtains.

Table 1
The Intrapersonal Argument

	State 1 (p)	State 2 ($1-p$)
$\mathcal{A}$	a	
$\mathcal{A}+$	$a+$	$z-$
$\mathcal{Z}$	z	z

Nebel argues that $\mathcal{A}+$ is better than $\mathcal{A}$ for Sally because Sally's welfare is higher if state 1 obtains, and her life would be worth living if state 2 obtains. This is supported by the following principle:[9]

> **Probable Addition Principle:** If, in every state of the world in which a person S would exist in Y, S would be better off in X, and if, in every other state, S's life would be worth living in X, then X is better than Y for S.[10]

axiology.

[8] Nebel (2019, p. 313).

[9] Nebel (2019, p. 315).

[10] One could object that if $\mathcal{A}$ is chosen and state 2 obtains, then Sally does not exist, and therefore there

Next, Nebel argues that $\mathcal{Z}$ must be better than $\mathcal{A}+$ when the probability of state 1 is very small. Suppose that it is one-in-a-googolplex. Then, it would be irresponsible for Sally's parents to choose $\mathcal{A}+$ instead of $\mathcal{Z}$, as $\mathcal{A}+$ has such a small probability of resulting in a better outcome ($a+$ instead of z) and a very high probability of resulting in a worse outcome ($z-$ instead of z). This is supported by the following principle:[11]

> **Minimal Prudence:** No matter how good some life would be[12], there is some small probability and some pair of mediocre lives such that certainty of the better mediocre life would be better for some person S than a gamble that might yield the very good life but would almost certainly yield the worse mediocre life.

Next, if $\mathcal{A}+$ is better than $\mathcal{A}$ (by the Probable Addition Principle), and $\mathcal{Z}$ is better than $\mathcal{A}+$ (by Minimal Prudence), it follows by transitivity that $\mathcal{Z}$ must be better than $\mathcal{A}$ for Sally—which is the Intrapersonal Repugnant Conclusion.[13] So, accepting the Probable Addition Principle, Minimal Prudence and the transitivity of *better than* leads to the Intrapersonal Repugnant Conclusion. This conclusion is not repugnant, but Nebel argues that accepting it leads to the Repugnant Conclusion. Next, I will summarize the second stage of his argument.

---

is no one for whom it would have been better had $\mathcal{A}+$ been chosen instead. Thus, it is not the case that $\mathcal{A}+$ must be better than $\mathcal{A}$ for Sally. Nebel (2019, §V) discusses similar concerns, but I will not address them here as my argument does not rely on them.

[11] Nebel (2019, p. 316).

[12] An exception here could be an infinitely good life. An agent who maximizes expected value would prefer a gamble with any probability of an infinitely good outcome.

[13] This argument is structurally analogous to Parfit's Mere Addition Paradox and Huemer's Benign Addition Paradox. See Parfit (1984, ch 19) and Huemer (2008, pp. 899–933).

1.2 From intrapersonal to interpersonal Repugnant Conclusion

Nebel considers a simplified case to show the derivation of the Repugnant Conclusion from its intrapersonal analogue. Consider these two outcomes (see table 2):

The Repugnant Conclusion:

$\mathbf{A_0}$: Ann has a very happy life at welfare level a.

Z: Bob, Cat and Dan have mediocre lives at welfare level z.

Suppose that the Intrapersonal Repugnant Conclusion is true and that (unrealistically) a 1/3 chance of having a very happy life at welfare level a is worse for a person than certainly having a mediocre life at welfare level z.[14] Nebel argues that this will lead to the conclusion that **Z** is better than $\mathbf{A_0}$ (i.e., the Repugnant Conclusion).

TABLE 2
THE REPUGNANT CONCLUSION

	Ann	Bob	Cat	Dan
$\mathbf{A_0}$	a			
Z		z	z	z

Nebel considers three prospects: A, $A*$ and Z (see table 3). Prospect A certainly results in outcome $\mathbf{A_0}$ (Ann has a very happy life at welfare level a), while prospect $A*$ results in either Bob, Cat or Dan getting a very happy life at welfare level a, each with a 1/3 probability. Nebel argues that A and $A*$ are equally good, given that they are egalitarian prospects that guarantee equally good outcomes.[15] Both prospects result in one person existing at welfare

[14] Nebel's argument can be generalized to more realistic instances of the Intrapersonal Repugnant Conclusion with probabilities smaller than 1/3. Nebel (2019, p. 323).

[15] Nebel (2019, p. 319).

level a, and the probabilities in $A*$ do not favor anyone over the others.[16] If the outcomes are equally good, and no one is unfairly advantaged in either prospect, then the prospects must be equally good.

Table 3
To the Repugnant Conclusion

	State 1 (1/3)				State 2 (1/3)				State 3 (1/3)			
	Ann	Bob	Cat	Dan	Ann	Bob	Cat	Dan	Ann	Bob	Cat	Dan
A	a				a				a			
$A*$		a					a					a
Z		z	z	z		z	z	z		z	z	z

Nebel argues that A and $A*$ are equally good, and that Z is better than $A*$.

Next, compare prospect $A*$ to prospect Z, which guarantees outcome **Z** (Bob, Cat and Dan all have mediocre lives at welfare level z). As we have already assumed that a 1/3 chance of existing with a very happy life is worse for a person than certainly existing with a mediocre life, it follows that $A*$ is worse than Z for each Bob, Cat and Dan. Consequently, it must be overall worse, Nebel argues. This is supported by the following principle:

> **Weak Pareto for Equal Risk:** For any egalitarian prospects X and Y, if X is better than Y for each person who might exist in either prospect, then X is better than Y.[17]

[16] According to some person-affecting views, these options might be incomparable in value because different people exist in their outcomes. For a discussion of the narrow person-affecting principle, see Parfit (1984, ch 16 and 18).

[17] Fleurbaey (2010, p. 656) and Nebel (2019, p. 320). 'Egalitarian' here should be understood in either *ex ante* or *ex post* sense amongst those who will exist—otherwise $A*$ would not be egalitarian.

This principle states that a prospect is overall better than another prospect if it is better for everyone who might exist in either prospect—at least when there is no risk of unfairness.

Nebel offers an inductive argument in favor of this Pareto principle.[18] First, he argues that if only a single person might exist, and X is better than Y for that person, then X must be overall better than Y. Because X is better than Y for that person, we ought to prefer X for the sake of that person. And because we ought to prefer X for the sake of the only person who might exist, from an impartial perspective, we ought to prefer X. Thus, the Pareto principle is true for all prospects in which only a single person might exist.

Then Nebel considers egalitarian prospects X and Y in which any number n of people might exist. He argues that if X is better than Y for n number of people, then X must also be better than Y for $n + 1$ number of people. It cannot be the case that the addition of one more person (for whom X is also better than Y) suddenly reverses the betterness relation between X and Y—if X is better than Y for the additional person (and everyone else), then X must remain better than Y. Nebel argues that if this inductive step was unjustified, it should be for some reason having to do with some relation between the $(n + 1)$th person and the others. If there was any risk of inequality, then the relational fact that some might be worse off than others could be blamed—but there is no such risk as the prospects are egalitarian. Therefore, Nebel argues that it is hard to see why the principle should be true for n but not for $n + 1$.[19] So, Weak Pareto for Equal Risk is true when only a single person might exist, and by the inductive step, it is also true when two people might exist, and again by the inductive step, it is true when three people might exist, and so on. No matter how many people might exist, if X is better than Y for each of those people, then X is better than

[18] Nebel (2019, p. 321).

[19] Nebel (2019, p. 322).

Y.

Nebel's inductive argument for Weak Pareto for Equal Risk requires us to think that the principle fails even when only a single person might exist or that the difference between its true and false instances lies in the addition of only a single possible person whose existence will not generate a trade-off between different people's interests.[20] Nebel argues that neither of these possibilities seems very plausible, and thus, he concludes that we ought to accept Weak Pareto for Equal Risk.

The derivation of the Repugnant Conclusion from the Intrapersonal Repugnant Conclusion thus goes like this: Certainly existing with a mediocre life is better for a person than a 1/3 chance of existing with a very happy life (by the Intrapersonal Repugnant Conclusion). Thus, Z is better than $A*$ for each person who might exist in either prospect. By Weak Pareto for Equal Risk, Z is therefore overall better than $A*$. Furthermore, $A*$ is equally good as A because they are egalitarian prospects that guarantee equally good outcomes. Consequently, as A and $A*$ are equally good, and Z is better than $A*$, Z must be better than A. Lastly, because Z guarantees outcome $\mathbf{Z}$ and A guarantees outcome $\mathbf{A_0}$, and Z is better than A, outcome $\mathbf{Z}$ must be better than outcome $\mathbf{A_0}$—and we have arrived at the Repugnant Conclusion.[21]

Nebel considers the second stage of his argument to be more compelling than the argument for the Intrapersonal Repugnant Conclusion, so he focuses on possible responses to the latter.[22] He discusses how Parfit's *Perfectionism* could respond. Perfectionism states that "even if some change brings a great net benefit to those who are affected, it is a change

[20] Nebel (2019, p. 322).

[21] Nebel (2019, p. 323).

[22] Nebel (2019, p. 324).

for the worse if it involves the loss of one of the best things in life."[23] Perfectionism could respond that, even if some prospect brought a great net benefit to a person, it is worse for her if it lowers her probability of enjoying the best things in life. Perfectionism would therefore deny that $\mathcal{Z}$ is better than $\mathcal{A}+$ for Sally. However, Nebel argues that it is irrational to prefer prospects that will almost certainly be worse for us in the pursuit of arbitrarily small chances of enjoying the best things in life—this would be an absurdly reckless decision theory.[24]

In this chapter, I will argue that Perfectionism (and other population axiologies) can respond to Nebel's challenge without being absurdly reckless by discounting small probabilities.[25,26] However, first I will discuss an independent motivation for this kind of discounting.

2 Discounting small probabilities

According to orthodox decision theory, a rational agent always maximizes expected utility. However, doing this would lead one to make highly counter-intuitive choices when presented with options that have a small probability of a huge payoff. One such case is the St. Petersburg paradox, a version of which was originally proposed by Nicolaus Bernoulli in

[23] Parfit (2004, p. 19).

[24] Nebel (2019, p. 324).

[25] As noted by an anonymous reviewer, someone might argue that we should not draw axiological conclusions from normative arguments about how we should treat very small probabilities. However, if this is true, then Nebel's argument never gets off the ground because we should not draw axiological conclusions from the value of prospects when such normative arguments are relevant to their value.

[26] Parfit regarded ignoring tiny chances as one of the five mistakes in 'moral mathematics'. He (1984, p. 75) writes: "When the stakes are very high, no chance, however small, should be ignored. The same is true when each chance will be taken very many times. In both these kinds of cases, each tiny chance should be taken to be just what it is, and included in the calculation of the expected benefit. We can usually ignore a very small chance. But we should not do so when we may affect a very large number of people, or when the chance will be taken a very large number of times." See Parfit (1984, pp. 73–75).

1713.[27,28] The modern version of the game is played by flipping a fair coin until it lands on heads. The prize is then $\$2^n$, where n is the number of coin flips. This game has infinite expected monetary value, so an agent who maximizes expected monetary value would pay any finite amount to play it. However, this seems counter-intuitive. As Nicolaus Bernoulli, agreeing with his friend Gabriel Cramer, writes: "[T]here is no person of good sense who wished to give merely 20 coins."[29]

Daniel Bernoulli (cousin of Nicolaus Bernoulli) argues that we should not be willing to pay any finite sum to play the St. Petersburg game because of the diminishing marginal utility of money.[30] He argues that the expected utility of the game is finite even though it has infinite expected monetary value. However, one can change the game slightly to bypass this objection by changing the prize from money to something that has no diminishing marginal utility, such as (possibly) days of life.[31,32] When the payoffs are utilities, the game

[27] The game was then simplified by Gabriel Cramer in 1728 and published by Daniel Bernoulli in 1738. See Pulskamp (2013) and Bernoulli (1954).

[28] Bostrom (2009) presents another case that involves a very small probability of a huge payoff.

[29] Pulskamp (2013, p. 6).

[30] More specifically, he argues that the utility of money equals the logarithm of the monetary value. See Bernoulli (1954). Cramer (Pulskamp, 2013, p. 4) also came close to suggesting that money has diminishing marginal utility: "One asks the reason for the difference between the mathematical calculation and the common value. I believe that it comes from this that the mathematicians value money in proportion to its quantity, and men of good sense in proportion to the usage that they may make of it."

[31] Monton (2019, p. 2). Relatedly, Menger (1967, pp. 217–218) shows that if utilities are unbounded, one can always create a *Super St-Petersburg game*, in which the payoffs grow sufficiently fast so that the expected utility of the game is infinite. See also Samuelson (1977, §2).

[32] Although one could argue that the longer one has lived, the less valuable extra days of life are. Temkin writes: "I believe that in many cases, though certainly not all, once people have experienced certain kinds of events 'enough' times in their lives, there will be a diminishing marginal value to subsequent similar experiences." See Temkin (2008, p. 208). One could also add that at some point, it is not even possible to have new valuable experiences that are different enough from one's past experiences such that this diminishing marginal utility does not happen—at some point everything worth experiencing has been experienced.

has infinite expected utility.[33] Nevertheless, few would sacrifice the rest of one's days as a payment to play the game in the hopes of living longer (though almost certainly dying soon), and this reluctance seems rational. Furthermore, if the game has infinite expected utility and we value gambles at their expected utilities, then we value the St. Petersburg game more than any of its possible (finite) payoffs—which seems clearly irrational.[34]

Nicolaus Bernoulli, in turn, argues that in order to solve this paradox, we ought to discount very small probabilities down to zero—let's call this *Probability Discounting*. He writes: "[T]he cases which have a very small probability must be neglected and counted for nulls, although they can give a very great expectation."[35] More recently, Smith and Monton have argued for the same idea.[36] Monton argues that in order to avoid the fate of the expected utility maximizers, we need to either limit the high utility numbers or discount the small probability numbers in cases that involve very small probabilities of huge payoffs.[37] However, he argues that introducing a utility cap would be ethically problematic because one can always add more agents into the utility calculation and that the utilities of those individuals matter regardless of how many agents already exist. Thus, bounding utility is not viable[38], which leaves the only other option: discounting very small probabilities. Monton then argues that very small probabilities need to be discounted down to zero instead

[33] Many decision theorists reject an unbounded utility function. However, even if utility is bounded, the expected utility of the St. Petersburg game can still be very high if the upper bound of utility is very high.

[34] Huemer (2016, pp. 34–35) and Russell and Isaacs (2021).

[35] Pulskamp (2013, p. 2); the German original von Spieß (1975).

[36] Smith (2014) considers it permissible to discount very small probabilities down to zero, while Monton (2019) argues that one is rationally required to do so. Smith argues that Probability Discounting is a way of getting a reasonable expected utility for the Pasadena game. See Nover and Hájek (2004).

[37] Monton (2019, p. 5).

[38] Standard axiomatizations of expected utility maximization, such as the von Neumann-Morgenstern utility theorem, require utility to be bounded—or else the continuity axiom is violated. See Kreps (1988, p. 63).

of merely reducing those probabilities because one can always increase the payoffs of the games by a sufficient amount to compensate for those reduced probabilities.[39]

To summarize, orthodox decision theory gives highly counter-intuitive recommendations in cases that involve very small probabilities of huge payoffs. In response to such cases, some have argued that we ought to discount very small probabilities down to zero.[40] Next, I will argue that Probability Discounting can also solve the Intrapersonal Addition Paradox.

3 A response to Nebel's inductive argument

In this section, I will argue that the inductive step of Nebel's inductive argument for Weak Pareto for Equal Risk is unjustified due to the cumulative nature of probabilities if one engages in Probability Discounting.

Recall that Sally's parents need to make a choice between $\mathcal{A}$, $\mathcal{A}+$ and $\mathcal{Z}$. For the purposes of this chapter, let's grant the Probable Addition Principle. Consequently, $\mathcal{A}+$ is better than $\mathcal{A}$ for Sally. Also, I accept Minimal Prudence; $\mathcal{Z}$ is better than $\mathcal{A}+$ for Sally when the probability of state 1 is very small because very small probabilities should be discounted down to zero. Sally's parents should ignore the small possibility of Sally getting a life at welfare level $a+$ and compare the options for their remaining outcomes: a life at welfare level

[39] Monton (2019, p. 5).

[40] Although there may also be some more fundamental justification for Probability Discounting. According to Monton, maximizing expected utility is a mistake because "you only live once", and the prescription to maximize expected utility does not take seriously the importance of how one's life actually goes. In contrast, Smith's argument is that decision theory tells us to ignore outcomes with zero probability, and because decision-making is a practical activity, infinite precision cannot be required. Smith also argues that Probability Discounting is a way of getting a unique expected value for the Pasadena game. See Nover and Hájek (2004) on the Pasadena game. For a discussion of other possible justifications, see Monton (2019). For further discussion of discounting small probabilities down to zero, see Smith (2014, 2016), Hájek (2014) and Isaacs (2016).

$z-$ in $\mathcal{A}+$ and a life at welfare level z in $\mathcal{Z}$. Because z is greater than $z-$, $\mathcal{Z}$ is better than $\mathcal{A}+$ for Sally. Consequently, the argument for the Intrapersonal Repugnant Conclusion goes through.[41] However, I will argue that this does not lead to the Repugnant Conclusion.

I will use the following principle in my argument:

> **Risky Non-Repugnance:** q chance (or greater) of obtaining at least one life at a high welfare level a is better than certainly obtaining any number of lives at a low welfare level z, where q is the smallest probability that should not be discounted down to zero.[42,43]

Risky Non-Repugnance states that a non-negligible probability of at least one very good life is better than certainty of any number of low positive lives. It can be supported with Perfectionism and some other types of Value Superiority. Value Superiority avoids the Repugnant Conclusion because, according to it, no quantity of low positive lives could ever be as good as some number of very good lives—very good lives are lexically superior to low positive lives. Then, one can argue that whatever the smallest probability that should not be discounted is, one should choose that probability of obtaining at least one very good life

[41] Actually, it might not: If one discounts the probability of state 1 down to zero, then the only outcome left in $\mathcal{A}$ is non-existence. Thus, one cannot use the Probable Addition Principle when comparing $\mathcal{A}$ and $\mathcal{A}+$ because one is comparing existence with non-existence.

[42] Alternatively, one could discount anything up to and including some small probability q.

[43] A related principle could be called *Reckless Risky Non-Repugnance*: any non-zero probability of obtaining one life at a high welfare level a is better than certainly obtaining any number of lives at a low welfare level z. Risky Non-Repugnance is also similar to what one might call *Intrapersonal Risky Non-Repugnance*: q chance of life at a high welfare level a is better for some person S than certainty of life at a low welfare level z. Another related principle could be called *Intrapersonal Risky Welfare-Level Superiority*: q chance of a life of some length t at a high momentary well-being level a is better for a person S than certainty of any length of life at a low momentary well-being level z. While Risky Non-Repugnance (and its reckless version) are about the contributive value of lives to the value of a population, the latter two principles are about what is good for individuals. One can also make reckless versions of them by replacing 'q chance' with 'any non-zero probability'.

instead of certainty of any number of low positive lives.[44]

However, one might think Risky Non-Repugnance is implausible because the smallest probability that should not be discounted down to zero might be very small. Say it is one-in-a-trillion. Then, according to Risky Non-Repugnance, a one-in-a-trillion chance of one very good life is better than certainty of any number of low positive lives. A few things can be said in favor of Risky Non-Repugnance. First, there are cases in which Risky Non-Repugnance does not seem counter-intuitive: q chance of one very good life is better than certainty of any number of 10-second-lives consisting of a barely positive emotion. No quantity of 10-second-lives could ever be as good as one very good life, and q chance of a very good life is still better than those 10-second-lives. Secondly, q might actually be higher than one-in-a-trillion. However, by definition, q is a probability that should not be discounted down to zero—it is a probability that we should pay attention to and consider non-negligible.

Lastly, finding an intuitively acceptable population axiology is notoriously difficult, and this task gets even harder when we take risk into account. Finding Risky Non-Repugnance

[44]Some versions of Value Superiority might accept the following principle instead of Risky Non-Repugnance:

> **Weak Risky Non-Repugnance:** There is some probability p (less than 1) and some number n of very good lives such that p chance of n very good lives is better than certainty of any number of low positive lives.

On these versions of Value Superiority, the value of additional positive lives at some welfare level *w* diminishes the more such lives there already are, and their total contributive value approaches some upper bound. As this bound is higher for very good lives than for low positive lives, some number of very good lives is better than any number of low positive lives. However, the probability of obtaining very good lives might have to be high for that prospect to be better than certainty of any number of low positive lives. Lazar and Lee-Stronach (2019) defend this kind of approach in the context of limited aggregation and risk. They argue against an *infinitist* approach (such as Risky Non-Repugnance), which posits an infinite value difference between higher and lower considerations.

counter-intuitive is not a decisive reason for rejecting it if one must bite the bullet anyway and the alternatives are even worse. Consider, for example, this implication of Expected Total Utilitarianism:

> **Risky Very Repugnant Conclusion:** For any number of very good lives that could be obtained for certain, there is a prospect that certainly gives many very bad lives together with a small probability that in addition there will exist some very large number of barely good lives, which is better, provided that the quantity of barely good lives is sufficient.

The quantity of barely good lives would, of course, have to be enormous for this to be true, as the goodness of those mediocre lives must be enough to outweigh the badness of the very bad lives and the risk introduced. Nevertheless, the world would almost certainly be arbitrarily bad. And, in the best-case scenario, it would only contain very bad and barely positive lives—yet Expected Total Utilitarianism would still recommend that option. In comparison, Risky Non-Repugnance does not seem counter-intuitive.[45]

Now, recall that Weak Pareto for Equal Risk states that we ought to prefer prospects that are better for everyone. Next, I will argue that—contrary to the inductive step—the addition of only a single possible person can make the difference between the true and false instances of Weak Pareto for Equal Risk. I will illustrate my argument using the choice Sally's parents have to make, and for simplicity, I will compare $\mathcal{A}$ and $\mathcal{Z}$ (instead of $\mathcal{A}+$ and $\mathcal{Z}$). Consider the following situation:

[45]However, views that imply Risky Non-Repugnance might have other—even more counter-intuitive—implications than Risky Non-Repugnance. Ultimately, we must compare complete theories against one another.

Iterated Sally's Parents' Choice: A great number of couples face the choice between $\mathcal{A}$ and $\mathcal{Z}$, and the probability of obtaining a very good life is independent every time if $\mathcal{A}$ is chosen repeatedly.

If $\mathcal{A}$ is chosen repeatedly, the probability of obtaining at least one very good life accumulates. Thus, with some k number of choices of $\mathcal{A}$, the probability of obtaining at least one very good life is less than the threshold for Probability Discounting (and thus should be discounted down to zero). But with $k + 1$ number of choices, that probability is above or equal to the threshold (and thus should *not* be discounted down to zero). If $\mathcal{A}$ is chosen enough times, the probability that at least one person gets a very good life is greater than or equal to q—a prospect that is better than certainty of any number of low positive lives, according to Risky Non-Repugnance.

The inductive step of the inductive argument for Weak Pareto for Equal Risk is unsound. Certainty of k individuals obtaining low positive lives is better than the prospect of them all having p chance of getting a very good life because the cumulative probability of obtaining at least one very good life is still rationally negligible. The former prospect is also better for each individual who might exist because the probability of them obtaining a very good life is also rationally negligible. However, certainty of $k + 1$ individuals obtaining low positive lives is worse than them all having a p chance of getting a very good life. This is because the cumulative probability of obtaining at least one very good life is no longer rationally negligible, and Risky Non-Repugnance judges that prospect to be better than $k + 1$ individuals obtaining low positive lives for certain.[46] So, Weak Pareto for Equal Risk is true for

[46] A weaker principle than RNR would be sufficient here, as the principle only needs to state that a q chance of a very good life is better than certainty of $k + 1$ low positive lives.

k individuals, but it is not true for $k + 1$ individuals. The accumulation of probabilities is the relational fact that renders the inductive step false.[47] Consequently, the value of a series of choices is not just an aggregation of the value of each individual choice.[48,49]

It is worth pointing out that those who accept a probability-discounting version of Expected Total Utilitarianism must also reject the inductive step. Consider the case of Sally again. Say that the very good a-life is sufficiently good such that the expected value of obtaining one such life with probability q is higher than the value of certainly obtaining $k + 1$ lives at welfare level z. As before, with up to k number of choices, $\mathcal{A}$ is judged overall worse than $\mathcal{Z}$ because the possibility of getting at least one very good life is rationally negligible. However, with $k + 1$ number of choices, $\mathcal{A}$ is judged overall better than $\mathcal{Z}$ because q chance of at least one very good a-life has a higher expected value than certainty of $k + 1$ lives at welfare level z. Thus, the addition of a single possible person can make the difference between the true and false instances of Weak Pareto for Equal Risk if one accepts a probability-discounting version of Expected Total Utilitarianism.

Lastly, the probabilities (of getting a very good life) are not independent in the second stage of Nebel's argument because either Bob, Cat or Dan would have a very good life for

[47] One could object that it is implausible that a very small difference in probabilities (from just below q to just above it) can make all the difference. One possible response is that the threshold q might be vague.

[48] There is also the question of what decision-makers should do when they know they face a series of choices involving a very small probability of a huge payoff. Should they refrain from discounting in the last choice of the series, even if they would discount in a similar one-off choice? To deal with iterated choices, probability discounters could accept *Resolute Choice*. A resolute agent chooses according to any plan they have adopted earlier as long as nothing unexpected has happened since then. Probability discounters can then form a plan to not discount in any of the choices in the series, even if they would discount in a similar one-off choice. See McClennen (1990) on Resolute Choice. However, Probability Discounting in combination with Resolute Choice leads to untenable results. See §3.2 in Chapter 5 of this thesis.

[49] As noted by an anonymous reviewer, other views on which the value of conjunction of acts is different from the sum of the value of the individual acts can also block Nebel's argument. For a discussion of this possibility in the context of limited aggregation and risk, see Tadros (2019).

certain if $A*$ were chosen. However, my argument provides a justification for rejecting the inductive step of Nebel's inductive argument. Next, I will argue that one should reject Ex Ante Pareto principles such as Weak Pareto for Equal Risk if one discounts very small probabilities down to zero.

4 An argument against Ex Ante Pareto from discounting small probabilities

Contrary to Ex Ante Pareto, a prospect can be impersonally better, even if it is worse for everyone. This happens when the probability of an individual receiving some good (or harm) is discounted down to zero, but—collectively—the probability that at least one person receives that good (or harm) is large enough to be taken into account. $\mathcal{Z}$ is better than $\mathcal{A}$ for Sally, but it would be impersonally worse if the choice of $\mathcal{Z}$ over $\mathcal{A}$ was repeated a great number of times. Also, Z is better than $A*$ for all Bob, Cat and Dan, but $A*$ is impersonally better than Z. In both cases, the probability of obtaining a very good life is rationally negligible for the individual, but the probability of someone obtaining such a life is non-negligible.

Accepting Ex Ante Pareto and engaging in Probability Discounting gets one in trouble. Consider the following case:

> **Celebratory Gunfire:** Someone shoots into the air in an area full of people during a celebration, which causes people to feel excitement for a few seconds. The probability of any particular individual being hit by the bullet when it falls is negligibly small, but there is a high probability that someone is hit by it.

In this case, the prospect of shooting into the air is *ex ante* better than not shooting for

everyone; each individual feels excitement for a few seconds, and the probability of any particular individual being hit by the bullet is rationally negligible. However, the goodness of everyone feeling excitement is not enough to outweigh the badness of the likely injury. Consequently, shooting into the air is *ex ante* impersonally worse than not shooting—which contradicts Ex Ante Pareto. If one accepts Probability Discounting, one should also hold that impersonally better prospects are possible, or one would permit the infliction of arbitrarily severe harms for little or no benefits.[50]

I have argued that sometimes we should not choose prospects that are better for everyone. However, it also seems that sometimes it is permissible to only care about what is good for particular individuals instead of what would be impersonally best. This seems appropriate when one attempts to benefit one's family, friends or oneself. So, it would be permissible for Sally's parents to choose $\mathcal{Z}$ over $\mathcal{A}$, even if a great number of parents faced the same choice, because they are concerned with the welfare of their future child instead of attempting to make the world better overall. This discrepancy between what is *ex ante* good for individuals and what is impersonally *ex ante* good is a price to pay for Probability Discounting.

5 Conclusion

I have argued that one can solve the Intrapersonal Addition Paradox if one discounts very small probabilities down to zero. The Repugnant Conclusion does not follow from its intrapersonal analogue because we should reject one of the premises of the argument, namely, Weak Pareto for Equal Risk. First, this principle is not supported by the inductive argument

[50] Pareto principles have also been challenged before. See for example Sen (1970), Mongin (1995) and Temkin (2000).

if one engages in Probability Discounting. This is because the inductive step is unjustified due to the cumulative nature of probabilities. Secondly, this principle licenses the infliction of arbitrarily severe harms for little or no benefits if combined with Probability Discounting. This happens when the probability of harming each individual is small, but there is a high probability that someone is harmed.

Furthermore, we have independent reasons for engaging in Probability Discounting because it enables one to get intuitively right responses in decision-theoretic problems that involve very small probabilities of huge payoffs. To conclude, in order to avoid ending up like the expected utility maximizers in these situations, we must discount very small probabilities down to zero. But then, we must give up Ex Ante Pareto. And then, we can solve the Intrapersonal Addition Paradox. This argument has implications for other ethical debates as well, as this solution is somewhat similar to the solution posed to the problem of aggregation and risk.[51] This chapter adds to the idea that this solution has a principled foundation in a more general claim in decision theory: very small probabilities have no prudential or moral significance.[52]

References

Arrhenius, G. (2000), 'An impossibility theorem for welfarist axiologies', *Economics and Philosophy* **16**(2), 247–266.

Barrington, M. (forthcoming), 'Superiority Discounting implies the Preposterous Conclusion', *Utilitas* .

[51] Thanks to an anonymous reviewer of Ethics for raising this point. For discussions of aggregation and risk, see Lazar (2018), Lazar and Lee-Stronach (2019), Tadros (2019), and Horton (2020).

[52] See Barrington (forthcoming) for criticism of combining Probability Discounting with a lexical view.

Bernoulli, D. (1954), 'Exposition of a new theory on the measurement of risk', *Econometrica* **22**(1), 23–36.

Bostrom, N. (2009), 'Pascal's Mugging', *Analysis* **69**(3), 443–445.

Fleurbaey, M. (2010), 'Assessing risky social situations', *Journal of Political Economy* **118**(4), 649–680.

Greaves, H. (2017), 'Population axiology', *Philosophy Compass* **12**(11), e12442.

Hájek, A. (2014), 'Unexpected expectations', *Mind* **123**(490), 533–567.

Horton, J. (2020), 'Aggregation, risk, and reductio', *Ethics* **130**(4), 514–529.

Huemer, M. (2008), 'In defence of repugnance', *Mind* **117**(468), 899–933.

Huemer, M. (2016), *Approaching infinity*, Palgrave Macmillan, New York.

Isaacs, Y. (2016), 'Probabilities cannot be rationally neglected', *Mind* **125**(499), 759–762.

Kosonen, P. (2021), 'Discounting small probabilities solves the Intrapersonal Addition Paradox', *Ethics* **132**(1), 204–217.

Kreps, D. M. (1988), *Notes on the Theory of Choice*, Westview Press, Boulder.

Lazar, S. (2018), 'Limited aggregation and risk', *Philosophy & Public Affairs* **46**(2), 117–159.

Lazar, S. and Lee-Stronach, C. (2019), 'Axiological absolutism and risk', *Noûs* **53**(1), 97–113.

McClennen, E. F. (1990), *Rationality and Dynamic Choice: Foundational Explorations*, Cambridge University Press, Cambridge.

McTaggart, J. M. E. (1927), *The Nature of Existence*, Vol. 2, Cambridge University Press, Cambridge.

Menger, K. (1967), The role of uncertainty in economics, *in* M. Shubik, ed., 'Essays in Mathematical Economics: In Honor of Oskar Morgenstern', Princeton University Press, Princeton, pp. 211–231.

Mongin, P. (1995), 'Consistent bayesian aggregation', *Journal of Economic Theory* **66**(2), 313–351.

Monton, B. (2019), 'How to avoid maximizing expected utility', *Philosophers' Imprint* **19**(18), 1–24.

Nebel, J. M. (2019), 'An intrapersonal addition paradox', *Ethics* **129**(1), 309–343.

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

Parfit, D. (1984), *Reasons and Persons*, Clarendon Press, Oxford.

Parfit, D. (2004), Overpopulation and the quality of life, *in* T. Tännsjö and J. Ryberg, eds, 'The Repugnant Conclusion: Essays on Population Ethics', Springer Netherlands, Dordrecht, pp. 7–22.

Pulskamp, R. J. (2013), 'Correspondence of Nicolas Bernoulli concerning the St. Petersburg Game'. Unpublished manuscript. Accessed through: https://web.archive.org/.
URL: *http://cerebro.xu.edu/math/Sources/NBernoulli/correspondence_petersburg_game.pdf*

Russell, J. S. and Isaacs, Y. (2021), 'Infinite prospects', *Philosophy and Phenomenological Research* **103**(1), 178–198.

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

Sen, A. (1970), 'The impossibility of a paretian liberal', *Journal of Political Economy* **78**(1), 152–157.

Smith, N. J. J. (2014), 'Is evaluative compositionality a requirement of rationality?', *Mind* **123**(490), 457–502.

Smith, N. J. J. (2016), 'Infinite decisions and rationally negligible probabilities', *Mind* **125**(500), 1199–1212.

Tadros, V. (2019), Localized restricted aggregation, *in* D. Sobel, P. Vallentyne and S. Wall, eds, 'Oxford Studies in Political Philosophy Volume 5', Oxford University Press, Oxford, pp. 171–203.

Temkin, L. S. (2000), Equality, priority, and the levelling down objection, *in* M. Clayton and A. Williams, eds, 'The Ideal of Equality', Macmillan, London, pp. 126–161.

Temkin, L. S. (2008), 'Is living longer living better?', *Journal of Applied Philosophy* **25**(3), 193–210.

Temkin, L. S. (2012), *Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning*, Oxford University Press, New York.

von Spieß, O. (1975), Zur vorgeschichte des Petersburger problems, *in* B. van der Wærden, ed., 'Die Werke von Jakob Bernoulli Band 3: Wahrscheinlichkeitsrechnung', Birkhäuser, Basel, pp. 557–567.

Chapter 4

*How to Discount Small Probabilities**

ABSTRACT: Maximizing expected value leads to counterintuitive choices in cases that involve tiny probabilities of huge payoffs. In response to such cases, some have argued that we ought to discount very small probabilities down to zero. In this chapter, I discuss how exactly this view can be formulated. I begin by showing that less plausible versions of discounting small probabilities violate dominance. Then, I show that more plausible formulations of this view avoid these dominance violations, but instead, they violate the axiom of Independence—and in a particularly counterintuitive way. As a result of this violation, those who discount small probabilities can be exploited by a money pump. Lastly, I discuss one possible way of avoiding exploitation by this money pump.

Orthodox decision theory claims that a rational agent always maximizes expected utility. However, this seems to imply counterintuitive choices in cases that involve very small probabilities of huge payoffs. In these cases, an option can be great in

*I wish to thank Tomi Francis, Andreas Mogensen, Teruji Thomas and the audience of the Global Priorities Institute seminar for valuable feedback and discussions.

expectation, even if the probability of obtaining a valuable outcome is tiny, as long as this valuable outcome is great enough. One example of such a case is *Pascal's Mugging*:[1]

> **Pascal's Mugging:** A stranger approaches Pascal and claims to be an Operator from the Seventh Dimension. He promises to perform magic that will give Pascal an extra thousand quadrillion happy days in the Seventh Dimension if he pays the mugger ten livres—money that the mugger will use for helping very many orphans in the Seventh Dimension.

Pascal thinks that the probability of the mugger telling the truth is very low. However, the potential payoff is so high that Pascal is forced to conclude that the expected utility of paying the mugger is positive. Furthermore, if Pascal gives a non-zero probability to the proposition that the mugger can reward him with any finite amount of utility, then the mugger can always increase the payoff until the offer has positive expected utility.[2] Consequently, maximizing expected utility (with unbounded utilities) requires paying the mugger—which seems counterintuitive.[3]

Another case that involves tiny probabilities of huge payoffs is the St. Petersburg game, a version of which was originally proposed by Nicolaus Bernoulli.[4] This

[1] Bostrom (2009). This case is based on informal discussions by various people, including Eliezer Yudkowsky (2007*b*).

[2] Contrary to this, see Hanson (2007), Yudkowsky (2007*a*) and Baumann (2009).

[3] This may not hold if utilities are bounded as standard axiomatizations of expected utility maximization (such as the von Neumann-Morgenstern utility theorem) require. See Kreps (1988, p. 63).

[4] The game was simplified by Gabriel Cramer in 1728 and published by Daniel Bernoulli in 1738. See Pulskamp (n.d.) and Bernoulli (1954).

game is played by flipping a fair coin until it lands on heads. The prize of this game is $\$2^n$, where n is the number of coin flips. This game has infinite expected monetary value, so agents who maximize expected monetary value would pay any finite amount to play the game. However, this seems counterintuitive.[5] Furthermore, if this game's (monetary) value is infinite, one would value it higher than any of its possible finite payoffs, which seems irrational.[6]

In response to cases like this, some have argued that we ought to discount very small probabilities down to zero—let's call this *Probability Discounting*. Nicolaus Bernoulli first proposed this idea in response to the St. Petersburg game. He writes: "[T]he cases which have a very small probability must be neglected and counted for nulls, although they can give a very great expectation. […] This is a remark which merits to be well examined."[7] Recently, Smith (2014) and Monton (2019) have also defended the idea of Probability Discounting. Monton (2019) argues that one ought to discount very small probabilities down to zero, while Smith (2014) argues that it is rationally permissible—but not required—to do so.[8] However, we

[5]Pulskamp (n.d., p. 6). Daniel Bernoulli (cousin of Nicolaus Bernoulli) argues that, due to the diminishing marginal utility of money, one should not pay any finite sum to play the St. Petersburg game. See Bernoulli (1954). However, one can change the game slightly to bypass this objection by changing the prize from money to something with no diminishing marginal utility, such as perhaps days of life. See Monton (2019, p. 2). Relatedly, Menger (1967, pp. 217–218) shows that if utilities are unbounded, one can always create a *Super St-Petersburg game*, in which the payoffs grow sufficiently fast so that the expected utility of the game is infinite. See also Samuelson (1977, §2).

[6]Huemer (2016, pp. 34–35) and Russell and Isaacs (2021).

[7]Pulskamp (n.d., p. 2). Other proponents of Probability Discounting include, for example, Buffon and Condorcet. See Hey et al. (2010) and Monton (2019, pp. 16–17).

[8]Smith argues that discounting small probabilities down to zero is a way of getting a unique expected value for the Pasadena game. See Nover and Hájek (2004). See Hájek (2014) and Isaacs (2016) for criticism of discounting small probabilities. Also, see Beckstead (2013, ch. 6), Beckstead and Thomas (2020), Goodsell (2021), Russell and Isaacs (2021) and Russell (2021) for discussions

do not yet have a well-specified and plausible theory that tells us how to discount small probabilities. As Monton writes: "I don't have a perfectly rational, reasonable decision theory to hand you just yet (sorry)."[9]

This chapter discusses how Probability Discounting can be formulated and what the most plausible version of it might look like. §1 discusses a simple version of Probability Discounting on which one should conditionalize on outcomes associated with tiny probabilities not occurring. I show that this view faces a problem with individuating outcomes, and it also violates Statewise Dominance. §2 discusses a version of Probability Discounting that considers very-small-probability outcomes as tiebreakers when prospects would otherwise be equally good. I show that this view also violates Statewise Dominance. §3 discusses a version of Probability Discounting on which one should conditionalize on very-small-probability *states* not occurring. I discuss three ways of specifying this view. I show that one violates Stochastic Dominance and Acyclicity within choice sets, another violates Pairwise Acyclicity, Contraction and Expansion Consistency and Stochastic Dominance, and the last violates Statewise Dominance. §4 discusses more plausible versions of Probability Discounting that avoid the earlier violations of dominance and Acyclicity. §5 shows that these views violate the axiom of Independence—and

---

of related issues, and see Wilkinson (2022) for a defense of *Probability Fanaticism*:

> **Probability Fanaticism:** For any probability $p > 0$ and any finite utility u, there is some large enough utility U such that probability p of U (and otherwise nothing) is better than certainty of u.

In this context, 'otherwise nothing' means retaining the status quo or baseline outcome.

[9] Monton (2019, p. 15).

in a particularly counterintuitive way. As a result of this violation, those who discount small probabilities are vulnerable to exploitation by a money pump for Independence.[10] Lastly, §6 discusses one possible way of avoiding exploitation by this money pump. I conclude that Probability Discounting faces significant problems that undermine its plausibility as a theory of instrumental rationality.

1 Naive Discounting

This section discusses a version of Probability Discounting on which one should conditionalize on outcomes associated with tiny probabilities not occurring. However, I show that this view faces the *Outcome Individuation Problem*, and it also violates Statewise Dominance. Therefore, it is implausible as a theory of instrumental rationality.

According to Probability Discounting, an agent is rationally required or permitted to discount very small probabilities down to zero. On this view, there is some discounting threshold t such that probabilities below this threshold are discounted down to zero, but probabilities at least as great as this threshold are not discounted.[11,12] But when are probabilities small enough to be discounted? Or, as

[10] Isaacs (2016) also presents a problem for Probability Discounting in a dynamic context, to which Smith (2016) and Monton (2019) respond by arguing that relevantly similar choices ought to be evaluated collectively. This response is not available in the Independence Money Pump I will discuss later.

[11] Alternatively, this threshold probability t and probabilities below it are discounted, while the probabilities above t are not discounted. Note that this threshold might also be vague.

[12] Smith (2014) holds that the threshold might not apply to simple prospects, that is, prospects that assign a non-zero probability to only finitely many outcomes. Also, Smith does not argue for one universal threshold applicable in all situations. Instead, he maintains that this threshold may

Buffon writes: "[O]ne can feel that it is a certain number of probabilities that equals the moral certainty, but what number is it?"[13] Some possible discounting thresholds have been suggested. For Buffon and Condorcet, the discounting thresholds were 1/10,000 and 1/144,768 (respectively), while for Monton, this threshold is approximately 1 in 2 quadrillion.[14] As Monton argues, the discounting threshold is plausibly subjective. There is no objective answer to Buffon's question. Instead, it is up to each individual where the discounting threshold is.[15]

So, on this view, one should discount small probabilities—but small probabilities of *what*? This chapter discusses versions of Probability Discounting that ignore very-small-probability outcomes or states.[16] I will begin with the former views. There are many ways of ignoring outcomes associated with small probabilities. One way to ignore the very-small-probability outcomes of some prospect $\mathcal{P}_1$ would be to treat $\mathcal{P}_1$ as interchangeable with a prospect $\mathcal{P}_2$, which really does

be different in different situations.

[13]Hey et al. (2010, p. 256).

[14]Buffon's discounting threshold was the probability of a 56-year-old man dying in 24 hours—an outcome reasonable people typically ignore. See Monton (2019, pp. 8–9). Condorcet's discounting threshold was the difference between the probability that a 47-year-old man would die in one day and the probability that a 37-year-old man would. See Monton (2019, pp. 16–17). Monton's discounting threshold is between $1/2^{50}$ and $1/2^{51}$, as he treats the probability of getting tails at least 50 times in a row (with a fair coin) as rationally negligible. Monton (2019, p. 17).

[15]The subjectivity of the discounting threshold may be reasonable for individuals' rational preferences. However, it seems less so in the context of ethics when we are asking which prospects are better or worse.

[16]Whether one ignores very-small-probability outcomes or states makes a difference in some cases; a very-small-probability state might result in some outcome that overall has a non-negligible probability (when one also considers the other states). If one ignores very-small-probability states, one would discount down to zero (or at least decrease) the probability of this outcome. In contrast, if one ignores very-small-probability outcomes, one would not discount down to zero nor decrease the probability of this outcome.

assign probability zero to these outcomes.[17] However, $\mathcal{P}_2$ cannot assign the same probabilities as $\mathcal{P}_1$ to the remaining outcomes. Otherwise, the sum of all the probabilities assigned to outcomes of $\mathcal{P}_2$ would be less than one.[18] Instead, the probabilities assigned by $\mathcal{P}_2$ can be obtained from those assigned by $\mathcal{P}_1$ by conditionalizing on the supposition that some outcome of non-negligible probability occurs, where 'non-negligible' means a probability that is at least as great as the discounting threshold.[19]

Let $X \succsim Y$ mean that X is at least as preferred as Y. Also, let $EU(X)_{pd}$ denote the expected utility of prospect X when tiny probabilities have been discounted down to zero (read as 'the probability-discounted expected utility of X', where 'pd' stands for 'probability-discounted'). A *prospect* is taken to be a situation that may result in different outcomes with different probabilities. Then, one of the simplest versions of Probability Discounting—let's call it *Naive Discounting*—states:

> **Naive Discounting:** For all prospects X and Y, $X \succsim Y$ if and only if $EU(X)_{pd} \geq EU(Y)_{pd}$, where $EU(X)_{pd}$ and $EU(Y)_{pd}$ are obtained by conditionalizing on the supposition that some outcome of non-negligible probability occurs.[20]

On Naive Discounting, one should conditionalize on very-small-probability outcomes not occurring—but what counts as an 'outcome'? In particular, Naive

[17] Smith (2014, p. 478).

[18] Smith (2014, p. 478).

[19] Smith (2014, p. 478).

[20] Note that $EU(X)_{pd}$ and $EU(Y)_{pd}$ are obtained by conditionalizing, potentially, on different events not occurring.

Discounting faces the following problem:[21]

> **Outcome Individuation Problem:** If we individuate outcomes with too much detail, all outcomes have negligible probabilities. Is there a privileged way of individuating outcomes that avoids this?

The most obvious non-arbitrary way of individuating outcomes is by their utilities:[22]

> **Individuation by Preference:** Outcomes should be distinguished as different if and only if one has a preference between them.

Following this principle, each final utility level that a prospect might result in is considered a distinct outcome, and the possibilities of these outcomes are ignored if their associated probabilities are below the discounting threshold. For example, this view recommends against paying the mugger in Pascal's Mugging because the probability of obtaining an outcome as good as a thousand quadrillion happy days is very unlikely.

However, individuating outcomes by their utilities might result in ignoring all possible outcomes of some prospect if all its final utility levels are very unlikely.

[21] See also Beckstead and Thomas (2020, p. 13).

[22] If an agent is indifferent between winning $1 and eating an apple, on this view these would be considered the same outcome. Suppose the total probability of winning $1 and eating an apple is above the discounting threshold. In that case, these possibilities are not ignored, even if both winning $1 and eating an apple have negligible probabilities. Contrast Individuation by Preference with a similar principle presented by Broome (1991, p. 103):

> **Principle of Individuation by Justifiers:** Outcomes should be distinguished as different if and only if they differ in a way that makes it rational to have a preference between them.

In response to such cases, agents might lower their discounting thresholds until at least some outcomes have non-negligible probabilities. However, in cases where all outcomes have a zero probability, it is not possible to do so (except, of course, by not discounting at all).[23] Imagine, for example, an ideally shaped dart thrown on a dartboard, where each point results in a different utility. The probability that the dart hits a particular point may be zero. But one should not ignore every possible outcome of throwing the dart. Nevertheless, one might argue that we need not worry about cases where all outcomes have a zero probability because they are rare in practice. In all (or near all) cases we care about, some outcomes have non-zero probabilities.

Some might be satisfied with the above solution to the *Outcome Individuation Problem*. However, besides this problem, Naive Discounting also violates dominance. Let $X \succ Y$ mean that X is strictly preferred (or simply 'preferred') to Y. Then, Naive Discounting violates the following dominance principle:[24]

> **Statewise Dominance:** If the outcome of prospect X is at least as preferred as the outcome of prospect Y in all states, then $X \succsim Y$. Furthermore, if in addition the outcome of X is strictly preferred to the outcome of Y in some possible state, then $X \succ Y$.

Statewise Dominance is very plausible.[25] If some prospect is sure to turn out at

[23] Beckstead and Thomas (2020, pp. 12–13).

[24] Savage (1951, p. 58) and Luce and Raiffa (1957, p. 287).

[25] Russell (2021, p. 13) writes on (strict) Statewise Dominance: "What if Statewise Dominance fails? In that case, I'm not sure what we're doing when we compare how good prospects are. […] [W]hat we ultimately care about is how well things turn out; choosing better prospects is supposed

least as well as another prospect, but it might turn out better, then that prospect should be better.

To see why Naive Discounting violates Statewise Dominance, consider the following prospects (see table 1):[26]

Naive Statewise Dominance Violation:

Prospect A Gives \$1,000,000 in state 1 and nothing in state 2.

Prospect B Gives nothing in both states.

Suppose the probability of state 1 is below the discounting threshold. After conditionalizing on the supposition that some outcome of non-negligible probability occurs, A is substituted by B. One would then be indifferent between A and B, even though the outcomes of A and B are equally good in state 2, but the outcome of A is better than the outcome of B in state 1. Therefore, Naive Discounting violates Statewise Dominance.

Table 1
A Violation of Statewise Dominance

	State 1	**State 2**
	$p <$ threshold	$1 - p$
A	\$1,000,000	\$0
B	\$0	\$0

to guide us toward achieving better outcomes. In light of this, if dominance reasoning is wrong, then I don't want to be right. If A is sure to turn out better than B, then this tells us precisely the thing that betterness-of-prospects is supposed to be a guide to."

[26] Monton (2019, pp. 20–21) discusses a similar dominance violation. He proposes that Probability Discounting be supplemented with dominance. On discounting small probabilities and dominance violations, also see Isaacs (2016), Smith (2016), Lundgren and Stefánsson (2020, pp. 912–914) and Beckstead and Thomas (2020, §2.3).

To summarize, Naive Discounting states that one should conditionalize on not obtaining very-small-probability outcomes. This view faces the *Outcome Individuation Problem*, which can be solved by individuating outcomes by their utilities (except in cases where all outcomes have a zero probability). However, Naive Discounting also faces another problem: It violates Statewise Dominance. This undermines its plausibility as a theory of instrumental rationality.[27]

2 Lexical Discounting

This section discusses a version of Probability Discounting that treats very-small-probability outcomes as tiebreakers when prospects would otherwise be equally good. This view avoids the previous violation of Statewise Dominance. However, I show that it violates Statewise Dominance in another case.

There is a straightforward solution to the previous case. Probability Discounting can avoid the earlier violation of Statewise Dominance if outcomes whose probabilities are below the discounting threshold are treated as tiebreakers. Then, A is better than B because A and B have equal probability-discounted expected utility but, in addition, A gives a negligible probability of a positive outcome (while B does not). More generally, in tied cases, prospects can be compared by their ex-

[27] As shown in Chapter 2 of this thesis, Expected Utility Theory also violates Statewise Dominance, on pain of violating Continuity. Monton (2019, §7) argues that violations of Statewise Dominance should not count against Probability Discounting, given that Expected Utility Theory violates Statewise Dominance too. Later in §4, I discuss versions of Probability Discounting that do not violate Statewise Dominance.

pected utilities without any discounting (like Expected Utility Theory would do).

On this proposal, prospects are first ranked by their probability-discounted expected utilities. Then, in cases of ties, these prospects are ranked by their expected utilities without discounting small probabilities. Formally this view—let's call it *Lexical Discounting*—states the following:

> **Lexical Discounting:** For all prospects X and Y, $X \succsim Y$ if and only if
>
> - $EU(X)_{pd} > EU(Y)_{pd}$ or
> - $EU(X)_{pd} = EU(Y)_{pd}$ and $EU(X) \geq EU(Y)$,
>
> where $EU(X)_{pd}$ and $EU(Y)_{pd}$ are obtained by conditionalizing on the supposition that some outcome of non-negligible probability occurs.[28]

It is slightly misleading to say that Lexical Discounting is a form of discounting small probabilities down to zero because small probabilities and their associated utilities are considered in cases of ties. The outcomes whose probabilities are (at and) above the discounting threshold just take lexical priority over the very-small-probability outcomes.[29]

[28] As before, note that $EU(X)_{pd}$ and $EU(Y)_{pd}$ are obtained by conditionalizing, potentially, on different events not occurring.

[29] It might be argued that because some small probabilities are much smaller than others, one should have multiple discounting thresholds that form probability ranges, where higher probability ranges take lexical priority over the lower ones. Beckstead and Thomas (2020, p. 24 n. 19) point out that Probability Discounting faces some of the same problems as Probability Fanaticism if it uses very-small-probability outcomes as tiebreakers. Having multiple discounting thresholds may help probability discounters avoid these problems. For brevity, I will only discuss views on which there is just one discounting threshold.

However, Lexical Discounting also violates Statewise Dominance. To see how this violation happens, consider the following case (table 2):

Lexical Statewise Dominance Violation:

Prospect A Gives \$10 in states 1 and 2, \$100 in state 3, and nothing in state 4.

Prospect B Gives \$10 in state 1, \$100 in states 2 and 3, and nothing in state 4.

The only difference between these prospects is that A gives \$10 in state 2, while B gives \$100 in that same state. The probability of states 1 and 4 is 0.49, and the probability of states 2 and 3 is 0.01. For simplicity, let the discounting threshold be (implausibly) 0.03. Then, all probabilities below 0.03 should be discounted down to zero, while probabilities at least as great as 0.03 should not be discounted down to zero. Let's also assume that the utility of money equals the monetary amount.

Table 2
A Violation of Statewise Dominance

	State 1	**State 2**	**State 3**	**State 4**
p	0.49	0.01	0.01	0.49
A	\$10	\$10	\$100	\$0
B	\$10	\$100	\$100	\$0

In this case, A gives a 0.5 probability of \$10 and a 0.01 probability of \$100 (and otherwise nothing). So, A's probability-discounted expected utility is $EU(A)_{pd} \approx 5.05$ after conditionalizing on not obtaining \$100 (as its associated probability is

below the discounting threshold).[30] B in turn gives a 0.49 probability of \$10 and a 0.02 probability of \$100 (and otherwise nothing). B's probability-discounted expected utility is $EU(B)_{pd} = 5$ after conditionalizing on not obtaining \$100 with it.[31] Given that the former is greater than the latter, A is better than B according to Lexical Discounting. However, as mentioned above, the only difference between A and B is that A gives \$10 in state 2, while B gives \$100 in that same state. Therefore, Lexical Discounting—too—violates Statewise Dominance.

This violation of Statewise Dominance happens because when one conditionalizes on not obtaining \$100 with A (state 3), the probability of state 3 is divided between states 1, 2 and 4. However, when one conditionalizes on not obtaining \$100 with B (states 2 and 3), the probability of states 2 and 3 is divided between states 1 and 4. Therefore, the probability of obtaining nothing is greater with B than with A after ignoring the possibility of obtaining \$100.

To summarize, Lexical Discounting states that outcomes whose probabilities are (at or) above the discounting threshold take lexical priority over very-small-probability outcomes in determining prospects' betterness ranking—very-small-probability outcomes are only treated as tiebreakers. However, like Naive Discounting, Lexical Discounting also violates Statewise Dominance. This makes it a less plausible candidate for a theory of instrumental rationality.

[30] $0.5/(1-0.01)\cdot 10 \approx 5.05$.

[31] $0.49/(1-0.02)\cdot 10 = 5$.

3 State Discounting

This section discusses a version of Probability Discounting on which one should conditionalize on very-small-probability states not occurring. Three versions of this view are presented. I show that one violates Stochastic Dominance and Acyclicity within choice sets, another violates Pairwise Acyclicity, Contraction and Expansion Consistency and Stochastic Dominance, and the last one violates Statewise Dominance.

3.1 Pairwise and Set-Dependent State Discounting

Again, there is a straightforward solution to the previous violation of Statewise Dominance. Earlier it was assumed that one should ignore (except in cases of ties) the possibility of obtaining outcomes associated with tiny probabilities. However, one might instead ignore very-small-probability *states*—call this view *State Discounting*.[32] One can also make a lexical version of this view:

> **State Discounting** For all prospects X and Y, $X \succsim Y$ if and only if
>
> - $EU(X)_{pd} > EU(Y)_{pd}$ or
> - $EU(X)_{pd} = EU(Y)_{pd}$ and $EU(X) \geq EU(Y)$,
>
> where $EU(X)_{pd}$ and $EU(Y)_{pd}$ are obtained by conditionalizing on the supposition that no state of negligible probability occurs.

[32] This view naturally captures the idea that one should ignore very small *changes* in probabilities instead of very small (absolute) probabilities. Thus, it allows one to ignore the possibility of making a difference to some outcome if the probability of doing so is negligible. See §4 in Chapter 6 of this thesis.

State Discounting recommends against paying the mugger in Pascal's Mugging because the state in which the mugger delivers a thousand quadrillion happy days is very unlikely to occur. In the previous violation of Statewise Dominance, State Discounting tells one to ignore states 2 and 3 as their associated probabilities are below the discounting threshold. Consequently, A and B have equal probability-discounted expected utility (as they give the same outcomes in states 1 and 4). However, B has greater expected utility without discounting, so it is better than A (assuming a lexical version of State Discounting). Thus, State Discounting avoids the previous violation of Statewise Dominance.[33]

However, notice that State Discounting faces an analogous problem to the *Outcome Individuation Problem*, namely, the

> **State Individuation Problem:** If one individuates states with too much detail, all states have negligible probabilities. Is there a privileged way of individuating states that avoids this?

As before, a possible solution is to individuate states by the utilities of their outcomes.[34]

There are different views about how states should be partitioned. On another version of State Discounting, prospects are always compared two at a time, and the possible states of the world are partitioned for every pairwise comparison sepa-

[33] However, as I will show later, one version of State Discounting violates Statewise Dominance in this case.

[34] As before, one problem with this is that, in some cases, all states might have probabilities below the discounting threshold. One could lower the threshold in such cases. However, this will not solve the problem in cases where all states have a zero probability.

rately. Alternatively, one could compare all available options at once and partition the states for every choice set separately. Let's call these views *Pairwise State Discounting* and *Set-Dependent State Discounting*, respectively (the difference between these views is illustrated with an example later).

> **Pairwise State Discounting:** States are partitioned by comparing two prospects at a time.

> **Set-Dependent State Discounting:** States are partitioned by comparing all available prospects at once.

Although these views avoid the earlier violations of Statewise Dominance, they violate the following principle instead:

> **Acyclicity:** If $X_1 \succ X_2 \succ \cdots \succ X_n$, then it is not the case that $X_n \succ X_1$.

To see why these views violate Acyclicity, consider the following case:

> **Acyclicity Violation:** A random number generator returns a number between 1 and 100.
>
> *Prospect A* Gives \$1000 with numbers 1 and 2 (probability 0.02); otherwise, it gives nothing.
>
> *Prospect B* Certainly gives \$10 no matter what number comes up.
>
> *Prospect C* Gives \$1000 with number 1 (probability 0.01) and otherwise it gives \$1.

Let the discounting threshold be 0.02. First, compare A and B. Individuating states by the utilities of their outcomes results in two states as shown in table 3. A is better than B because neither state has a non-negligible probability, and A's expected utility is greater than that of B.[35] Next, compare B and C. In this case, individuating states by the utilities of their outcomes results in states shown in table 4. As the probability of state 1* is below the discounting threshold, one should ignore the possibility of state 1* occurring. Once one does that, B is better than C, as it gives a better outcome in state 2* ($10 vs. $1).

Table 3
A is better than B

	State 1	**State 2**
Output	1 or 2 (p=0.02)	3 to 100 (p=0.98)
A	$1000	$0
B	$10	$10

Table 4
B is better than C

	State 1*	**State 2***
Output	1 (p=0.01)	2 to 100 (p=0.99)
B	$10	$10
C	$1000	$1

Now we have that A is better than B, which is better than C. It follows by Acyclicity that C is not better than A. However, when we compare A and C pairwise, we notice that C is better than A. In this case, individuating states by the utilities of their outcomes results in states shown in table 5. As states 1** and 2** have probabilities below the discounting threshold, the agent should ignore the possibilities of these states. Moreover, when the agent does that, C is better than A because it gives a better outcome in state 3**. So, we have a violation of Acyclicity: A is better than B, which is better than C, which is better than A.

[35] $EU(A)_{pd} = 0.02 \cdot 1000 = 20$ and $EU(B)_{pd} = 10$.

TABLE 5

C IS BETTER THAN A

	State 1**	**State 2****	**State 3****
Output	1 (p=0.01)	2 (p=0.01)	3 to 100 (p=0.98)
A	\$1000	\$1000	\$0
C	\$1000	\$1	\$1

Let's now go back to Pairwise and Baseline State Discounting. If we partition states for each pair of options in a way that depends on the particular two options being compared (in line with Pairwise State Discounting), then State Discounting violates Acyclicity within choice sets. Consequently, it is not clear what one ought to choose when all A, B and C are available, as there is no most-preferred alternative.[36] However, if we partition states in a way that depends on the overall choice set (in line with Set-Dependent State Discounting), then there is no violation of Acyclicity within choice sets (see table 6). In this case, states 1*** and 2*** have probabilities below the discounting threshold, so one should ignore the possibilities of these states. And when one does that, B is the best prospect as it gives the best outcome in state 3***, and C is the second-best prospect as it gives a better outcome than A in that state.

However, Set-Dependent State Discounting violates Acyclicity across choice sets (as shown in tables 3, 4 and 5). In particular, it was shown that Set-Dependent State Discounting violates Pairwise Acyclicity, that is, it violates Acyclicity when we compare two options at a time (when each choice set only includes two options).

[36] Fishburn (1991, p. 116).

TABLE 6

NO VIOLATION OF ACYCLICITY

	State 1* **	**State 2*** **	**State 3*** **
Output	1 (p=0.01)	2 (p=0.01)	3 to 100 (p=0.98)
A	\$1000	\$1000	\$0
B	\$10	\$10	\$10
C	\$1000	\$1	\$1

It is odd that adding or removing options can influence which events one ignores. For example, when comparing A and B, Set-Dependent State Discounting does not ignore the possibility of the random number generator outputting number 1 or 2. However, when C is also available, Set-Dependent State Discounting ignores these possibilities. Consequently, the value of A decreases significantly when C is also available, as one then ignores the possibility of obtaining \$1000 with A.

This case shows that Set-Dependent State Discounting violates the following principles that many find plausible:[37]

> **Contraction Consistency:** For all prospects X and Y, if it is permissible to choose X from the set $\{X, \dots, Y\}$, then it is permissible to choose X from any subset of the set $\{X, \dots, Y\}$.

[37] Sen (1977, pp. 63–66). More generally, Contraction Consistency implies Acyclicity. Suppose that one violates Acyclicity. Then, there is a sequence of prospects such that $X_1 \succ X_2 \succ \cdots \succ X_n \succ X_1$. Suppose that some prospect X_i is chosen from the choice set that includes all these prospects. Next, consider the choice set that includes only X_i and X_{i-1} (if $X_i = X_1$, then this choice set includes X_1 and X_n). Given that $X_{i-1} \succ X_i$, one would now choose X_{i-1} (or X_n if $X_i = X_1$). This is a violation of Contraction Consistency. Thus, if a view does not violate Contraction Consistency, then it does not violate Acyclicity. See Sen (1977, p. 67).

Strong Expansion Consistency: For all prospects X, Y and Z, if it is permissible to choose X from the set $\{X, \ldots, Y\}$, then if it is permissible to choose Y from the set $\{X, \ldots, Y, \ldots, Z\}$, it is permissible to choose X from the set $\{X, \ldots, Y, \ldots, Z\}$.

Set-Dependent State Discounting violates Contraction Consistency as it is permissible (indeed rationally required) to choose B when all A, B and C are options. However, when only A and B are options, it is no longer permissible to choose B (because then one is rationally required to choose A as one no longer ignores the possibility of obtaining \$1000 with A). On the other hand, Set-Dependent State Discounting violates Strong Expansion Consistency because it is permissible (indeed rationally required) to choose A when only A and B are options. But when A, B and C are options, it is permissible to choose B but not permissible to choose A.

Next, let $X = \{x_1,\ p_1; \ldots;\ x_n,\ p_n\}$ stand for prospect X that gives non-zero probabilities p_1, p_2, ..., p_n of outcomes x_1, x_2, ..., x_n. Then, both versions of State Discounting violate the following principle:[38]

Stochastic Dominance: Prospect $X = \{x_1,\ p_1; \ldots;\ x_n,\ p_n\}$ is pre-

[38] Buchak (2013, p. 42). More precisely, the definition given here is for *first-order stochastic dominance*, an idea that was introduced to statistics by Mann and Whitney (1947) and Lehmann (1955), and to economics by Quirk and Saposnik (1962). The name 'first-degree stochastic dominance' is due to Hadar and Russell (1969, p. 27).

ferred to prospect $Y = \{y_1,\ q_1; \ldots;\ y_n,\ q_n\}$ if, for all outcomes o,

$$\sum_{\{i \mid x_i \succsim o\}} p_i \geq \sum_{\{j \mid y_j \succsim o\}} q_j,$$

and for some outcome u,

$$\sum_{\{i \mid x_i \succsim u\}} p_i > \sum_{\{j \mid y_j \succsim u\}} q_j.$$

A violation of Stochastic Dominance happens if, for all outcomes, some prospect X gives an at least as high probability of an at least as great outcome as some other prospect Y does, and for some outcome, X gives a greater probability of an at least as great outcome as Y does—yet Y is judged better than or equally as good as X.

To see why both versions of State Discounting violate Stochastic Dominance, consider the following case:

Two Coins:

Prospect A Gives \$10 if a coin lands on heads (probability 0.5), nothing if it lands on tails (probability 0.49), and \$100 if it lands on the edge (probability 0.01).

Prospect B Gives \$10 if another coin lands on heads (probability 0.49), nothing if it lands on tails (probability 0.49), and \$100 if it lands on the edge (probability 0.02).

Let the discounting threshold be 0.03. These prospects give the same probabilities

of the same outcomes as the prospects in *Lexical Statewise Dominance Violation* (table 2), but instead of four states, we now have nine different states due to having two coins. Let 'H' stand for 'heads', 'T' for 'tails' and 'E' for 'edge'. Also, let '(X, Y)' stand for the first coin landing on 'X' and the second one on 'Y'. All states in which either coin lands on the edge have probabilities below the discounting threshold (given that the probabilities of either coin landing on the edge are alone already below the discounting threshold). Only four states have probabilities above the discounting threshold: (H, H), (H, T), (T, H) and (T, T) (see table 7).

TABLE 7
A VIOLATION OF STOCHASTIC DOMINANCE

	State 1	**State 2**	**State 3**	**State 4**
	H, H	H, T	T, H	T, T
$p^\star$	*0.253*	*0.253*	*0.247*	*0.247*
A	\$10	\$10	\$0	\$0
B	\$10	\$0	\$10	\$0

$p^\star$=probability conditional on one of states 1, 2, 3 or 4 occurring.

After conditionalizing on one of these four states occurring, the probability-discounted expected utility of A is greater than that of B: Now the only difference between these prospects is that A gives \$10 in state 2 (and nothing in state 3), while B gives \$10 in state 3 (and nothing in state 2), and state 2 has a greater probability than state 3.[39] Thus, A is better than B according to both versions of State Discounting. However, this is a violation of Stochastic Dominance. Before dis-

[39] A's probability-discounted expected utility is $EU(A)_{pd} \approx 5.05$. B's probability-discounted expected utility, in turn, is $EU(B)_{pd} = 5$.

counting, both A and B give a 0.51 probability of at least \$10, but B gives a greater probability of at least \$100 (0.02 vs. 0.01). So, for all outcomes, B gives an at least as high probability of an at least as great outcome as A does, and for some outcome, B gives a greater probability of an at least as great outcome as A does. Thus, B stochastically dominates A, and Pairwise and Set-Dependent State Discounting violate Stochastic Dominance as they claim that A is better than B.[40]

3.2 Baseline State Discounting

According to the previous versions of State Discounting, states might be partitioned differently depending on what other options are available. This leads to a violation of Acyclicity. However, states might also be partitioned in a way that does not depend on the other available options. This can be done by comparing each prospect to some baseline or status quo prospect—let's call this *Baseline State Discounting*.

> **Baseline State Discounting:** States are partitioned by comparing every prospect to a status quo prospect (each separately).[41]

[40] Note that the prospects in *Two Coins* are stochastically equivalent with the earlier prospects in *Lexical Statewise Dominance Violation*; both prospects give the same probabilities of the same outcomes. However, both Pairwise and Set-Dependent State Discounting judged B as better than A in the earlier case, but A as better than B in this case. Thus, on these views, the probabilities and the utilities associated with them are not the only determinants of the value of prospects. It also matters which states result in the different outcomes and what the probabilities of those states are. In general, Stochastic Equivalence and Statewise Dominance imply Stochastic Dominance. See Russell (2021, §2).

[41] On Baseline State Discounting, one might sometimes ignore some events e_1 and e_2 when comparing some prospect X to the status quo prospect, but not ignore them when comparing another prospect Y to the status quo prospect. This can happen if both the status quo prospect and prospect Y result in the same outcome with e_1 as with e_2, but prospect X results in a different

According to this view, one should ignore the very-small-probability states of some prospect X when states are partitioned by comparing X to a baseline or status quo prospect, which corresponds to doing nothing.

However, Baseline State Discounting violates Statewise Dominance in the same way as Lexical Discounting does. Consider again *Lexical Statewise Dominance Violation* (table 2). This time, let's specify the events that result in each outcome:

> **Random Number:** A random number generator returns a number between 1 and 100.
>
> *Prospect A* Gives \$10 if a number between 1 and 50 is returned, \$100 if number 51 is returned, and nothing if a number between 52 and 100 is returned.
>
> *Prospect B* Gives \$10 if a number between 1 and 49 is returned, \$100 if 50 or 51 is returned, and nothing if a number between 52 and 100 is returned.

In this case, the baseline prospect is (presumably) certainly getting nothing. When A is compared to this baseline prospect, state individuation by utilities results in three states as shown in table 8. As the probability of state 2 is below the discounting threshold of 0.03, the possibility of this state is ignored. Consequently, the

outcome with e_1 than with e_2. Consequently, e_1 and e_2 result in two different states when prospect X is compared to the status quo, but only one state when Y is compared to the status quo. If the combined probability of these states is above the discounting threshold, but the probabilities of these states taken individually are below the discounting threshold, then e_1 and e_2 will get ignored with prospect X but not with prospect Y.

probability-discounted expected utility of A is $EU(A)_{pd} \approx 5.05$.[42]

TABLE 8
A VS. THE BASELINE

	State 1	**State 2**	**State 3**
Output	1–50 (p=0.5)	51 (p=0.01)	52–100 (p=0.49)
A	\$10	\$100	\$0
Baseline	\$0	\$0	\$0

Next, compare B to the baseline prospect. This time state individuation by utilities results in the three states shown in table 9. Again, the probability of state 2* is below the discounting threshold, so the possibility of this state is ignored. Then, the probability-discounted expected utility of B is $EU(B)_{pd} = 5$.[43]

TABLE 9
B VS. THE BASELINE

	State 1*	**State 2***	**State 3***
Output	1–49 (p=0.49)	50 or 51 (p=0.02)	52–100 (p=0.49)
B	\$10	\$100	\$0
Baseline	\$0	\$0	\$0

A's probability-discounted expected utility is greater than that of B (5.05 vs. 5), so A is better than B. However, B statewise dominates A because the only difference between these prospects is that A gives \$10 if the random number generator returns the number 50, while B gives \$100 in that case. Consequently, Base-

[42] $EU(A)_{pd} = 0.5/0.99 \cdot 10 \approx 5.05$.
[43] $EU(B)_{pd} = 0.49/0.98 \cdot 10 = 5$.

line State Discounting violates Statewise Dominance when states are partitioned in the usual way corresponding to possible states of the world (such as 'number 50 is returned'). This violation of Statewise Dominance happens because the possible states of the world that Baseline State Discounting ignores are not the same for every prospect. For example, when comparing A to the baseline prospect, the possibility of the random number generator returning the number 50 is not ignored, but when B is compared to the baseline prospect, this possibility is ignored (tables 8 and 9).

To summarize, instead of ignoring very-small-probability outcomes, Probability Discounting might ignore very-small-probability states. State Discounting faces the *State Individuation Problem*, which can be solved by individuating states by the utilities of their outcomes. I have discussed three ways of formulating State Discounting. Pairwise State Discounting always compares two options at a time, even if the choice set includes other options as well. It ignores very-small-probability states in every pairwise comparison. However, Pairwise State Discounting violates Stochastic Dominance and Acyclicity within choice sets. Set-Dependent State Discounting compares all available options simultaneously and ignores very-small-probability states in every choice set. This view violates Pairwise Acyclicity, Contraction and Expansion Consistency and Stochastic Dominance. Finally, Baseline State Discounting ignores very-small-probability states of some prospect X, when states are partitioned by comparing X to a baseline prospect. This view violates Statewise (and hence also Stochastic) Dominance. To conclude, all three versions of State Discounting violate plausible principles of rationality, which undermines

their plausibility as theories of instrumental rationality.[44]

4 Stochastic and Tail Discounting

This section discusses more plausible versions of Probability Discounting that avoid the earlier violations of dominance (and Acyclicity). However, §5 shows that these views violate the axiom of Independence and are therefore vulnerable to exploitation by a money pump.

4.1 Stochastic Discounting

One version of Probability Discounting—let's call it *Absolutist Stochastic Discounting*—works like this: To obtain the probability-discounted expected utility of a prospect, first add the lowest possible positive utility, weighted by the probability of getting at least that much utility.[45] Next, add the difference between the lowest utility and the next lowest utility, weighted by the probability of getting at least the higher amount of utility. Then, add the difference between this utility and the next lowest utility, weighted by the probability of getting at least that much utility, and so on until the next probability is below the discounting threshold.[46] Then, ignore the

[44]Someone might adopt a view on which one should first filter one's options by Statewise and Stochastic Dominance and then choose following some version of Probability Discounting from amongst the remaining options. This view avoids the dominance violations, but it also seems *ad hoc*. However, some may find the benefit of a greater fit with our intuitions worth the cost in terms of simplicity.

[45]Note that this view requires an objective zero point on the utility-scale.

[46]This is similar to an alternative way of calculating the expected utility of a prospect discussed by Buchak (2014, p. 1100).

rest of the utility levels (whose probabilities are below the discounting threshold). Negative utilities are then treated similarly, and their expectation is summed with the expectation of positive utilities to obtain the value of a prospect. (This is equivalent to calculating the probability-discounted expected utility of a prospect in the same way as Expected Utility Theory would calculate expected utilities with the following exception: The greatest positive and negative utilities [whose utility levels have negligible cumulative probability] have been replaced by the greatest positive or negative utility whose utility level has a non-negligible cumulative probability [respectively for positive and negative utilities]).[47]

According to Absolutist Stochastic Discounting, there is an objective neutral level. On this view, one should ignore the possibility of very high or very low utility levels when the cumulative probability of ending up with at least or at most that much utility (respectively for positive and negative utilities) is negligible. This view recommends against paying the mugger in Pascal's Mugging if there is only a tiny probability that one will get an outcome at least as good as a thousand quadrillion happy days. However, it does not recommend against paying the mugger if there is a non-negligible probability of obtaining an outcome that is at least as great as a

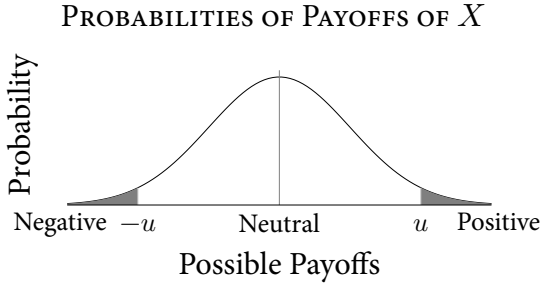
[47]That is, the following formula for calculating the probability-discounted expected utility of positive outcomes of prospect X is equivalent to the formulation given later (see *Positive outcomes*):

$$EU(X)_{pd,\ pos} = \sum_{i=1}^{m} p(E_i)u(x_i) + \Big(\sum_{i=m+1}^{n} p(E_i) \Big) u(x_m),$$

where x_m is the greatest positive utility that has a non-negligible cumulative probability, and x_n is the greatest positive utility possible with prospect X. (Negative utilities are treated similarly.)

thousand quadrillion happy days for some reason unrelated to the mugger's offer.[48]

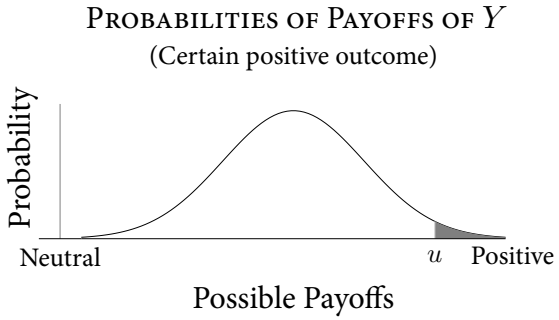
Suppose that some prospect X has possible outcomes whose values are normally distributed with a mean of zero (when the outcomes are ordered in terms of betterness). Absolutist Stochastic Discounting tells one to ignore the highest positive and the lowest negative utility levels of X. This is equivalent to substituting the values of the outcomes in the grey areas (see the graph below) with the values of u and $-u$ (respectively for positive and negative outcomes), where u and $-u$ are the best positive and the worst negative utility levels that have non-negligible cumulative probabilities.[49] For example, suppose X gives a negligible probability p of $u + \epsilon$ utility. Then, when calculating the probability-discounted expected utility of X, $u + \epsilon$ is substituted with u (that is, the contribution of a p chance of $u + \epsilon$ utility to the probability-discounted expected utility of X is $p \cdot u$.).

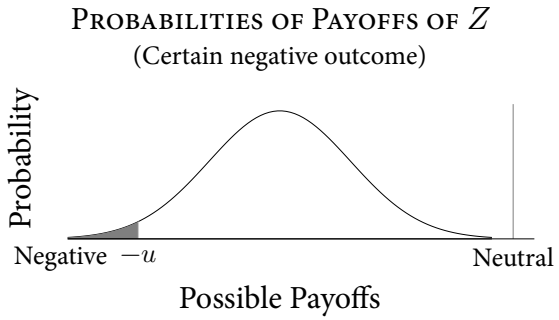


[48] For example, an agent who thinks there is a non-negligible probability of going to Heaven would not ignore the possibility of a great payoff in Pascal's Mugging. More generally, such an agent would not discount small probabilities very often (if ever); the non-negligible probability of going to Heaven makes it the case that there is a non-negligible probability of ending up with at least u amount of utility for all positive values of u.

[49] One is not simply ignoring the possibilities of the outcomes in the grey areas because their probabilities contribute to the cumulative probabilities of the utility levels with lower magnitudes.

The following prospects Y and Z can only result in positive or negative outcomes, respectively. Consequently, Absolutist Stochastic Discounting tells one to ignore Y's highest positive utility levels and Z's lowest negative utility levels. This is equivalent to substituting the values of the best positive outcomes of Y (the grey area in the left image below) with u and the values of the worst negative outcomes of Z (the grey area in the right image below) with $-u$ where u and $-u$ are the best positive and the worst negative utility levels of Y and Z (respectively) that have non-negligible cumulative probabilities.





Call the versions of Probability Discounting that have the same structure as Absolutist Stochastic Discounting *Stochastic Discounting*. Besides Absolutist Stochastic Discounting, there is another way of understanding Stochastic Discounting.

This view is similar to Baseline State Discounting because it compares each prospect to a baseline prospect—so it might be called *Baseline Stochastic Discounting*. On this view, one calculates the amount by which the baseline/status quo utility level is increased or decreased by the different possible outcomes of a prospect. Then, to obtain the probability-discounted expected utility of a prospect, one first adds the lowest possible gain (i.e., positive change to the baseline), weighted by the probability of gaining at least that much. Next, one adds the difference between the lowest gain and the next lowest gain, weighted by the probability of gaining at least the higher amount. Then, one adds the difference between this gain and the next lowest gain, weighted by the probability of gaining at least that much, and so on, until the next probability is below the discounting threshold. Then, one ignores the rest of the possible gains (whose probabilities are below the discounting threshold). Losses (i.e., negative changes to the baseline) are then treated similarly, and their expectation is summed with the expectation of gains to obtain the value of a prospect.[50]

Unlike the previous version of Stochastic Discounting, this view recommends against paying the mugger in Pascal's Mugging, even if there is a non-negligible probability of gaining an equally great payoff for some reason unrelated to the mugger's offer. This is so because once one has 'subtracted' the baseline prospect from

[50]One can also make a version of Stochastic Discounting that is analogous to Pairwise State Discounting in that it compares prospects to other available prospects pairwise—call this *Pairwise Stochastic Discounting*. On this view, one considers the utility difference in each state between two prospects and ignores the largest differences when the cumulative probability of states with differences at least that large is negligible. (Again, one does not entirely ignore these differences because the probabilities of these differences contribute to the cumulative probability of lesser differences.)

the mugger's offer, gains at least as great as a thousand quadrillion happy days have a negligible cumulative probability.

On both versions of Stochastic Discounting, the probability-discounted expected utility of positive outcomes is calculated as follows (here 'positive outcomes' are either gains if one accepts Baseline Stochastic Discounting or final utilities if one accepts Absolutist Stochastic Discounting):[51]

> **Positive outcomes:** For all prospects X, such that X gives non-zero probabilities of positive outcomes
> $X_{pos} = \{E_1,\ x_1;\ E_2,\ x_2; \ldots;\ E_m,\ x_m; \ldots;\ E_n,\ x_n\}$, and
> $0 < u(x_1) \leq \ldots \leq u(x_m) \leq \ldots \leq u(x_n)$, the probability-discounted expected utility of positive outcomes of X is

$$\begin{aligned} EU(X)_{pd,\ pos} = &\Big(\sum_{i=1}^{n} p(E_i)\Big)u(x_1) + \Big(\sum_{i=2}^{n} p(E_i)\Big)(u(x_2) - u(x_1)) \\ &+ \Big(\sum_{i=3}^{n} p(E_i)\Big)(u(x_3) - u(x_2)) \\ &+ \cdots + \Big(\sum_{i=m}^{n} p(E_i)\Big)(u(x_m) - u(x_{m-1})), \end{aligned}$$

[51] Technically, this formulation requires the following qualifications: If all probabilities of positive utility levels are non-negligible, then in order to obtain $EU(X)_{pd,\ pos}$, one simply sums up the positive utilities weighted by their probabilities (without discounting). And if all probabilities of positive utility levels are negligible, then $EU(X)_{pd,\ pos} = 0$ (on Baseline Stochastic Discounting) or the value of the baseline (on Absolutist Stochastic Discounting). Furthermore, this formula assumes that, amongst the possible positive utility levels, there is one that is the lowest. But this may not always be true. Consider for example a St. Petersburg-style prospect in which the possible payoff halves with each additional coin flip (i.e., it gives probability 1/2 of 2 utilities, probability 1/4 of one utility, probability 1/8 of 0.5 utilities and so on.) One can calculate the probability-discounted expected utility of such prospects as discussed on p. 159, that is, the same way as Expected Utility Theory would do with the following exception: The greatest positive utilities (whose utility levels have negligible cumulative probability) are replaced by the greatest positive utility whose utility level has a non-negligible cumulative probability.

where

$$\sum_{i=m}^{n} p(E_i) \geq t > \sum_{i=m+1}^{n} p(E_i),$$

where t is the discounting threshold.

The probability-discounted expected utility of prospect X is then obtained by summing the probability-discounted expected utilities of its positive and negative outcomes:[52]

$$EU(X)_{pd} = EU(X)_{pd,\ pos} + EU(X)_{pd,\ neg}.$$

Stochastic Discounting can use very-small-probability utility levels as tiebreak-

[52]The probability-discounted expected utility of negative outcomes is calculated as follows:

Negative outcomes: For all prospects X, such that X gives non-zero probabilities of negative outcomes $X_{neg} = \{E_{-1},\ x_{-1};\ E_{-2},\ x_{-2}; \ldots;\ E_{-m},\ x_{-m}; \ldots;\ E_{-n},\ x_{-n}\}$, and $0 > u(x_{-1}) \geq \ldots \geq u(x_{-m}) \geq \ldots \geq u(x_{-n})$, the probability-discounted expected utility of negative outcomes of X is

$$\begin{aligned} EU(X)_{pd,\ neg} = &\Big(\sum_{i=-1}^{-n} p(E_i)\Big)u(x_{-1}) + \Big(\sum_{i=-2}^{-n} p(E_i)\Big)(u(x_{-2}) - u(x_{-1})) \\ &+ \Big(\sum_{i=-3}^{-n} p(E_i)\Big)(u(x_{-3}) - u(x_{-2})) \\ &+ \cdots + \Big(\sum_{i=-m}^{-n} p(E_i)\Big)(u(x_{-m}) - u(x_{-m+1})), \end{aligned}$$

where

$$\sum_{i=-m}^{-n} p(E_i) \geq t > \sum_{i=-m-1}^{-n} p(E_i),$$

where t is the discounting threshold.

ers to rank prospects with equal probability-discounted expected utility. It can then be stated as follows:

> **Stochastic Discounting:** For all prospects X and Y, $X \succsim Y$ if and only if
>
> - $EU(X)_{pd} > EU(Y)_{pd}$ or
> - $EU(X)_{pd} = EU(Y)_{pd}$ and $EU(X) \geq EU(Y)$,
>
> where, for all prospects X, it holds that
>
> $$EU(X)_{pd} = EU(X)_{pd,\, pos} + EU(X)_{pd,\, neg}.$$

The following case illustrates the difference between these versions of Stochastic Discounting:[53]

> **Absolutist vs. Baseline Stochastic Discounting:**
>
> *Prospect A* Gives a 0.01 probability of $-$\$1000 (the agent loses \$1000) and otherwise \$10.
>
> *Prospect B* Certainly gives \$1.

Let the discounting threshold be 0.02. Absolutist and Baseline Stochastic Discounting treat this case similarly if the agent does not have any money when facing this choice. Both versions imply that the agent should ignore the possibility of losing

[53] Note that in this example prospects are defined in terms of monetary gains and losses rather than final consequences, such as wealth levels.

$1000 with A. Consequently, A is better than B because its probability-discounted expected utility is greater.[54]

Next, suppose the agent already possesses $2000. Then, A gives a 0.01 probability of ending up with $1000 and a 0.99 probability of ending up with $2010. With B, the agent certainly ends up with $2001. Baseline Stochastic Discounting treats this case similarly as the case where the agent does not have any money on their bank account. In contrast, Absolutist Stochastic Discounting calculates the values of these prospects using the amounts of money the agent could end up with. So, on this view, the agent ought *not* ignore the possibility of losing $1000 with A; if they lose $1000, then they will end up with $1000 overall, and the probability of ending up with $1000 or more is 1. So, according to Absolutist Stochastic Discounting, B is better than A because because its probability-discounted expected utility is $EU(B)_{pd} = 2001$, while A's probability-discounted expected utility is $EU(A)_{pd} \approx 2000$.[55]

Absolutist Stochastic Discounting has the (possible) disadvantage of requiring an objective neutral utility level. Baseline Stochastic Discounting does not require one because it ignores very large changes to the baseline—the baseline serves the same purpose as the objective neutral level on the absolutist view. Also, Absolutist Stochastic Discounting entails that sometimes one might not ignore the possibility of a huge loss even if there is only a tiny probability of it occurring. This can happen when the agent ends up with a positive outcome regardless of the loss and the

[54] $EU(A)_{pd} = 0.99 \cdot 10 = 9.9$ and $EU(B)_{pd} = 1$.

[55] $EU(A)_{pd} = 1000 + 0.99(2010 - 1000) \approx 2000$.

probability of obtaining an outcome that is at least as good as that is non-negligible. Similarly, it can also happen if the probability of ending up with a worse utility level is non-negligible for some reason not related to the prospect in question. So, Absolutist Stochastic Discounting sometimes lets tiny probabilities of huge losses dictate one's course of action (and similarly for payoffs). Therefore, it does not capture the motivation behind Probability Discounting as well as Baseline Stochastic Discounting does.[56]

Now, recall the earlier violations of Statewise and Stochastic Dominance (*Lexical Statewise Dominance Violation*, *Random Number* and *Two Coins*):

> *Prospect A* Gives a 0.5 probability of \$10 and a 0.01 probability of \$100 (and otherwise nothing).
>
> *Prospect B* Gives a 0.49 probability of \$10 and a 0.02 probability of \$100 (and otherwise nothing).

[56]It is also worth pointing out the following features of Absolutist Stochastic Discounting: First, egoistic agents who are offered the "same" prospects (when one ignores the baseline utility levels) and have the same discounting threshold can reach different conclusions about which option is best. This can happen when they have different baseline utility levels, because one of these agents might end up with an overall positive utility level in a state where the other agent ends up with a negative one. Secondly, if Absolutist Stochastic Discounting takes into account past value, then what happened in the past can influence what altruistic agents now ought to do. For example, if one learns that the past was much better than one thought, then the overall moral value of the world would be much higher. Consequently, one might no longer ignore some tiny probability of a large loss because even if the loss occurred, the value of the world would still be positive (and there is a non-negligible probability of obtaining an outcome that is at least as good as that). This is similar to the Egyptology objection to the Average View in population ethics. See McMahan (1981, p. 115) and Parfit (1984, p. 420). Also, even if one only takes into account future value (or value in one's future light cone), what happens in distant places can affect what altruistic agents here ought to do. Wilkinson (2022, §6) shows that views that reject Probability Fanaticism must violate separability or Stochastic Dominance. Absolutist Stochastic Discounting violates the former. Given that Baseline Stochastic Discounting ignores background uncertainty (and thus satisfies separability), it must sometimes violate Stochastic Dominance.

Again, let the discounting threshold be 0.03. Unlike the earlier versions of Probability Discounting, both versions of Stochastic Discounting imply that B is better than A. B gives a 0.51 probability of at least \$10 and a 0.02 probability of at least \$100, so its probability-discounted expected utility is $EU(B)_{pd} = 5.1$.[57] A in turn gives a 0.51 probability of at least \$10 and a 0.01 probability of at least \$100, so its probability-discounted expected utility is also $EU(A)_{pd} = 5.1$.[58] As A and B have equal probability-discounted expected utility, these prospects are then compared by their expected utilities without discounting. Consequently, B is better than A, and both versions of Stochastic Discounting avoid the earlier violations of Statewise and Stochastic Dominance.

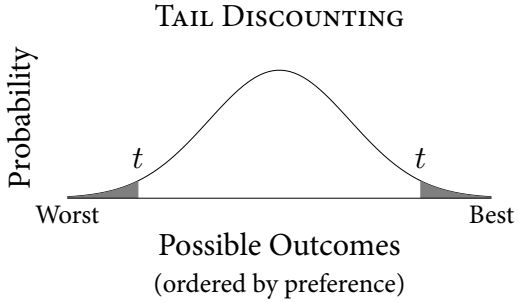
4.2 Tail Discounting

There is a similar view to Absolutist Stochastic Discounting called *Tail Discounting*.[59] According to Tail Discounting, one should ignore both the left and the right 'tails' of the distribution of possible outcomes of some prospect X when these outcomes are ordered by one's preference. Suppose the possible outcomes of some prospect are normally distributed when they are ordered from the least to the most preferred. Then, Tail Discounting advises one to ignore the grey areas under the curve:

[57] $EU(B)_{pd} = 0.51 \cdot 10 = 5.1$.

[58] $EU(A)_{pd} = 0.51 \cdot 10 = 5.1$.

[59] Beckstead and Thomas (2020, §2.3).



Call the outcomes that fall in the middle of the distribution of possible outcomes 'normal outcomes'. An outcome is normal if and only if there is a non-negligible probability of getting at least and at most as good an outcome.[60] Tail Discounting then states the following:

> **Tail Discounting:** For all prospects X and Y, $X \succsim Y$ if and only if
>
> - $EU(X)_{pd} > EU(Y)_{pd}$ or
> - $EU(X)_{pd} = EU(Y)_{pd}$ and $EU(X) \geq EU(Y)$,
>
> where $EU(X)_{pd}$ and $EU(Y)_{pd}$ are obtained by conditionalizing on the supposition that some normal outcome occurs.[61]

[60] For example, in the St. Petersburg game, all payoffs up to some large payoff o are normal, where o depends on one's discounting threshold.

[61] Formally this view states the following:

> **Tail Discounting (formal):** In order to determine $EU(X)_{pd}$, first order the possible outcomes of some prospect X from the least to the most preferred. Then, conditionalize on obtaining some outcome in the middle part of the distribution such that the following necessary conditions hold for all outcomes o that are not ignored:
>
> i The probability of obtaining an outcome that is at least as good as o is above the discounting threshold and
>
> ii the probability of obtaining an outcome that is at most as good as o is above the discounting threshold.
>
> If some outcome o fulfills the above necessary conditions, and

Tail Discounting has the advantage over Absolutist Stochastic Discounting that it does not require an objective neutral level. However, similarly to Absolutist Stochastic Discounting, Tail Discounting recommends paying the mugger in Pascal's Mugging if there is a non-negligible probability of obtaining an outcome at least as great as a thousand quadrillion happy days. This is because then a thousand quadrillion happy days falls in the middle part of the distribution of possible outcomes, which is not ignored.[62]

Again, recall the earlier violations of Statewise and Stochastic Dominance (*Lexical Statewise Dominance Violation*, *Random Number* and *Two Coins*). Tail Discounting also implies that B is better than A. B gives nothing with a 0.49 probability, \$10 with a 0.49 probability and \$100 with a 0.02 probability. Consequently, its probability-discounted expected utility is $EU(B)_{pd} \approx 5.1$.[63] A in turn gives

- the probability of obtaining an outcome that is better than o is below the discounting threshold, then decrease the probability of obtaining o until the total discounted probability of outcomes that are at least as good as o equals the discounting threshold (and conditionalize to make sure the remaining probabilities add up to 1), and
- if the probability of obtaining an outcome that is worse than o is below the discounting threshold, then decrease the probability of obtaining o until the total discounted probability of outcomes that are at most as good as o equals the discounting threshold (and conditionalize to make sure the remaining probabilities add up to 1).

[62] One might also make a version of Tail Discounting similar to Baseline Stochastic Discounting. On this view—let's call it *Baseline Tail Discounting*—one compares every prospect to a baseline prospect as follows: First, calculate the difference in utilities a prospect makes in each state of the world (compared to the baseline prospect). Then, order these differences from the greatest loss to the greatest gain. Then, ignore the right and left tails of this distribution by conditionalization. Also, one can make a version of Tail Discounting similar to Pairwise Stochastic Discounting (i.e., *Pairwise Tail Discounting*). On this view, one compares prospects pairwise instead of comparing every prospect to a baseline prospect.

[63] $EU(B)_{pd} = (0.49 - 0.01)/0.94 \cdot 10 \approx 5.1$. The divisor '0.94' comes from subtracting the

nothing with probability 0.49, \$10 with probability 0.5 and \$100 with probability 0.01. So, its probability-discounted expected utility is also $EU(A)_{pd} \approx 5.1$.[64] As A and B have equal probability-discounted expected utility, these prospects are then compared by their expected utilities without discounting. B has greater expected utility than A without discounting, so B is better than A—and Tail Discounting avoids the earlier violations of Statewise and Stochastic Dominance.

To summarize, I have discussed three versions of Probability Discounting in this section. Absolutist Stochastic Discounting states that one should ignore the possibility of a very high (or a very low) utility level in cases where the cumulative probability of such utility levels is below the discounting threshold. Baseline Stochastic Discounting works similarly, but it operates on gains and losses instead of final utilities. Finally, Tail Discounting states that one should ignore the 'tails' of the distribution of possible outcomes of some prospect when these outcomes are ordered from the least to the most preferred. All these views avoid the earlier violations of Statewise and Stochastic Dominance (and Acyclicity). However, next, I will raise a diachronic problem for these views.

---

discounting threshold of 0.03 from both tails of the distribution. '0.01' is subtracted from 0.49 to make sure that the discounting threshold of 0.03 is ignored in the right tail as well; the probability of obtaining \$100 is 0.02, so merely ignoring the possibility of \$100 would mean one ignores the '0.02 part' of the right tail. More generally, on Tail Discounting, one discounts a little bit of each 'tail' with every prospect (until the discounting threshold is ignored from both tails). See footnote 61.

[64] $EU(A)_{pd} = (0.5 - 0.02)/0.94 \cdot 10 \approx 5.1$.

5 Independence

This section shows that Stochastic and Tail Discounting violate the axiom of Independence.[65] As a result of this violation, these views are vulnerable to exploitation by a money pump. In the next section, I discuss one possible way of avoiding exploitation by this money pump.

5.1 A violation of Independence

Both Stochastic and Tail Discounting violate the axiom of Independence. Let XpY be a risky prospect with a p chance of prospect X obtaining and a $1-p$ chance of prospect Y obtaining. Then, Independence states that

> **Independence:** If $X \succ Y$, then $XpZ \succ YpZ$ for all probabilities $p \in (0, 1]$.[66]

Informally, Independence is the idea that every outcome contributes to the value of a prospect in a way that does not depend on the alternative outcomes.

The basic problem for Probability Discounting is that by mixing gambles, one can arbitrarily reduce the probabilities of different states or outcomes within the compound lottery until these probabilities end up below the discounting threshold. Therefore, mixtures of gambles can end up being valued differently than the gambles that are mixed together. For example, consider the following case:

[65] As these views differ from Expected Utility Theory, they must violate at least one of its axioms. In addition to violating Independence, they also violate Continuity. See §1 in Chapter 5 of this thesis.

[66] Jensen (1967, p. 173).

A Violation of Independence:

Prospect A Certainly gives nothing.

Prospect B Gives a 0.5 probability of \$1 and otherwise $-$\$1,000,000.

Prospect C Certainly gives \$1.

Next, let $p = 0.02$. Then, we have the following mixed prospects (see table 10):

Prospect ApC Gives a 0.98 probability of \$1 and otherwise nothing.

Prospect BpC Gives a 0.99 probability of \$1 and a 0.01 probability of $-$\$1,000,000.

TABLE 10
A VIOLATION OF INDEPENDENCE

p	0.01	0.01	0.98
ApC	\$0	\$0	\$1
BpC	$-$\$1,000,000	\$1	\$1

First, consider what Stochastic Discounting says about these prospects (Baseline and Absolutist Stochastic Discounting treat this case similarly if the agent possesses nothing when making this choice). Let the discounting threshold be 0.02. ApC gives a 0.98 probability of gaining at least \$1 (and otherwise nothing), so its probability-discounted expected utility is $EU(ApC)_{pd} = 0.98$. BpC, in turn, gives a 0.99 probability of gaining at least \$1 and a 0.01 probability of losing at least \$1,000,000. The probability of losing at least \$1,000,000 is below the discounting

threshold, so this possibility is ignored. Thus, BpC's probability-discounted expected utility is $EU(BpC)_{pd} = 0.99$. So, according to Stochastic Discounting, BpC is better than ApC, given that its probability-discounted expected utility is greater than ApC's.

However, the difference between them is that ApC gives a 0.02 probability of nothing, while BpC gives a 0.01 probability of gaining \$1 and a 0.01 probability of losing \$1,000,000 instead (columns 1 and 2 in table 10). Note that BpC is better than ApC no matter how bad the negative outcome is (in column 1) as long as the good outcome (in column 2) is at least slightly positive.

Next, consider what Tail Discounting says about these prospects. Now let the discounting threshold be 0.01. Then, Tail Discounting also implies that BpC is better than ApC. After ignoring both tails of the distribution of possible outcomes of ApC, it probability-discounted expected utility is $EU(ApC)_{pd} \approx 0.99$.[67] And after ignoring the tails of the distribution of possible outcomes of BpC, its probability-discounted expected utility is $EU(BpC)_{pd} = 1$.[68] Thus, we have that BpC is better than ApC.

Some might consider this implication already worrisome on its own, but it is also a violation of Independence—and there is a money pump against theories that violate Independence.[69] Both Stochastic and Tail Discounting consider A better than B. It is better to get nothing certainly than to take a 50–50 gamble between gaining \$1 and losing \$1,000,000. Thus, we have the following violation of Inde-

[67] $(0.98 - 0.01)/0.98 \cdot 1 \approx 0.99$.

[68] $(0.99 - 0.01)/0.98 \cdot 1 = 1$.

[69] See Gustafsson (forthcoming, §5).

pendence:

$A \succ B$, and $BpC \succ ApC$ for some probability $p \in (0, 1]$.

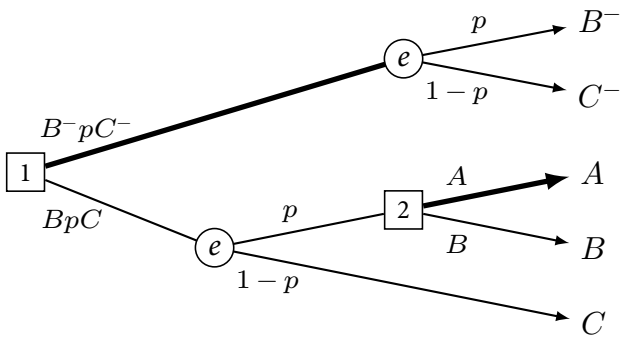
This renders Stochastic and Tail Discounting vulnerable to exploitation by a money pump for Independence.

5.2 The Independence Money Pump

A money-pump argument intends to show that agents who violate some alleged requirement of rationality are vulnerable to making a combination of choices that leads to a sure loss. If vulnerability to this kind of exploitation is a sign of irrationality, then Stochastic and Tail Discounting are untenable as theories of instrumental rationality.

Consider the following decision problem:[70]

THE INDEPENDENCE MONEY PUMP



$A \succ B$, and $BpC \succ B^{-}pC^{-} \succ ApC$.

[70] This money pump is from Gustafsson (2021, p. 31 n21). Also see Gustafsson (forthcoming, §5), Hammond (1988*b*, pp. 292–293) and Hammond (1988*a*, pp. 43–45).

In this decision tree, the squares represent choice nodes, and the circles represent chance nodes. Going up at a choice node means accepting a trade, and going down means refusing a trade.[71] The agent starts with prospect BpC: a 0.99 probability of \$1 and a 0.01 probability of $-$\$1,000,000. At node 1, they are offered a trade from BpC to B^-pC^-, which is like BpC except that the agent has ϵ less money. If the agent turns down this trade and BpC results in the agent going up at the chance node e, the agent will be offered a trade from B to A at node 2. Both chance nodes depend on the same chance event e.

An agent can use *backward induction* to reason about this case. This means that they consider what they would choose at later choice nodes and take those predictions into account when making choices at earlier choice nodes.[72] As the agent prefers A to B, they would accept the trade at node 2. They would rather certainly get nothing than take a 50–50 gamble between gaining \$1 and losing \$1,000,000. Then, by using backward induction at node 1, the prospect of turning down the trade is effectively ApC, and the prospect of accepting the trade is B^-pC^-. Given that the agent prefers BpC over ApC, there must be some price they would be willing to pay to get the former rather than the latter. So, there is some ϵ amount of money such that the agent prefers B^-pC^- over ApC. Then, for some ϵ, they go up at node 1. However, they then end up with B^-pC^-, even though they could have kept BpC had they gone down at both choice nodes.[73] They have given up money

[71] Rabinowicz (2008, p. 152).

[72] Selten (1975) and Rosenthal (1981, p. 95).

[73] Note that if one accepts a baseline or pairwise version of Stochastic or Tail Discounting, then the fact that the agent starts with BpC might matter. For example, if BpC is considered the baseline prospect, then according to the baseline versions, the value of ApC is calculated by comparing

for the exploiter.

Furthermore, as the chance nodes depend on the same event, the prospect of going up at node 1 is statewise dominated by the prospect of going down at both choice nodes. Let a *plan* be a specification of a sequence of choices to be taken by an agent at each choice node that can be reached from that node while following this specification. Stochastic and Tail Discounting advise that an agent make a plan that results in a worse outcome in every state than another available plan. This is a violation of a sequential version of Statewise Dominance:

> **Sequential Statewise Dominance:** If the outcome of plan X is at least as preferred as the outcome of plan Y in all states, and the outcome of X is strictly preferred to the outcome of Y in some possible state, then $X \succ Y$.

Moreover, by choosing $B^{-}pC^{-}$, the agent has paid to give up their power to choose A rather than B if event e occurs. The agent has therefore paid for having their freedom of choice taken away from them.[74]

To summarize, Stochastic and Tail Discounting violate Independence—and in a particularly counterintuitive way.[75] The violation of Independence is particu-

it to BpC in every state. In the state in which event e does not happen, both prospects result in C. Thus, the value of ApC in that state is zero. Compared to BpC, ApC gives a 0.01 probability of gaining a million, a 0.01 probability of losing \$1, and otherwise, it gives nothing. Consequently, its probability-discounted expected utility is $EU(ApC)_{pd} = 0$. These prospects are then compared by their expected utilities without discounting. Consequently, ApC is better than BpC—and we have avoided the Independence violation. However, if the agent does not start with BpC but instead is offered BpC, then they would choose $B^{-}pC^{-}$ for the reasons explained in the main text. They have therefore chosen a dominated prospect.

[74] Rabinowicz (2021, pp. 530–531).

[75] Stochastic and Tail Discounting also violate the following version of Independence:

larly counterintuitive because BpC is considered better than ApC no matter how bad the negative outcome (losing $1,000,000) is. In addition, this implication renders those who accept these views vulnerable to exploitation in the Independence Money Pump. Next, I will discuss one possible way of avoiding exploitation in this case.

6 Avoiding exploitation in the Independence Money Pump

Those who accept Stochastic or Tail Discounting might avoid exploitation in the Independence Money Pump if they use policies of decision-making that prevent dynamic inconsistency. One such decision policy is Resolute Choice.[76] However, I will show that using Resolute Choice in the Independence Money Pump leads to

---

Independence for Constant Outcome: For all probabilities $p \in (0, 1]$, $XpU \succ YpU$ if and only if $YpV \succ XpV$ (McClennen, 1990, p. 45).

In addition to the earlier prospects ApC and BpC, consider the following prospects:

A Violation of Independence for Constant Outcome:

Prospect ApD Gives a 0.98 probability of −$1,000,000 and otherwise nothing.

Prospect BpD Gives a 0.99 probability of −$1,000,000 and a 0.01 probability of $1.

Both Stochastic and Tail Discounting imply that BpC is better than ApC, but ApD is better than BpD. For example, according to Stochastic Discounting, $EU(ApD)_{pd} = 0.98 \cdot (-1,000,000) = -980,000$ and $EU(BpD)_{pd} = 0.99 \cdot (-1,000,000) = -990,000$ (with a discounting threshold of 0.02). There is also a money-pump argument against preferences like this. See Gustafsson (forthcoming, §5) and Raiffa (1968, pp. 83–85).

[76] Backward induction was initially proposed as a decision policy to avoid exploitation. However, as we saw earlier, backward induction got one in trouble in the Independence Money Pump.

untenable results. Furthermore, I will argue that even if there is a way of avoiding exploitation in the Independence Money Pump, Stochastic and Tail Discounting cannot escape the untenable implication they face if combined with Resolute Choice. This makes them (and Probability Discounting more generally) less plausible as theories of instrumental rationality.

A resolute agent chooses in accordance with any plan they have adopted earlier as long as nothing unexpected has happened since the adoption of the plan.[77] Resolute Choice violates Decision-Tree Separability, which states that what is rational at a choice node does not depend on what has happened in the past—only the future matters. With Resolute Choice, one can commit to choosing BpC at node 1 of the Independence Money Pump and then stick to that plan. One then makes a plan that one will not trade B for A at node 2, even though one would usually prefer the latter prospect over the former one. But this seems wrong. Choosing B over A would mean choosing a 0.5 probability of losing \$1,000,000 and otherwise gaining \$1 over certainty of nothing. No reasonable view recommends this.

However, someone might object that the whole point of Resolute Choice is that, by adhering to a plan, the agent makes choices that they would view as unreasonable if they occurred outside the scope of the plan as stand-alone decisions. Therefore, the agent agrees that no reasonable view sanctions the choice if the choice happens outside a plan. Their view is that such a choice can be reasonable when licensed by adhering to the best available plan. However, choosing a 0.5 probabil-

[77] Strotz (1955-1956) and McClennen (1990, pp. 12–13). See Steele (2007), Steele (2018) and Gustafsson (forthcoming, §7) for criticism of Resolute Choice.

ity of losing \$1,000,000 and otherwise gaining \$1 over certainty of nothing is beyond the scope of what is reasonable even for someone who is resolute. We might also change the example so that one loses arbitrarily much instead of losing just \$1,000,000. Furthermore, the probability of this loss can be arbitrarily close to 1.[78] It would not be rational to commit to choosing that prospect over certainty of nothing. So, Resolute Choice is untenable in combination with Stochastic and Tail Discounting as a solution to the Independence Money Pump.

Stochastic and Tail Discounting violate the axiom of Independence in a particularly counterintuitive way. This case is, therefore, worrisome independently of the exploitation. The violation of Independence is particularly counterintuitive because BpC is considered better than ApC no matter how bad the negative outcome ($-$\$1,000,000) is as long as the good outcome (\$1) is at least slightly positive.

Suppose that at node 1 of the Independence Money Pump, the agent is offered another option: to lock in their choice at node 2 without knowing whether e has happened. Stochastic and Tail Discounting would recommend locking in the choice of B because then, at node 1, the agent faces BpC, which is better than ApC and $B^{-}pC^{-}$. However, this seems wrong. First, this would mean that the agent willingly avoids costless information by locking in their choice at node 2 without knowing whether e has happened.[79] More importantly, they would lock in the

[78] For example, let BpC be a prospect that gives a $0.02 - \epsilon$ probability of losing arbitrarily much; otherwise, it gives \$1 (probability $0.98 + \epsilon$). Then, BpC would still be better than ApC because it gives a higher probability of \$1. However, prospect B would almost certainly give an arbitrarily large loss and only a small probability of \$1.

[79] More generally, agents who violate Independence avoid costless information. See for example Wakker (1988), Hilton (1990) and Machina (1989, p. 1638–1639).

choice of a lottery that gives a 0.5 probability of $-$\$1,000,000 and otherwise \$1 (B) over certainty of nothing (A). Limiting one's future choices in this way seems irrational. Even if the agent does not accept Resolute Choice, they would still lock in the same choice of B over A if offered the option at node 1. This makes Probability Discounting less plausible even if some decision policy helps probability discounters avoid exploitation in the Independence Money Pump. Even absent exploitation, choosing B over A, or locking in the choice of B over A, seems irrational.[80]

To summarize, those who accept Stochastic or Tail Discounting might be able to avoid exploitation in the Independence Money Pump if they use policies of decision-making that prevent dynamic inconsistency. I have argued that these views give untenable recommendations if combined with Resolute Choice. I also argued that even if there is a way of avoiding exploitation in the Independence Money Pump), Stochastic and Tail Discounting cannot avoid the untenable result they face if combined with Resolute Choice. This makes them—and Probability Discounting more generally—less plausible as theories of instrumental rationality.[81]

[80] However, some might argue that saying that it is irrational to lock in B over A in this context simply amounts to saying that it is irrational to prefer BpC to ApC—which is begging the question.

[81] What should one do now? One could, for example, bite the bullet and accept one version of Probability Discounting discussed in this chapter, find a more plausible version of Probability Discounting, bound utilities, conditionalize on one's knowledge before maximizing expected utility (see for example Francis and Kosonen [n.d.]) or accept Probability Fanaticism (see for example Beckstead and Thomas [2020] and Wilkinson [2022]). However, note that, independently of Probability Discounting, agents with unbounded utilities are also vulnerable to money pumps because they violate countable generalizations of the Independence axiom. See Russell and Isaacs (2021).

7 Conclusion

Maximizing expected utility implies counterintuitive choices in cases that involve tiny probabilities of huge payoffs. In response to such cases, some have argued that we should deviate from Expected Utility Theory by discounting small probabilities to zero. I have discussed how exactly this view can be formulated. First, I argued that less plausible versions of Probability Discounting violate dominance. More specifically, I showed that Naive Discounting, Lexical Discounting and Baseline State Discounting violate Statewise Dominance. I also showed that Pairwise State Discounting violates Stochastic Dominance and Acyclicity within choice sets and that Set-Dependent State Discounting violates Pairwise Acyclicity, Contraction and Expansion Consistency and Stochastic Dominance.

Then, I showed that more plausible versions of Probability Discounting, namely Stochastic Discounting and Tail Discounting, avoid these dominance violations. However, they violate the axiom of Independence and do so in a particularly counterintuitive way. As a result of this violation, those who accept these views can be exploited in the Independence Money Pump. I then argued that these views cannot use Resolute Choice to avoid exploitation because this would have untenable implications. Lastly, I argued that even if there is a way of avoiding exploitation in the Independence Money Pump, Stochastic and Tail Discounting cannot avoid the untenable result they face if combined with Resolute Choice. This makes them—and Probability Discounting more generally—less plausible as theories of instrumental rationality. All in all, I have discussed possible ways of formulating Probability

Discounting. All of these theories have significant problems, and it is yet to be seen whether there is a perfectly rational, reasonable decision theory that deviates from Expected Utility Theory by discounting small probabilities down to zero.

References

Baumann, P. (2009), 'Counting on numbers', *Analysis* **69**(3), 446–448.

Beckstead, N. (2013), On the overwhelming importance of shaping the far future, PhD thesis, Rutgers, the State University of New Jersey.

Beckstead, N. and Thomas, T. (2020), A paradox for tiny probabilities and enormous values. Unpublished manuscript. Global Priorities Institute Working Paper No.10.

URL: *https://globalprioritiesinstitute.org/nick-beckstead-and-teruji-thomas-a-paradox-for-tiny-probabilities-and-enormous-values/*

Bernoulli, D. (1954), 'Exposition of a new theory on the measurement of risk', *Econometrica* **22**(1), 23–36.

Bostrom, N. (2009), 'Pascal's Mugging', *Analysis* **69**(3), 443–445.

Broome, J. (1991), *Weighing Goods: Equality, Uncertainty and Time*, Blackwell, Oxford.

Buchak, L. (2013), *Risk and Rationality*, Oxford University Press, Oxford.

Buchak, L. (2014), 'Risk and tradeoffs', *Erkenntnis* **79**(6), 1091–1117.

Fishburn, P. C. (1991), 'Nontransitive preferences in decision theory', *Journal of Risk and Uncertainty* **4**(2), 113–134.

Francis, T. and Kosonen, P. (n.d.), Ignore outlandish possibilities. Unpublished manuscript.

Goodsell, Z. (2021), 'A St Petersburg paradox for risky welfare aggregation', *Analysis* **81**(3), 420–426.

Gustafsson, J. E. (2021), 'The sequential dominance argument for the Independence axiom of Expected Utility Theory', *Philosophy & Phenomenological Research* **103**(1), 21–39.

Gustafsson, J. E. (forthcoming), *Money-Pump Arguments*, Cambridge University Press, Cambridge.

Hadar, J. and Russell, W. R. (1969), 'Rules for ordering uncertain prospects', *The American Economic Review* **59**(1), 25–34.

Hájek, A. (2014), 'Unexpected expectations', *Mind* **123**(490), 533–567.

Hammond, P. J. (1988*a*), 'Consequentialist foundations for expected utility', *Theory and Decision* **25**(1), 25–78.

Hammond, P. J. (1988*b*), 'Orderly decision theory: A comment on professor Seidenfeld', *Economics and Philosophy* **4**(2), 292–297.

Hanson, R. (2007), 'Pascal's Mugging: Tiny probabilities of vast utilities'.

URL: *https://www.lesswrong.com/posts/a5JAiTdytou3Jg749/pascal-s-mugging-tiny-probabilities-of-vast-utilities?commentId=Q4ACkdYFEThA6EE9P*

Hey, J. D., Neugebauer, T. M. and Pasca, C. M. (2010), Georges-Louis Leclerc de Buffon's 'Essays on moral arithmetic', *in* A. Sadrieh and A. Ockenfels, eds, 'The Selten School of Behavioral Economics: A Collection of Essays in Honor of Reinhard Selten', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 245–282.

Hilton, R. W. (1990), 'Failure of Blackwell's Theorem under Machina's generalization of expected-utility analysis without the independence axiom', *Journal of Economic Behavior & Organization* **13**(2), 233–244.

Huemer, M. (2016), *Approaching Infinity*, Palgrave Macmillan, New York.

Isaacs, Y. (2016), 'Probabilities cannot be rationally neglected', *Mind* **125**(499), 759–762.

Jensen, N. E. (1967), 'An introduction to Bernoullian utility theory: I. Utility functions', *The Swedish Journal of Economics* **69**(3), 163–183.

Kreps, D. M. (1988), *Notes on the Theory of Choice*, Westview Press, Boulder.

Lehmann, E. L. (1955), 'Ordered families of distributions', *The Annals of Mathematical Statistics* **26**(3), 399–419.

Luce, R. D. and Raiffa, H. (1957), *Games and Decisions: Introduction and Critical Survey*, Wiley, New York.

Lundgren, B. and Stefánsson, H. O. (2020), 'Against the *De Minimis* principle', *Risk Analysis* **40**(5), 908–914.

Machina, M. J. (1989), 'Dynamic consistency and non-expected utility models of choice under uncertainty', *Journal of Economic Literature* **27**(4), 1622–1668.

Mann, H. B. and Whitney, D. R. (1947), 'On a test of whether one of two random variables is stochastically larger than the other', *The Annals of Mathematical Statistics* **18**(1), 50–60.

McClennen, E. F. (1990), *Rationality and Dynamic Choice: Foundational Explorations*, Cambridge University Press, Cambridge.

McMahan, J. (1981), 'Problems of population theory', *Ethics* **92**(1), 96–127.

Menger, K. (1967), The role of uncertainty in economics, *in* M. Shubik, ed., 'Essays in Mathematical Economics: In Honor of Oskar Morgenstern', Princeton University Press, Princeton, pp. 211–231.

Monton, B. (2019), 'How to avoid maximizing expected utility', *Philosophers' Imprint* **19**(18), 1–24.

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

Parfit, D. (1984), *Reasons and Persons*, Clarendon Press, Oxford.

Pulskamp, R. J. (n.d.), Correspondence of Nicolas Bernoulli concerning the St. Petersburg game. Unpublished manuscript. Accessed through:

https://web.archive.org/.

URL: *http://cerebro.xu.edu/math/Sources/NBernoulli/correspondence_petersburg_game.pdf*

Quirk, J. P. and Saposnik, R. (1962), 'Admissibility and measurable utility functions', *The Review of Economic Studies* **29**(2), 140–146.

Rabinowicz, W. (2008), Pragmatic arguments for rationality constraints, *in* M. C. Galavotti, R. Scazzieri and P. Suppes, eds, 'Reasoning, Rationality, and Probability', CSLI, Stanford, CA, pp. 139–163.

Rabinowicz, W. (2021), Between sophistication and resolution—wise choice, *in* R. Chang and K. Sylvan, eds, 'The Routledge Handbook of Practical Reason', Routledge, Abingdon, Oxon ; New York, NY, pp. 526–540.

Raiffa, H. (1968), *Decision Analysis: Introductory Lectures on Choices under Uncertainty*, Addison–Wesley, Reading, MA.

Rosenthal, R. W. (1981), 'Games of perfect information, predatory pricing and the chain-store paradox', *Journal of Economic Theory* **25**(1), 92–100.

Russell, J. S. (2021), 'On two arguments for fanaticism'. Global Priorities Institute Working Paper 17–2021.

URL: *https://globalprioritiesinstitute.org/on-two-arguments-for-fanaticism-jeff-sanford-russell-university-of-southern-california/*

Russell, J. S. and Isaacs, Y. (2021), 'Infinite prospects', *Philosophy and Phenomenological Research* **103**(1), 178–198.

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

Savage, L. J. (1951), 'The theory of statistical decision', *Journal of the American Statistical Association* **46**(253), 55–67.

Selten, R. (1975), 'Reexamination of the perfectness concept for equilibrium points in extensive games', *International Journal of Game Theory* **4**(1), 25–55.

Sen, A. (1977), 'Social choice theory: A re-examination', *Econometrica* **45**(1), 53–88.

Smith, N. J. J. (2014), 'Is evaluative compositionality a requirement of rationality?', *Mind* **123**(490), 457–502.

Smith, N. J. J. (2016), 'Infinite decisions and rationally negligible probabilities', *Mind* **125**(500), 1199–1212.

Steele, K. (2007), Precautionary Decision-Making: An Examination of Bayesian Decision Norms in the Dynamic Choice Context, PhD thesis, University of Sydney.

Steele, K. (2018), Dynamic decision theory, *in* S. O. Hansson and V. F. Hendricks, eds, 'Introduction to Formal Philosophy', Springer, Cham, pp. 657–667.

Strotz, R. H. (1955-1956), 'Myopia and inconsistency in dynamic utility maximization', *The Review of Economic Studies* **23**(3), 165–180.

Wakker, P. (1988), 'Nonexpected utility as aversion of information', *Journal of Behavioral Decision Making* **1**(3), 169–175.

Wilkinson, H. (2022), 'In defence of fanaticism', *Ethics* **132**(2), 445–477.

Yudkowsky, E. (2007*a*), 'A comment on Pascal's Mugging: Tiny probabilities of vast utilities'.

URL: *https://www.lesswrong.com/posts/a5JAiTdytou3Jg749/pascal-s-mugging-tiny-probabilities-of-vast-utilities?commentId=kqAKXskjohx4SSyp4*

Yudkowsky, E. (2007*b*), 'Pascal's Mugging: Tiny probabilities of vast utilities'.

URL: *http://www.overcomingbias.com/2007/10/pascals-mugging.html*

Chapter 5

*Probability Discounting and Money Pumps**

ABSTRACT: In response to cases that involve very small probabilities of huge payoffs, some argue that we ought to discount very small probabilities down to zero. However, this chapter shows that doing so violates Independence and Continuity, and as a result of these violations, those who discount small probabilities can be exploited by money pumps. Various possible ways of avoiding exploitation will be discussed. However, echoing the previous chapter, this chapter concludes that the money pump for Independence undermines the plausibility of discounting small probabilities.

On the standard decision theory, a rational agent always maximizes expected utility. However, this seems to lead to counterintuitive choices in cases that involve very small probabilities of huge payoffs. Consider, for example, the following case:[1]

*I wish to thank Andreas Mogensen and Teruji Thomas for valuable feedback and discussions.

[1]Bostrom (2009). This case is based on informal discussions by different people, including Eliezer Yudkowsky (2007). Another case that involves very small probabilities of huge payoffs is the St. Petersburg game. See for example Peterson (2020).

> **Pascal's Mugging:** Someone approaches Pascal and claims to be an Operator from the Seventh Dimension. The stranger promises to perform magic that will give Pascal a thousand quadrillion happy days in the Seventh Dimension if Pascal pays the mugger ten livres—money that the mugger will use for helping orphans in the Seventh Dimension.

Pascal thinks the probability of the mugger telling the truth is very low. However, the potential payoff is so high that the expected utility of paying the mugger is positive. Furthermore, as long as Pascal has a non-zero credence in the proposition that the mugger is able and willing to reward him with any finite amount of utility, the mugger can increase the payoff until the offer has positive expected utility.[2] At some point, maximizing expected utility (with unbounded utilities) requires paying the mugger. And more generally, it leads to

> **Probability Fanaticism:** For any tiny probability $p > 0$, and for any finite utility u, there is some large enough utility U such that probability p of U (and otherwise nothing) is better than certainty of u.[3]

In response to cases like this, some have argued that we ought to discount very small probabilities down to zero—let's call this *Probability Discounting*. For example, Monton (2019) argues that one ought to discount very small probabilities down to zero, while Smith (2014) argues that it is rationally permissible, but not re-

[2]This may not be possible if utility is bounded as standard axiomatizations of expected utility maximization require. See for example Kreps (1988, p. 63) and §1 and §2.1 in Chapter 1 of this thesis.

[3]Wilkinson (2022, p.449).

quired, to do so.[4] There are many ways of making Probability Discounting precise. Let $X \succsim Y$ mean that X is at least as preferred as Y. Also, let $EU(X)_{pd}$ denote the expected utility of prospect X when small probabilities have been discounted down to zero (read as 'the probability-discounted expected utility of X'). Also, let a *negligible probability* be a probability below the discounting threshold, that is, a probability that should be discounted down to zero. Then, one of the simplest versions of Probability Discounting—let's call it *Naive Discounting*—states:[5]

> **Naive Discounting:** For all prospects X and Y, $X \succsim Y$ if and only if $EU(X)_{pd} \geq EU(Y)_{pd}$, where $EU(X)_{pd}$ and $EU(Y)_{pd}$ are obtained by conditionalizing on the supposition that some outcome of non-negligible probability occurs.

Given that Probability Discounting differs from Expected Utility Theory, it has to violate at least one of the following axioms that together entail Expected Utility Theory: Completeness, Transitivity, Independence and Continuity.[6] Furthermore, violating these axioms renders probability discounters vulnerable to exploitation

[4]Smith argues that discounting small probabilities allows one to get a reasonable expected utility for the Pasadena game (see [Nover and Hájek 2004]). On Smith's view, the discounting threshold could be chosen lower than any relevant probability in cases that involve finitely many possible outcomes. So, in effect, discounting small probabilities might not apply to cases involving a finite number of possible outcomes. See Hájek (2014), Isaacs (2016) and Lundgren and Stefánsson (2020) for criticism of discounting small probabilities. Also see Beckstead (2013, ch. 6), Beckstead and Thomas (2020), Goodsell (2021), Russell and Isaacs (2021), Russell (2021) and Wilkinson (2022) for discussions of issues surrounding Probability Fanaticism.

[5]See Chapter 4 for a discussion of some possible versions of Probability Discounting.

[6]von Neumann and Morgenstern (1947), Jensen (1967, pp. 172–182) and Hammond (1998, pp. 152–164). This chapter assumes the von Neumann-Morgenstern framework with its lotteries with given probabilities, rather than the Savage framework, where subjective probabilities must be constructed alongside utilities, requiring the use of a different and more expansive set of axioms.

as there are money-pump arguments for each of these axioms.[7]

This chapter shows that some versions of Probability Discounting, such as Naive Discounting, violate Independence and Continuity. They are therefore vulnerable to exploitation in the money pumps for Independence and Continuity.[8] Here's the structure of the chapter: §1 discusses three ways in which Probability Discounting might violate Continuity. This section also shows that probability discounters are vulnerable to exploitation in a money pump for Continuity. Lastly, it discusses some ways of avoiding exploitation in that case. §2 shows that Probability Discounting violates Independence. As a result, probability discounters are vulnerable to exploitation in a money pump for Independence. §3 discusses possible ways of avoiding exploitation in the Independence Money Pump. It concludes that there is no plausible way to do this. The chapter concludes that the Independence Money Pump greatly undermines the plausibility of Probability Discounting.

1 Continuity

This section discusses three ways in which Probability Discounting might violate Continuity. First, it shows that views that discount probabilities below some discounting threshold violate Continuity. Next, it shows that views that discount

[7] Gustafsson (forthcoming). It has also been argued that even agents who conform to Expected Utility Theory can be exploited in some cases with an infinite series of trade offers. Gustafsson (forthcoming, §8) argues that such agents can avoid exploitation if they use backward induction.

[8] Isaacs (2016) also presents a problem for probability discounters in a dynamic context, to which Smith (2016) and Monton (2019) respond by arguing that relevantly similar choices ought to be evaluated collectively. This response does not help avoid exploitation in the cases discussed in this chapter.

probabilities up to some discounting threshold violate another version of Continuity. Finally, it shows that views that ignore very-small-probability *outcomes* must violate either Continuity or Statewise Dominance.[9] As a result of violating Continuity, Probability Discounting is vulnerable to exploitation in a money pump for Continuity. Some ways of avoiding exploitation in this money pump will be discussed.

1.1 The Continuity Money Pump

As mentioned earlier, Continuity is one of the axioms that together entail Expected Utility Theory. Let $X \succ Y$ mean that X is strictly preferred (or simply 'preferred') to Y.[10] Also, let XpY be a risky prospect with a p chance of prospect X obtaining and a $1-p$ chance of prospect Y obtaining. Continuity then states the following:

> **Continuity:** If $X \succ Y \succ Z$, then there are probabilities p and $q \in (0,1)$ such that $XpZ \succ Y \succ XqZ$.

Views that discount probabilities below some threshold violate Continuity. To see how the Continuity violation happens, consider the following prospects:[11]

Continuity Violation:

[9] Instead of ignoring very-small-probability outcomes, one might ignore very-small-probability *states*. See §3 in Chapter 4 of this thesis on State Discounting.

[10] Some prospect X is strictly preferred to another prospect Y when X is weakly preferred to Y, but Y is not weakly preferred to X.

[11] Naive Discounting, Lexical Discounting, State Discounting, Stochastic Discounting and Tail Discounting (discussed in Chapter 4 of this thesis) all violate Continuity in this case (if the discounting threshold is the lowest probability not discounted down to zero).

Prospect A_t Gives probability t of some very good outcome (and otherwise nothing).

Prospect B Certainly gives a good outcome.

Prospect C Certainly gives nothing.

Let t be the discounting threshold. Then, all probabilities less than t will be discounted down to zero, but probabilities at least as great as t will not be discounted. Also, suppose that A_t is better than B, which is better than C; a non-negligible probability of a very good outcome (and otherwise nothing) is better than a certain good outcome, which is better than certainly getting nothing.

Next, consider the following mixed lottery (see table 1):

Prospect A_tpC Gives probability p of A_t and probability $1 - p$ of C (i.e., probability $t \cdot p$ of a very good outcome and otherwise nothing).

Given that t is the discounting threshold, t multiplied by any probability $p < 1$ must be below the discounting threshold. Consequently, $t \cdot p$ is discounted down to zero, and A_tpC only gives a negligible probability of a positive outcome. And, given that B certainly gives a good outcome, B must be better than A_tpC for all probabilities $p \in (0, 1)$. So, now we have that A_t is better than B, which is better than C, but B is better than A_tpC for all probabilities $p \in (0, 1)$—which is a violation of Continuity.

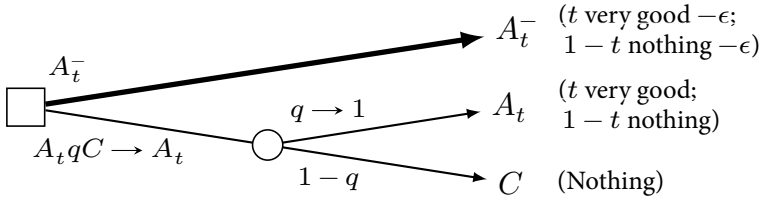
TABLE 1

A VIOLATION OF CONTINUITY

	$p \cdot t$	$1 - p \cdot t$
$A_t pC$	Very good	Nothing
B	Good	Good

There is also a money-pump argument for Continuity. A money-pump argument intends to show that agents who violate some alleged requirement of rationality would make a combination of choices that lead to a sure loss. In so far as vulnerability to this kind of exploitation is a sign of irrationality, Probability Discounting is untenable as a theory of instrumental rationality. The money-pump argument for Continuity goes as follows:[12]

THE CONTINUITY MONEY PUMP



$A_t \succ A_t^- \succ A_t pC$ for all probabilities $p \in (0, 1)$.

In this decision tree, the square represents a choice node and the circle represents a chance node. Going up at a choice node means accepting a trade and going down means refusing a trade.[13] The agent starts with $A_t qC$. $A_t qC$ is arbitrarily similar

[12] See Gustafsson (forthcoming, §6). Gustafsson calls this the Lexi-Pessimist Money Pump. Gustafsson (forthcoming, §6) also presents another money pump against preferences that violate Continuity in a different way.

[13] Rabinowicz (2008, p. 152).

to A_t; it results in the same outcome as A_t with a probability arbitrarily close to one. However, no matter how close q is to one, $A_t qC$ will only give a negligible probability of a positive outcome. Next, the agent is offered A_t^- in exchange for $A_t qC$. A_t^- is like A_t except that the agent has some amount ϵ less money. A_t^- gives the threshold probability of a positive outcome, while $A_t qC$ only gives a negligible probability of a positive outcome. Thus, the agent prefers A_t^- over $A_t qC$, no matter how close q is to one. Consequently, the agent accepts the trade. However, this means that the exploiter gets a fixed payment with only an arbitrarily small chance of having to give up anything. The situation is therefore arbitrarily close to pure exploitation.

To summarize, views on which all probabilities below some discounting threshold are ignored violate Continuity, and they are therefore vulnerable to exploitation in the Continuity Money Pump.

1.2 Mixture Continuity

The previous Continuity violation happens because the discounting threshold multiplied by any probability below one results in a probability below the discounting threshold. This happens because the discounting threshold is the lowest probability not discounted down to zero. Hence, the set of non-discounted values is closed (i.e., it is an interval of the form [t, 1]). However, instead of the discounting threshold being the lowest probability not discounted down to zero, it might be the highest probability that is discounted. In that case, there is no lowest non-negligible probability, and the set of non-discounted values is open on one side (i.e., it is an interval

of the form (t, 1]). So, A_t will only have positive probability-discounted expected utility if it gives at least a $t + \varepsilon$ probability of a positive outcome, where ε is positive but arbitrarily close to zero. But in that case, one can always find some probability p (that may be very close to one), such that $p(t + \varepsilon) > t$. In other words, for all probabilities above the discounting threshold, there is some probability p such that their product is still above the discounting threshold. Consequently, Probability Discounting can avoid the previous violation of Continuity by letting the discounting threshold be the highest probability discounted down to zero.

However, this view violates another version of Continuity:

> **Mixture Continuity:** For all prospects X, Y and Z, the set of probabilities $\{ p \in [0, 1] \}$ with property $XpZ \succsim Y$ and the set of probabilities $\{ q \in [0, 1] \}$ with property $Y \succsim XqZ$ are closed.[14]

In effect, this principle states that if prospect XpZ is at least as good as prospect Y with some probability p, then there must be some highest and some lowest probability with which XpZ is at least as good as Y. (Similarly, if prospect Y is at least as good as prospect XqZ, then there must be some highest and some lowest probability with which Y is at least as good as XqZ). To see how the view under consideration violates Mixture Continuity, consider the following prospects:[15]

---

[14] This is axiom 2 in Herstein and Milnor (1953, p. 293). Another way to state Mixture Continuity is as follows: If $\lim_{i \to \infty} p_i = p$ and each $Xp_iZ \succsim Y$, then $XpZ \succsim Y$. Similarly, if $\lim_{i \to \infty} p_i = p$ and $Y \succsim Xp_iZ$, then $Y \succsim XpZ$.

[15] This case is also a violation of the following version of Continuity that can be derived from Mixture Continuity (Herstein and Milnor, 1953, pp. 293–294):

> **Continuity (weak-preference):** If $X \succsim Y \succsim Z$, then there is a probability $p \in (0, 1)$ such that $Y \sim XpZ$.

Mixture Continuity Violation:

Prospect A Certainly gives a very good outcome.

Prospect B Certainly gives a good outcome.

Prospect C Certainly gives nothing.

In this case, A is better than B, which is better than C. Moreover, suppose that the very good outcome is sufficiently great so that ApC is at least as good as B for all $p > t$. Given that t is discounted down to zero, it is not the case that AtC is at least as good as B. So, there is no lowest probability p with which ApC is at least as good as B. For all $p > t$, ApC is at least as great as B; when $p = t$, ApC is worse than B. This is a violation of Mixture Continuity.[16]

Furthermore, even though this view avoids the first Continuity violation, it is still vulnerable to the Continuity Money Pump. Let $A_{t+\varepsilon}$ be a prospect that gives a probability $t+\varepsilon$ of a very good outcome (and otherwise it gives nothing). $A_{t+\varepsilon}$ has positive probability-discounted expected utility for all $\varepsilon > 0$, no matter how close ε is to zero. Also, let $A_{t+\varepsilon}pC$ be a prospect that gives a probability $p(t+\varepsilon)$ of a very good outcome (and otherwise it gives nothing). If ε is very close to zero, $A_{t+\varepsilon}pC$ will only have positive probability-discounted expected utility if p is very close to one—otherwise the probability of a positive outcome would be at most t, and thus,

In Mixture Continuity Violation, A is better than B, which is better than C. However, there is no probability $p \in (0, 1)$ such that $B \sim ApC$. When $p > t$, ApC is better than B (we can suppose so); when $p \leq t$, ApC is worse than B because it only gives a negligible probability of a positive outcome.

[16] As before, Naive Discounting, Lexical Discounting, State Discounting, Stochastic Discounting and Tail Discounting all violate Mixture Continuity in this way (if the discounting threshold is the highest probability discounted down to zero).

discounted down to zero. As ε can be arbitrarily close to zero, $A_{t+\varepsilon}pC$ does not have positive probability-discounted expected utility with probabilities arbitrarily close to one; as long as $p(t+\varepsilon)$ is at most t, $A_{t+\varepsilon}pC$ is at most marginally better than nothing. Consequently, even when p is very close to one, probability discounters would be willing to pay some fixed amount in order to trade $A_{t+\varepsilon}pC$ for $A_{t+\varepsilon}$ in the Continuity Money Pump. So, if we fix p, no matter how close to one, we can find a version of the Continuity Money Pump where the exploiter wins with probability p as long as we choose ε sufficiently close to zero. Therefore, an exploiter can get a fixed payment (up to the value of $A_{t+\varepsilon}$) from the agent with only an arbitrarily small chance $(1-p)$ of having to give up anything.

To summarize, views on which probabilities up to some discounting threshold are ignored violate Mixture Continuity. They are also vulnerable to exploitation in the Continuity Money Pump.

1.3 Continuity and Statewise Dominance

As discussed earlier, Continuity violations can be avoided by using an open set of non-discounted probabilities (although this does not help avoid violations of Mixture Continuity). However, I will show that Probability Discounting must violate either Continuity or Statewise Dominance. Consider the following prospects (see table 2):

Continuity or Statewise Dominance Violation:

Prospect D Gives a very good outcome in state 1, a good outcome in

state 2 and nothing in state 3.

Prospect D^{--} Gives the same outcome as D minus ϵ in state 1, a good outcome in state 2 and nothing in state 3.

Prospect C Certainly gives nothing.

Let the probability of state 1 be b, which is below the discounting threshold. Also, let the probability of state 2 be a_1, which is above the discounting threshold. Finally, let the probability of state 3 be a_2, which is also above the discounting threshold.

TABLE 2
A VIOLATION OF CONTINUITY OR STATEWISE DOMINANCE

	State 1 $b < t$	**State 2** $a_1 > t$	**State 3** $a_2 > t$
D	Very good	Good	Nothing
D^{--}	Very good$-\epsilon$	Good	Nothing
C	Nothing	Nothing	Nothing

Given that the only difference between D and D^{--} is what happens in a very-small-probability state (i.e., state 1), D and D^{--} have equal probability-discounted expected utility. However, considering these prospects equally good would be a violation of the following principle:[17]

[17]Note that, strictly speaking, Statewise Dominance is undefined in the framework of decision theory under risk (such as the von Neumann-Morgenstern framework), as this notion belongs to decision theory under uncertainty, where there is an explicit underlying state space (such as in the Savage framework).

> **Statewise Dominance:** If the outcome of prospect X is at least as preferred as the outcome of prospect Y in all states, then $X \succsim Y$. Furthermore, if in addition the outcome of X is strictly preferred to the outcome of Y in some possible state, then $X \succ Y$.[18]

In state 1, D gives a better outcome than D^{--}, while in the other states, they give the same outcomes. So, by Statewise Dominance, D is better than D^{--}.

To avoid violating Statewise Dominance in this way, probability discounters might use Statewise Dominance to rank prospects that have equal probability-discounted expected utility.[19] D would then be considered better than D^{--}, even though their probability-discounted expected utilities are the same. However, this will lead to a violation of Continuity. Consider the following mixed lottery:

> *Prospect* DpC Gives probability p of D and probability $1 - p$ of C.

Any decrease in the probability of the good outcome in D will make it the case that D's probability-discounted expected utility is less than that of D^{--}. Thus, DpC's probability-discounted expected utility is less than that of D^{--} for all probabilities $p \in (0, 1)$. Consequently, D^{--} is better than DpC for all probabilities $p \in (0, 1)$. Now we have that D is better than D^{--}, which is better than C, but D^{--} is better than DpC for all probabilities $p \in (0, 1)$—which is a violation of Continuity. So, using Statewise Dominance to rank prospects that have equal probability-discounted expected utility leads to a violation of Continuity.

[18] Savage (1951, p. 58) and Luce and Raiffa (1957, p. 287).

[19] See Monton (2019, §7).

However, one might argue that if p is very close to one, the agent should ignore the possibility of obtaining C with DpC. After all, $1 - p$ would then be below the discounting threshold, and C would have a negligible probability. Furthermore, we might rank prospects that have equal probability-discounted expected utility with their expected utilities (without discounting small probabilities).[20] Consequently, DpC would have the same probability-discounted expected utility as D and D^{--}. And, with some values of p, DpC has greater expected utility (without discounting small probabilities). So, if we rank prospects with their expected utilities without discounting in cases where they have equal probability-discounted expected utility, then DpC is better than D^{--} with some probability $p \in (0, 1)$. Thus, there is no violation of Continuity; it is not the case that DpC is worse than D^{--} for all $p \in (0, 1)$.

However, if one ignores *outcomes* whose associated probabilities are below the discounting threshold, then one should not ignore the possibility of obtaining C; C certainly results in obtaining nothing, and D gives a non-negligible probability of obtaining nothing as well. As the probability of obtaining nothing with DpC is non-negligible, C should not be ignored. Consequently, DpC is worse than D^{--} for all probabilities $p \in (0, 1)$—and we have not avoided violating Continuity. As long as one ignores outcomes whose associated probabilities are below the discounting threshold, one must violate Statewise Dominance or Continuity.[21]

[20] See §2 in Chapter 4 of this thesis.

[21] Naive Discounting, Lexical Discounting, Stochastic Discounting and Tail Discounting violate Continuity or Statewise Dominance in this way. However, if one accepts State Discounting, one might be able to ignore the possibility of obtaining C with DpC if its associated state has a negligible probability. See §3 in Chapter 4 of this thesis on State Discounting. State Discounting might

Preferences like this are also vulnerable to a money pump. However, this money pump is not as profitable for the exploiter as the previous one because the exploiter only gets a negligible probability of gaining something. Let D^- be a prospect that is like D except that the agent has less money in state 1 (of table 2), but the outcome in state 1 is still preferred to the outcome of D^{--} in state 1. D, D^- and D^{--} have equal probability-discounted expected utility. But, by Statewise Dominance, we have that D is better than D^-, which is better than D^{--}. The setup of the money pump is similar to the Continuity Money Pump. The agent starts with DqC, which is arbitrarily similar to D. The agent is then offered D^- in exchange for DqC. The agent prefers D^- to DqC, no matter how close q is to one because DqC's probability-discounted expected utility is less than that of D^-; any decrease in the probability of a good outcome in D will result in its probability-discounted expected utility being lower than that of D^-. Thus, the exploiter gets a negligible probability of payment from the agent with only an arbitrarily small chance of having to give up anything. This money pump is not as profitable to the exploiter as the previous ones because there is only a small probability that they will get the payment. However, there is only an arbitrarily small probability that they will lose something, so this scheme is still profitable to the exploiter in expectation.

---

therefore avoid violating Continuity—at least if the discounting threshold is the highest probability discounted down to zero. However, it would still violate Mixture Continuity in the same way as discussed in §1.2. Furthermore, the different versions of State Discounting violate either Acyclicity or Statewise Dominance. See §3 in Chapter 4 of this thesis.

1.4 Vulnerability to the Continuity Money Pump

Probability discounters are vulnerable to exploitation in the Continuity Money Pump because arbitrarily small increases in probability, from just below the discounting threshold to just above it, can make a large difference to the value of a prospect. One partial solution would be to reduce probabilities just above the discounting threshold, but not all the way down to zero—let's call this *Regressive Discounting*.[22] Probability discounters would still choose A_t^- in the Continuity Money Pump. But they would not be willing to pay as much for it as they would without reducing probabilities above the discounting threshold.

However, even if probabilities above the discounting threshold are reduced, it may be possible to compensate for those reduced probabilities by increasing the utility numbers.[23] So, probability discounters would still pay a significant sum to get A_t^- instead of $A_t qC$. Nevertheless, unlike in the Independence Money Pump (discussed later), at least probability discounters would be paying for something, namely, for a small increase in the probability of a positive outcome (from just below the discounting threshold to just above it). Therefore, this money pump is not as worrisome as the Independence Money Pump.[24] Furthermore, it might be

[22]Reducing probabilities just above the discounting threshold is discussed in Monton (2019, §6.3).

[23]If utility is bounded, the expected utility of a t chance of any positive outcome might be low. However, then Probability Discounting would be redundant, as Expected Utility Theory would no longer have counterintuitive implications in cases that involve very small probabilities of huge payoffs—at least if the upper bound is not very high and the lower bound not very low.

[24]Resolute Choice, Myopic Choice and Self-Regulation (discussed later) do not help in the Continuity Money Pump because this money pump is not dynamic like the Independence Money Pump. Also, Avoid Exploitable Plans and Avoid Dominated Plans (discussed later) do not help avoid exploitation because A_t^- is not dominated by $A_t qC$ as these prospects give slightly different proba-

argued that agents who maximize expected utility with an unbounded utility function are also vulnerable to schemes that are arbitrarily close to exploitation, and, indeed, this is what Pascal's Mugging illustrates. They will accept gambles that are arbitrarily close to a certain loss as long as the payoff in the small-probability state is great enough. However, unlike probability discounters, they will not pay a fixed amount for arbitrarily small changes in probabilities. The Continuity Money Pump illustrates how probability discounters, who wish to ignore very small probabilities, do care a great deal about very small *changes* in probabilities.[25,26]

To summarize, this section discussed three ways in which Probability Discounting might violate Continuity. First, it showed that views that discount probabilities below some threshold violate Continuity. Next, it showed that views that discount probabilities up to some threshold violate Mixture Continuity. Lastly, it showed that views that ignore very-small-probability outcomes must violate either Continuity or Statewise Dominance. Preferences that violate Continuity in these ways are vulnerable to exploitation by a money pump. However, the Continuity Money Pump is not as worrisome as the money pump for Independence because, in the

bilities.

[25] Similarly, Beckstead and Thomas (2020, §3.3) point out that Probability Discounting implies the following principle:

> **Threshold Timidity:** There is some discounting threshold such that, for any finite, positive payoffs x and y, getting x with probability below the threshold is never better than getting y with probability above the threshold—no matter how much better x is than y and no matter how close together the two probabilities may be.

Threshold Timidity states that, close to the discounting threshold, decreasing probability is infinitely more important than increasing expected utility.

[26] One possible response to the objection that probability discounters care about arbitrarily small changes in probabilities is that the discounting threshold is vague.

former, the agent is at least paying for something: a small increase in probability from just below the discounting threshold to just above it. Next, I will discuss the Independence Money Pump, which is a case of pure exploitation.

2 Independence

This section shows that Probability Discounting violates Independence. Then, it shows how violating Independence renders probability discounters vulnerable to exploitation in a money pump for Independence. §3 discusses possible ways of avoiding exploitation in this case.

2.1 A violation of Independence

To see how Probability Discounting violates Independence, consider the following prospects:[27]

> *Prospect* A_q Gives probability q of some very good outcome (and otherwise nothing).
>
> *Prospect* B Certainly gives a good outcome.
>
> *Prospect* C Certainly gives nothing.

Let q be a probability that is above the discounting threshold but less than one. Suppose that the very good outcome is sufficiently great so that A_q is better than

[27] Naive Discounting, Lexical Discounting, State Discounting, Stochastic Discounting and Tail Discounting all violate Independence in this case. See §5 and 6 of Chapter 4 on the latter two views and Independence.

B. Next, consider the following mixed lotteries (see table 3):

Independence Violation:

Prospect $A_q pC$ Gives a probability p of A_q and a probability $1-p$ of C (i.e., probability $p \cdot q$ of a very good outcome and otherwise nothing).

Prospect BpC Gives a probability p of B and a probability $1-p$ of C (i.e., probability p of a good outcome and otherwise nothing).

Given that B certainly gives a positive outcome, while A_q gives only a probability q of a positive outcome, we can mix A_q and B with C so that A_q mixed with C (i.e., $A_q pC$) gives only a negligible probability of a positive outcome but B mixed with C (i.e., BpC) gives a non-negligible probability of a positive outcome. This is so because there must be some probability $p \in (0,1)$ such that the result of q multiplied by p is below the discounting threshold, but p itself is above that threshold. Suppose that the outcomes in question are monetary and that the utility of money equals the monetary amount. Then, there must be some p such that the probability-discounted expected utility of $A_q pC$ is zero, but BpC has positive probability-discounted expected utility. In that case, Probability Discounting judges $A_q pC$ to be worse than BpC.

Table 3

A Violation of Independence

	p		$1-p$
	$p \cdot q$	$p(1-q)$	$1-p$
$A_q pC$	Very good	Nothing	Nothing
BpC	Good	Good	Nothing

Now, we have that A_q is better than B, but $A_q pC$ is worse than BpC for some $p \in (0, 1]$. This is a violation of the following axiom of Expected Utility Theory:

> **Independence:** If $X \succ Y$, then $XpZ \succ YpZ$ for all probabilities $p \in (0, 1]$.[28]

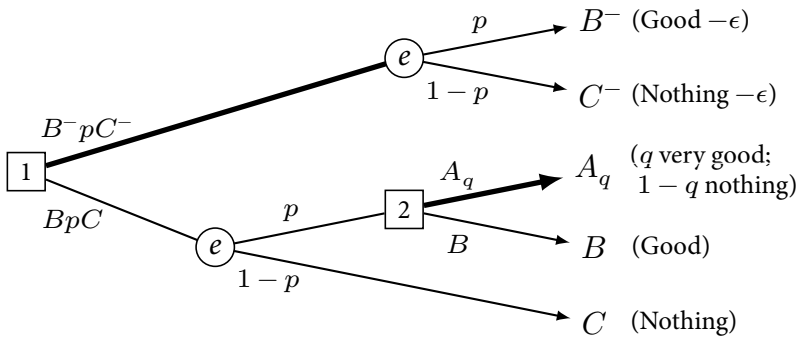
Informally, Independence is the idea that a lottery's contribution to the value of a mixed lottery does not depend on the other lotteries. The previous violation of Independence happens because, by mixing gambles together, one can reduce the probabilities of states or outcomes until their probabilities end up below the discounting threshold. As A_q gives a lower probability of a positive outcome than B does, with some values of p, $A_q pC$ only gives a negligible probability of a positive outcome, while BpC still gives a non-negligible probability.

[28] Jensen (1967, p. 173).

2.2 The Independence Money Pump

Violating Independence renders probability discounters vulnerable to exploitation in the Independence Money Pump. It goes as follows:[29]

THE INDEPENDENCE MONEY PUMP



$A_q \succ B$, and $BpC \succ B^-pC^- \succ A_qpC$.

The agent starts with prospect BpC: probability p of a good outcome and otherwise nothing. At node 1, the agent is offered a trade from BpC to B^-pC^-. B^-pC^- is just like BpC except that the agent has less money. If the agent turns down this trade and BpC results in the agent going up at chance node e, then at node 2, the agent will be offered a trade from B (certain good outcome) to A_q (probability q of a very good outcome and otherwise nothing). Both chance nodes depend on the same chance event e.

The agent can use *backward induction* to reason about this decision problem. This means that the agent considers what they would choose at later choice nodes and then takes those predictions into account when making choices at earlier choice

[29]This money pump is from Gustafsson (2021, p. 31n21). Also see Hammond (1988a, pp. 292–293), Hammond (1988b, pp. 43–45) and Gustafsson (forthcoming, §5).

nodes.[30] As the agent prefers A_q to B, they would accept the trade at node 2. By using backward induction at node 1, the agent can reason that the prospect of turning down the trade at node 1 is effectively A_qpC, and the prospect of accepting the trade is B^-pC^-. Given that the agent prefers BpC to A_qpC, it seems plausible that there is some price ϵ that they would be willing to pay to get the former instead of the latter. So, the agent pays that price and ends up with B^-pC^-. But they have ended up with B^-pC^- even though they could have kept BpC for free had they gone down at both choice nodes. Therefore, they have given up money for the exploiter.[31]

To summarize, this section showed that Probability Discounting violates Independence. This Independence violation happens because, by mixing gambles together, one can reduce the probabilities of states or outcomes until their associated probabilities are below the discounting threshold. As a result of violating Independence, probability discounters are vulnerable to exploitation in the Independence Money Pump. The next section discusses some possible ways of avoiding exploitation in this decision problem.

[30] Selten (1975) and Rosenthal (1981, p. 95).

[31] Also, as the chance nodes depend on the same event e, going up at node 1 is statewise dominated by going down at both choice nodes. See Gustafsson (forthcoming, §5).

3 Avoiding exploitation in the Independence Money Pump

This section discusses possible ways of avoiding exploitation in the Independence Money Pump. It argues that none of the standard views, such as Resolute Choice and Self-Regulation, work. It also argues that even if vulnerability to exploitation is not a sign of irrationality, Probability Discounting has untenable implications in a version of the Independence Money Pump that might result in a loss. But before discussing Resolute Choice and Self-Regulation, I will begin by discussing a foolish decision policy that nevertheless gives the right recommendation in the Independence Money Pump.

3.1 Myopic Choice

One decision policy that might help probability discounters is *Myopic Choice*. Myopic Choice advises an agent to choose at each choice node the option that currently seems best with no regard to what one will choose at later choice nodes.[32] But Myopic Choice is unjustifiable. It is irrational not to take one's future choices into account when making decisions. Nevertheless, one might be tempted to accept it as it gives the right recommendation in the Independence Money Pump.[33]

[32] Strotz (1955-1956) and von Auer (1998, p. 111). Myopic Choice is distinct from Naive Choice, on which one should choose the best available plan with no regard to whether one will in fact follow that plan. Similarly as a resolute agent (and unlike a myopic chooser), a naive chooser makes plans. However, unlike a resolute chooser, a naive chooser may not follow such plans.

[33] Myopic Choice is subject to problems that have nothing to do with Probability Discounting. For example, suppose a myopic agent starts with prospect *Bad*, which they can exchange for

If one accepts Myopic Choice, one will turn the offer down at node 1, thinking that one is choosing BpC. But if one ends up in node 2, one will choose A_q over B. So, with Myopic Choice, one can avoid getting money pumped in the Independence Money Pump.

However, probability discounters who use Myopic Choice are vulnerable to monetary exploitation in another decision problem. Recall the earlier prospects A_q, B and C:

> *Prospect* A_q Gives probability q of some very good outcome (and otherwise nothing).
>
> *Prospect* B Certainly gives a good outcome.
>
> *Prospect* C Certainly gives nothing.

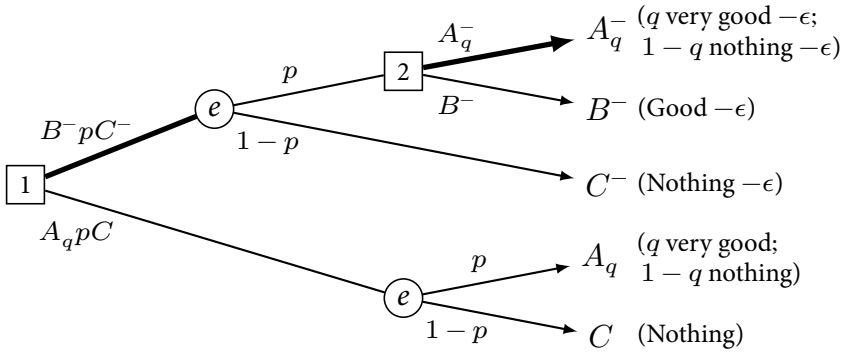
Now consider a reversed version of the Independence Money Pump:[34]

---

prospect *Terrible*. If they decide to do so, they are then offered prospect *Excellent*. A myopic chooser would refuse *Terrible*, and thus, they would end up with *Bad* when they could have had *Excellent*. However, probability discounters might accept some other (more plausible) principle that behaves like Myopic Choice in the Independence Money Pump. For example, they could restrict Myopic Choice to cases that include negligible probabilities.

[34]Note that backward induction gets one out of trouble in this money pump; by backward induction, accepting the trade at node 1 will effectively lead to prospect $A_q^- pC^-$, while rejecting the offer leads to $A_q pC$. Therefore, one should reject the offer.

The Reverse Independence Money Pump



$A_q \succ A_q^- \succ B^-$, and $B^-pC^- \succ A_qpC \succ A_q^-pC^-$.

Unlike in the Independence Money Pump, this time the agent starts with A_qpC. At node 1, they are offered B^-pC^-, which is like BpC except that the agent has ϵ less money. A myopic agent would choose B^-pC^-, given that they prefer it over A_qpC; there should be some amount ϵ that the agent is willing to pay to get BpC instead of A_qpC, given that they prefer the former. Then, if they end up in node 2, they are offered A_q^- in exchange for B^-. A_q^- is like A_q except that the agent has ϵ less money. Given that the agent prefers A_q over B (and B^-), it is again plausible that they are willing to pay some amount to get A_q instead of B (or B^-). So, the agent accepts the offer. They have now been money pumped; the agent chose $A_q^-pC^-$ even though they could have kept A_qpC for free by refusing the offer at node 1. So, Myopic Choice makes probability discounters vulnerable to exploitation in a reversed version of the Independence Money Pump.[35]

[35]The following case—let's call it *Discounter's Ruin*—also suggests that Probability Discounting should not be combined with Myopic Choice. Let the discounting threshold be (implausibly) just below 0.01. Now consider the following prospects:

3.2 Resolute Choice

Myopic Choice does not help probability discounters avoid monetary exploitation. But perhaps Resolute Choice will? A resolute agent chooses in accordance with any plan they have adopted earlier as long as nothing unexpected has happened since the adoption of the plan.[36] If one accepts Resolute Choice, one can make a plan that one will not trade B for A_q in node 2 of the Independence Money Pump. Even though one would usually prefer A_q over B, one is now committed to keeping B regardless. Consequently, one can safely refuse the trade at node 1, as one is then choosing BpC over $B^{-}pC^{-}$; one will not get money pumped nor end up with the inferior prospect A_qpC.

However, combining Probability Discounting with Resolute Choice gives untenable results in another case. Consider the following prospects:

Prospect C Certainly gives nothing.

Prospect E Gives probability r of some very bad outcome and prob-

---

Discounter's Ruin:

Prospect A Gives a 0.01 chance of \$10 and otherwise −1¢.

Prospect B Gives a 0.01 chance of \$10,000 and otherwise −\$10,

Prospect C Gives a 0.01 chance of \$10,000,000 and otherwise −\$10,000,

and so on for some large but finite number of prospects.
First, an agent is offered A, followed by an offer of B in case the agent wins \$10. As B is better than \$10, the agent would accept the offer. Then, if the agent wins \$10,000 with B, they are offered C. Again, the agent prefers C over \$10,000, so they would accept the offer. And so on for some large but finite number of offers. If one accepts all the offers, one would effectively choose an option that almost certainly results in a negative outcome and gives only a very small probability—a probability way below the discounting threshold—of a positive outcome. Thus, one would, effectively, not discount small probabilities down to zero.

[36]Strotz (1955-1956) and McClennen (1990, pp. 12–13). See Steele (2007), Steele (2018) and Gustafsson (forthcoming, §7) for criticism of Resolute Choice.

ability $1 - r$ of a barely positive outcome.

Prospect F Certainly gives a barely positive outcome.

Let r be a probability above the discounting threshold but less than $1 - r$ (i.e., less than 0.5). Suppose the very bad outcome in E is sufficiently bad so that C is better than E; certainly getting nothing is better than a non-negligible chance of a very bad outcome and otherwise a barely positive outcome.

Next, consider the following mixed lotteries (see table 4):

Independence Violation (Negative):

Prospect CpF Gives a probability p of C and a probability $1 - p$ of F (i.e., probability p of nothing and otherwise a barely positive outcome).

Prospect EpF Gives a probability p of E and a probability $1 - p$ of F (i.e., probability $p \cdot r$ of a very bad outcome and otherwise a barely positive outcome).

Given that r is less than $1 - r$, there must be some (relatively small) probability $p \in (0, 1)$ such that the result of r multiplied by p is below the discounting threshold, but the result of $1-r$ multiplied by p is above the discounting threshold. In that case, the possibility of obtaining a very bad outcome with EpF is ignored. However, given that $p(1 - r)$ is above the discounting threshold, EpF gives a greater probability of a barely positive outcome than CpF.[37] Consequently, EpF is better

[37] This is true whether one ignores very-small-probability outcomes or states. Thus, this argu-

than CpF. But now we have another violation of Independence: C is better than E, but EpF is better than CpF.[38] This violation of Independence happens because the probability of a very bad outcome is above the discounting threshold in E but below the discounting threshold in the mixed lottery EpF. Thus, the possibility of a very bad outcome is not ignored in E, but it is ignored in EpF.

Table 4
Independence Violation (Negative)

	p		$1-p$
	$p \cdot r$	$p(1-r)$	$1-p$
CpF	Nothing	Nothing	Barely positive
EpF	Very bad	Barely positive	Barely positive

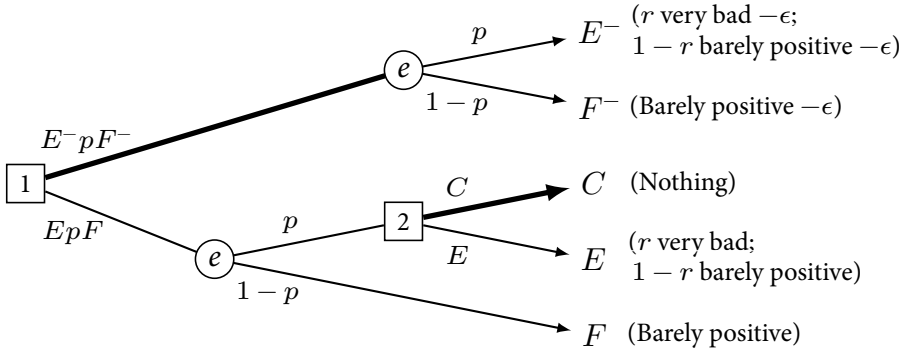
Let's go back to Resolute Choice and the Independence Money Pump. Recall that a probability discounter who uses Resolute Choice would commit to keeping B in node 2 of the Independence Money Pump (and thus avoid getting money pumped). In other words, they would commit to keeping a prospect that certainly gives a good outcome instead of trading it for a non-negligible chance of a very good outcome (and otherwise nothing). This does not seem untenable; one might bite the bullet and accept this implication. However, the same is not true in the

ment applies to all Naive Discounting, Lexical Discounting, State Discounting, Stochastic Discounting and Tail Discounting. If one ignores very-small-probability states and we take columns 2–4 in table 4 to correspond to states, then one ought to ignore column 2 (and not ignore columns 3 and 4). If one ignores very-small-probability outcomes, one ought to ignore the possibility of obtaining a very bad outcome with EpF (and not ignore the possibilities of the other outcomes). Either way, EpF gives a greater probability of a barely positive outcome than CpF.

[38] This violation of Independence is similar to the one discussed in Chapter 4 of this thesis.

following version of the Independence Money Pump:[39]

The Independence Money Pump (Negative)



$C \succ E$, and $EpF \succ E^{-}pF^{-} \succ CpF$.

In this case, the agent starts with EpF. At node 1, they are offered $E^{-}pF^{-}$, which is like EpF except that the agent has ϵ less money. If the agent refuses the trade and ends up in node 2, they are offered C in exchange for E. The agent prefers C over E; it is better to certainly get nothing that to choose a prospect that gives a non-negligible probability r of some very bad outcome and otherwise a barely positive outcome. Given that the agent would choose C over E at node 2, by using backward induction at node 1, the agent realizes that the choice is effectively between $E^{-}pF^{-}$ and CpF. And, similarly as before, the agent prefers $E^{-}pF^{-}$ over CpF, so they accept the offer. But then they have paid for something they could have kept for free.

In this case, a resolute agent can avoid getting money pumped if they commit to keeping E at node 2. However, unlike in the earlier money pump, this time the

[39] As before, the structure of this money pump is from Gustafsson (2021, p. 31n21).

resolute choice seems unreasonable: The agent would choose a prospect that gives a non-negligible probability r of some very bad outcome and otherwise a barely positive outcome over the certainty of getting nothing. Earlier, we assumed that r is above the discounting threshold but less than $1 - r$. So, it could be, for example, 0.49. Then, the agent would choose a prospect that gives a 0.49 probability of a very bad outcome and otherwise a barely positive outcome over certainly getting nothing. Furthermore, note that the very bad outcome can be arbitrarily bad, while the barely positive outcome can be arbitrarily close to getting nothing. No reasonable theory recommends making this choice.

Appeals to Resolute Choice seem to provide a general means of answering dynamic choice arguments against various patterns of preferences. However, Probability Discounting combined with Resolute Choice leads to disastrous results. Thus, Probability Discounting combined with Resolute Choice is untenable as a theory of instrumental rationality.

3.3 Self-Regulation

Another decision policy that has been proposed as a solution to money pumps is *Self-Regulation*.[40] Self-Regulation forbids (if possible) choosing options that may lead via a rationally permissible route to a final outcome that is unchoiceworthy by the agent's own lights.[41] The idea is that one ought not choose options that may (following one's preferences) lead to an outcome that one would not choose

[40] Self-Regulation helps avoid exploitation in money pumps against cyclic preferences. See Ahmed (2017). See Gustafsson (forthcoming, §2) for criticism of Self-Regulation.

[41] Ahmed (2017, p. 1001).

in a direct choice of all final outcomes. Unlike Resolute Choice, Self-Regulation is forward-looking.[42] When an agent's present choices determine the options available to them in the future, they should now choose so that their future choices lead to what they now consider acceptable in light of what is now available.[43] If the agent now wants to avoid some final outcome O, and they know what they are going to do at later choice nodes, then they should (if possible) now choose in such a way that, given those later choices, they will not end up with O.[44]

Self-Regulation in its original formulation does not help in the Independence Money Pumps, as it was intended for money pumps that do not involve chance.[45] However, the Independence Money Pumps involve chance nodes, so the agent does not know what the final holdings will be. One way to adapt Self-Regulation to cases that involve chance is to apply it to plans. A plan specifies a sequence of choices to be taken by an agent at each choice node that can be reached from that node while following this specification. Self-Regulation with respect to plans then states the following:

---

[42] Self-Regulation also differs from backward induction. Suppose that one's preferences are cyclic, so that A is better than B, which is better than C, which is better than A (and also suppose that A^- is better than B). Then, in all decision problems, Self-Regulation forbids (if possible) choosing an option that would lead to A^- via a rationally permissible path because one would not choose A^- in a direct choice of A, B, C and A^- (as A^- is dominated by A). But one might then end up with B, which is worse than A^-. In that case, backward induction would recommend choosing the option that leads to A^- (because A^- is better than B). However, Self-Regulation would recommend choosing the option that leads to B (because A^- would not be chosen in a direct choice of all final outcomes). See Ahmed (2017).

[43] Ahmed (2017, p. 1013).

[44] Ahmed (2017, p. 1003).

[45] Rabinowicz (2021, n. 13) writes: "[H]e [Ahmed, 2017] only shows how self-regulation allows the agents with cyclic preferences to avoid dynamic inconsistency. It is unclear whether and how this approach can be extended to agents who violate Independence."

> **Self-Regulation for Plans (i.e., Avoid Unchoiceworthy Plans):** If possible, one ought not choose options that may (following one's preferences) lead one to follow a plan that one would not choose in a direct choice of all plans (assuming one was able to commit to following some available plan).

Self-Regulation for Plans is a partial characterization of what it means to follow one's preferences: It involves, if possible, not choosing options that may, following one's preferences, lead one to follow an unchoiceworthy plan. A forward-looking choice rule C is self-regulating if and only if it tells you, at each node x, to choose a safe option whenever one is available. An option is 'safe' if and only if subsequently acting in accordance with C will lead you to follow a plan that is permissible at x.

The available plans at node 1 of the Independence Money Pump correspond to prospects A_qpC, BpC and B^-pC^-. One would not choose A_qpC or B^-pC^- in a direct choice between these plans. Therefore, one should not (if possible) choose any option that may lead via a rationally permissible route to one following A_qpC or B^-pC^-. However, both accepting and rejecting the trade at node 1 of the Independence Money Pump lead the agent to follow one of these plans via rationally permissible routes. Rejecting the offer leads one to follow A_qpC; accepting the offer leads one to follow B^-pC^-. So, Self-Regulation for Plans is silent in this case because it is not possible to make choices that do not lead to unchoiceworthy plans via rationally permissible routes. Thus, Self-Regulation for Plans does not help avoid exploitation in the Independence Money Pump.

3.4 Alternative decision policies

Instead of accepting Self-Regulation for Plans, one might restrict the set of forbidden plans and accept the following decision rule:

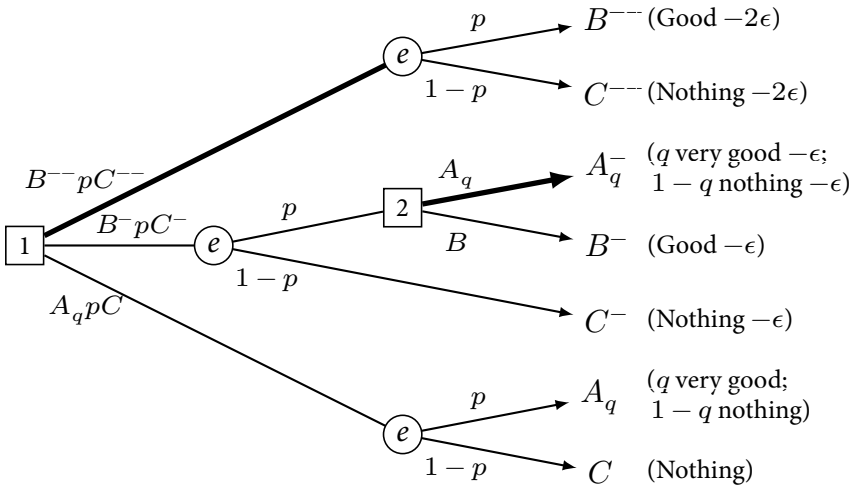
> **Avoid Exploitable Plans:** If possible, one ought not choose options that may (following one's preferences) lead one to pay for a plan that one could keep for free.

Avoid Exploitable Plans forbids accepting the trade at node 1 of the Independence Money Pump because accepting it would be paying for something that one could keep for free. However, Avoid Exploitable Plans does not forbid choosing A_q over B (or C over E) at node 2 because doing so would not be paying for a plan that one could keep for free. Thus, at node 2, an agent using Avoid Exploitable Plans would choose A_q over B (and C over E), given that they prefer the former. So, if one uses Avoid Exploitable Plans, one can avoid getting money pumped and also avoid the conclusion that one should keep E at node 2 of the version of the Independence Money Pump that includes negative payoffs.

But in the following decision problem, someone using Avoid Exploitable Plans would pay a higher price for something they could have obtained cheaper:[46]

[46]It might be objected that expected utility maximizers must also end up worse off than they could have been in some cases with an infinite series of trades. See for example Gustafsson (forthcoming, §8). However, expected utility maximizers might argue that there is a difference between not choosing the best option and paying more than one needs to, as the latter involves freely giving up what one already possesses while the former does not. But this kind of status quo bias may not be rationally justified.

The Three-Way Independence Money Pump



$A_q \succ A_q^- \succ B^- \succ B^{--}$, and $B^-pC^- \succ B^{--}pC^{--} \succ A_qpC \succ A_q^-pC^-$.

In this case, the agent starts with A_qpC. At node 1, they are offered B^-pC^- and $B^{--}pC^{--}$. $B^{--}pC^{--}$ is like B^-pC^- except that the agent has even less money (-2ϵ vs. $-\epsilon$). If the agent chooses B^-pC^- and ends up in node 2, then they are offered A_q^- in exchange for B^-. As the agent prefers A_q^- to B^-, they would accept the offer. So, choosing B^-pC^- at node 1 means effectively choosing $A_q^-pC^-$, given one's later choices. An agent who uses Avoid Exploitable Plans would therefore choose $B^{--}pC^{--}$; they prefer $B^{--}pC^{--}$ over A_qpC and $A_q^-pC^-$, and choosing it does not mean the agent is paying for something they could keep for free (as the agent starts with A_qpC). However, as B^-pC^- is also available, the agent has paid more than they needed to for BpC. They could have paid just ϵ instead of 2ϵ had they chosen B^-pC^- at node 1 (and then kept B^- at node 2).

The focus on avoiding monetary exploitation may be misplaced. Instead, one

might prefer adopting a decision rule that forbids all dominated plans whether or not they involve monetary exploitation:[47]

> **Avoid Dominated Plans:** If possible, one ought not choose options that may (following one's preferences) lead one to pay more for a plan that one could obtain for less money.

Avoid Dominated Plans forbids accepting the offer at node 1 of the Independence Money Pump because $B^{-}pC^{-}$ is dominated by BpC. Also, with this decision rule, one can refuse both offers of the Three-Way Independence Money Pump and keep A_qpC. One should refuse $B^{--}pC^{--}$ because it is dominated by $B^{-}pC^{-}$. And, one should refuse $B^{-}pC^{-}$ because choosing it means one is effectively choosing $A_q^{-}pC^{-}$, and $A_q^{-}pC^{-}$ is dominated by A_qpC. So, one should keep A_qpC. Avoid Dominated Plans thus allows an agent to avoid paying too much in this decision problem.

However, Avoid Dominated Plans seems a too narrow decision policy. Self-Regulation for Plans forbids choices that lead to plans that are unchoiceworthy by the agent's own lights. In contrast, Avoid Dominated Plans only forbids choices that lead to dominated plans but allows choices that lead to unchoiceworthy plans (such as A_qpC). It seems difficult to motivate such a decision policy. Why would it be irrational to choose an option that leads to a dominated plan (such as $B^{--}pC^{--}$) but not irrational to choose an option that leads to an unchoiceworthy plan (such as

[47] Avoid Dominated Plans is formulated in terms of monetary dominance: One should avoid plans that one can obtain for less money. But one should surely avoid plans that are dominated in other ways as well. More generally, one should avoid plans that are dominated with respect to anything valuable.

$A_q pC$)? Allowing the latter but forbidding the former seems arbitrary. Moreover, it leads one to something that is worse than the dominated plan, namely, $A_q pC$.

Furthermore, if we change the probabilities in the Independence Money Pump slightly, then Avoid Dominated Plans no longer avoids exploitation, at least entirely. Now, instead of $B^- pC^-$, the agent faces $B^- qC^-$, where q is arbitrarily close to p (and $q < p$). Then, given that $B^- qC^-$ and BpC do not give the exact same probabilities of the relevant outcomes, Avoid Dominated Plans no longer forbids accepting the trade at node 1; it is not the case that $B^- qC^-$ is like BpC except that the agent has less money, so Avoid Dominated Plans is silent. Consequently, a probability discounter who uses Avoid Dominated Plans will choose $B^- qC^-$ even though they could have kept BpC for free, and q is arbitrarily close to p. They have therefore given a fixed payment ϵ for an arbitrarily small increase in the probability of a positive outcome. So, Avoid Exploitable Plans is vulnerable to a scheme that is arbitrarily close to exploitation.[48]

3.5 How worrisome are the Independence Money Pumps?

Probability discounters might argue that these money pumps are not worrisome because, for example, the agent only really faces prospects $A_q pC$ and $B^- pC^-$ at node 1 of the Independence Money Pump, given that they would choose A_q at node 2. Thus, given the agent's preferences, in a way BpC is not even available to the agent. So, by choosing $B^- pC^-$, the agent does not end up paying for something they could have kept for free. However, a money-pump argument is supposed

[48] This objection also applies to Avoid Exploitable Plans.

to show that a given set of preferences is irrational because they lead to the agent paying for something they could have kept for free (if they had some other preferences). Therefore, it is not an adequate defense of those preferences that, given those preferences, the agent did not have any other option but to pay for something they could have kept for free. The target of the money pump is the structure of preferences. If one's preferences lead one to pay for something one could have kept for free (if one had some other preferences), then the money pump has succeeded in showing that those preferences are irrational.

Furthermore, even if being exploited is not a sign of irrationality as this argument claims, the violation of Independence in the case that includes negative payoffs (see table 4) is worrisome independently of the exploitation it leads to. This violation of Independence is particularly counterintuitive because EpF is considered better than CpF no matter how bad the very bad outcome is as long as the barely positive outcome is at least slightly positive. Moreover, the agent would choose to lock in a choice of keeping E (at node 2) if that was somehow possible at node 1.[49] This means they would lock in a choice of a prospect that gives a 0.49 probability of a very bad outcome and otherwise a barely positive outcome over certainly getting nothing. This seems irrational. So, even if probability discounters do not accept Resolute Choice, they would still make the same choice of E over C if offered the chance to lock in the choice at node 1.[50] This makes Probability Discounting less

[49]The agent would, therefore, also avoid costless information. More generally, agents who violate Independence avoid costless information. See for example Wakker (1988), Hilton (1990) and Machina (1989, p. 1638–1639).

[50]See §6 in Chapter 4 of this thesis.

plausible as a theory of instrumental rationality.[51]

To conclude, this section discussed possible ways of avoiding exploitation in the Independence Money Pump. First, it showed that, although Myopic Choice avoids exploitation in the Independence Money Pump, it does not avoid exploitation in the Reverse Independence Money Pump. Resolute Choice, in turn, leads to untenable results in the negative version of the Independence Money Pump, and Self-Regulation for Plans does not avoid exploitation in the Independence Money Pump. An agent who uses Avoid Exploitable Plans would pay too much for a plan in the Three-Way Independence Money Pump. Avoid Dominated Plans solves the Three-Way Independence Money Pump, but it is vulnerable to a scheme that is arbitrarily close to pure exploitation.

It was also argued that locking in the choice of E over C at node 2 of the negative version of the Independence Money Pump is irrational—and that this is something probability discounters would do regardless of whether they accept Resolute Choice or not. So, even if vulnerability to exploitation is not a sign of irrationality, Probability Discounting has untenable implications in the negative version of the Independence Money Pump. All in all, what we learn from these money pumps is that the various possible ways of avoiding exploitation do not ultimately work.[52] In

[51]It is worth pointing out that, independently of Probability Discounting, agents with unbounded utilities are also vulnerable to money pumps because they violate countable generalizations of the Independence axiom. See Russell and Isaacs (2021).

[52]The money pump arguments against Probability Discounting should be persuasive even for those who reject Independence for other reasons (e.g., due to the Allais paradox), as they might use Resolute Choice to avoid exploitation in the money pumps for Independence. However, as argued above, this solution is not available to probability discounters. In contrast, the Continuity Money Pump is not particularly worrying for probability discounters who already violate Continuity for other reasons.

addition, we learn that Probability Discounting gives untenable implications even if exploitation is not a sign of irrationality.

4 Conclusion

Probability Discounting is one way to avoid fanatical choices in cases that involve tiny probabilities of huge payoffs. However, it faces some serious problems. First, this chapter discussed three ways in which Probability Discounting might violate Continuity. It was shown that views that discount probabilities below some discounting threshold violate Continuity. Secondly, it was shown that views that discount probabilities up to some discounting threshold violate Mixture Continuity. Lastly, it was shown that views that ignore very-small-probability outcomes must violate either Continuity or Statewise Dominance. As a result of these Continuity violations, Probability Discounting is vulnerable to exploitation in the Continuity Money Pump.

In addition to violating Continuity, Probability Discounting also violates Independence, which renders probability discounters vulnerable to exploitation in the Independence Money Pump. Some possible ways of avoiding exploitation in the Independence Money Pump were discussed. However, these either failed to avoid exploitation in some version of the Independence Money Pump or they had otherwise untenable implications. It was also argued that even if vulnerability to exploitation is not a sign of irrationality, Probability Discounting has untenable implications in the negative version of the Independence Money Pump.

To conclude, this chapter has shown that Probability Discounting is vulnerable to exploitation in the money pumps for Independence and Continuity. The former is more worrisome than the latter, and it is difficult to see how Probability Discounting can respond to this challenge.

References

Ahmed, A. (2017), 'Exploiting cyclic preference', *Mind* **126**(504), 975–1022.

Beckstead, N. (2013), On the overwhelming importance of shaping the far future, PhD thesis, Rutgers, the State University of New Jersey.

Beckstead, N. and Thomas, T. (2020), 'A paradox for tiny probabilities and enormous values'. Global Priorities Institute Working Paper No.10.
URL: *https://globalprioritiesinstitute.org/nick-beckstead-and-teruji-thomas-a-paradox-for-tiny-probabilities-and-enormous-values/*

Bostrom, N. (2009), 'Pascal's Mugging', *Analysis* **69**(3), 443–445.

Goodsell, Z. (2021), 'A St Petersburg paradox for risky welfare aggregation', *Analysis* **81**(3), 420–426.

Gustafsson, J. E. (2021), 'The sequential dominance argument for the Independence axiom of Expected Utility Theory', *Philosophy & Phenomenological Research* **103**(1), 21–39.

Gustafsson, J. E. (forthcoming), *Money-Pump Arguments*, Cambridge University Press, Cambridge.

Hájek, A. (2014), 'Unexpected expectations', *Mind* **123**(490), 533–567.

Hammond, P. J. (1988a), 'Orderly decision theory: A comment on professor seidenfeld', *Economics and Philosophy* **4**(2), 292–297.

Hammond, P. J. (1988b), 'Consequentialist foundations for expected utility', *Theory and Decision* **25**(1), 25–78.

Hammond, P. J. (1998), Objective expected utility: A consequentialist perspective, *in* S. Barberà, P. J. Hammond and C. Seidl, eds, 'Handbook of Utility Theory Volume 1: Principles', Kluwer, Dordrecht, pp. 143–211.

Herstein, I. N. and Milnor, J. (1953), 'An axiomatic approach to measurable utility', *Econometrica* **21**(2), 291–297.

Hilton, R. W. (1990), 'Failure of Blackwell's theorem under Machina's generalization of expected-utility analysis without the Independence axiom', *Journal of Economic Behavior & Organization* **13**(2), 233–244.

Isaacs, Y. (2016), 'Probabilities cannot be rationally neglected', *Mind* **125**(499), 759–762.

Jensen, N. E. (1967), 'An introduction to Bernoullian utility theory: I. Utility functions', *The Swedish Journal of Economics* **69**(3), 163–183.

Kreps, D. M. (1988), *Notes on the Theory of Choice*, Westview Press, Boulder.

Luce, R. D. and Raiffa, H. (1957), *Games and Decisions: Introduction and Critical Survey*, Wiley, New York.

Lundgren, B. and Stefánsson, H. O. (2020), 'Against the De Minimis principle', *Risk Analysis* **40**(5), 908–914.

Machina, M. J. (1989), 'Dynamic consistency and non-expected utility models of choice under uncertainty', *Journal of Economic Literature* **27**(4), 1622–1668.

McClennen, E. F. (1990), *Rationality and Dynamic Choice: Foundational Explorations*, Cambridge University Press, Cambridge.

Monton, B. (2019), 'How to avoid maximizing expected utility', *Philosophers' Imprint* **19**(18), 1–24.

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

Peterson, M. (2020), The St. Petersburg paradox, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', fall 2020 edn, Metaphysics Research Lab, Stanford University.

Rabinowicz, W. (2008), Pragmatic arguments for rationality constraints, *in* M. C. Galavotti, R. Scazzieri and P. Suppes, eds, 'Reasoning, Rationality, and Probability', CSLI, Stanford, CA, pp. 139–163.

Rabinowicz, W. (2021), Between sophistication and resolution—wise choice, *in* R. Chang and K. Sylvan, eds, 'The Routledge Handbook of Practical Reason', Routledge, Abingdon, Oxon; New York, NY, pp. 526–540.

Rosenthal, R. W. (1981), 'Games of perfect information, predatory pricing and the chain-store paradox', *Journal of Economic Theory* **25**(1), 92–100.

Russell, J. S. (2021), 'On two arguments for fanaticism'. Global Priorities Institute Working Paper 17–2021.

URL: *https://globalprioritiesinstitute.org/on-two-arguments-for-fanaticism-jeff-sanford-russell-university-of-southern-california/*

Russell, J. S. and Isaacs, Y. (2021), 'Infinite prospects', *Philosophy and Phenomenological Research* **103**(1), 178–198.

Savage, L. J. (1951), 'The theory of statistical decision', *Journal of the American Statistical Association* **46**(253), 55–67.

Selten, R. (1975), 'Reexamination of the perfectness concept for equilibrium points in extensive games', *International Journal of Game Theory* **4**(1), 25–55.

Smith, N. J. J. (2014), 'Is evaluative compositionality a requirement of rationality?', *Mind* **123**(490), 457–502.

Smith, N. J. J. (2016), 'Infinite decisions and rationally negligible probabilities', *Mind* **125**(500), 1199–1212.

Steele, K. (2007), Precautionary Decision-Making: An Examination of Bayesian Decision Norms in the Dynamic Choice Context, PhD thesis, University of Sydney.

Steele, K. (2018), Dynamic decision theory, *in* S. O. Hansson and V. F. Hendricks, eds, 'Introduction to Formal Philosophy', Springer, Cham, pp. 657–667.

Strotz, R. H. (1955-1956), 'Myopia and inconsistency in dynamic utility maximization', *The Review of Economic Studies* **23**(3), 165–180.

von Auer, L. (1998), *Dynamic Preferences, Choice Mechanisms, and Welfare*, Springer-Verlag, Berlin.

von Neumann, J. and Morgenstern, O. (1947), *Theory of Games and Economic Behavior*, 2 edn, Princeton University Press, Princeton.

Wakker, P. (1988), 'Nonexpected utility as aversion of information', *Journal of Behavioral Decision Making* **1**(3), 169–175.

Wilkinson, H. (2022), 'In defence of fanaticism', *Ethics* **132**(2), 445–477.

Yudkowsky, E. (2007), 'Pascal's Mugging: Tiny probabilities of vast utilities'.
URL: *http://www.overcomingbias.com/2007/10/pascals-mugging.html*

Chapter 6

Tiny Probabilities and the Value of the Far Future[*]

ABSTRACT: Morally speaking, what matters the most is the far future—at least according to Longtermism. The reason why the far future is of utmost importance is that our acts' expected influence on the value of the world is mainly determined by their consequences in the far future. The case for Longtermism is straightforward: Given the enormous number of people who might exist in the far future, even a tiny probability of affecting how the far future goes outweighs the importance of our acts' consequences in the near term. However, it seems that there is something wrong with a theory that lets very small probabilities of huge payoffs dictate one's course of action. If, instead, we discount very small probabilities down to zero, we may have a response to Longtermism provided that its truth depends on tiny probabilities of vast value. Contrary to this, I will argue that discounting small probabilities does not undermine Longtermism.

[*]I wish to thank Gustav Alexandrie, Andreas Mogensen, Teruji Thomas, Hayden Wilkinson, participants of GPI's Early Career Conference Programme 2021 and the audience of the 8th Oxford Workshop on Global Priorities Research for valuable feedback.

Morally speaking, what matters the most is the far future—at least according to the following view:[1]

> **Longtermism:** In the most important decision situations, our acts' expected influence on the value of the world is mainly determined by their possible consequences in the far future.

On this view, the far future is of utmost importance. In the most important decision situations, we can often simply ignore our acts' effects in the near future and instead focus on their effects in the distant future. Longtermism follows naturally from additive views of value, such as total utilitarianism. Given the enormous number of people who might exist in the far future, even a tiny probability of affecting how the far future goes outweighs the importance of our acts' consequences in the near term.[2] So, if we are in a position to foreseeably affect the far future, our influence in the near term is outstripped by our influence in the far future.[3] However, one might reasonably doubt that we can have probabilistic evidence for some acts resulting in better outcomes than the alternatives hundreds or thousands of years from now.

One way in which we might beneficially influence the far future—it has been

[1] MacAskill (2019) and Greaves and MacAskill (2021). Greaves and MacAskill (2021, p. 2) define Longtermism as the view according to which we should be particularly concerned with ensuring that the far future goes well, and *Strong Longtermism* as the view on which the impact on the far future is the most important feature of our actions today. They defend axiological and deontic versions of this thesis. The former states that far-future effects are the most important determinant of the value of our options, while the latter states that they are the most important determinant of what we ought to do. See Greaves and MacAskill (2021, p. 3). For discussions of related topics, see for example Bostrom (2003), Beckstead (2013) and Ord (2020).

[2] Greaves and MacAskill (2021, p. 1).

[3] 'Foreseeably' in this context means probabilistic evidence rather than certainty or knowledge. We need not be able to foresee the effects of our actions so long as we can assign probabilities to the possible outcomes, conditional on the available acts, such that the expected value is favorable.

argued—is by mitigating existential risks.[4] Existential risks are risks that threaten the destruction of humanity's long-term potential. Such risks might be posed by, for example, synthetic pathogens, artificial intelligence (AI) systems, asteroids or climate change. Extinction risks are one type of existential risk. Because humanity's future is potentially very long, even relatively small reductions in the net probability of existential catastrophe correspond to enormous increases in expected moral value.[5] So, it can be argued that even very small reductions of existential risk have an expected moral value greater than that of the provision of any near-term good, such as the direct benefit of saving one billion present-day lives.[6]

However, it seems that there is something wrong with a theory that lets tiny probabilities of huge value dictate one's course of action. At least, such a theory would give counterintuitive recommendations. Consider, for example, the following case:[7]

> **Pascal's Mugging:** A stranger approaches Pascal and claims to be an Operator from the Seventh Dimension. He promises to perform magic that will give Pascal an extra thousand quadrillion happy days in the Seventh Dimension if Pascal pays the mugger ten livres—money

[4]Bostrom (2013). I will focus on existential risk mitigation as it seems one of the best candidates for longtermist interventions in terms of importance and tractability. Some longtermists focus instead on positively influencing humanity's trajectory conditional on survival.

[5]Bostrom (2013).

[6]Bostrom (2013, pp. 18–19). Greaves and MacAskill (2021, p. 11) write that even if there are 10^{14} lives to come (one of their more conservative estimates), a one-millionth of one percentage point reduction in the near-term extinction risk would be equivalent to the value of a million lives saved. On their main estimate of 10^{24} expected future lives, this becomes 10^{16} lives saved.

[7]Bostrom (2009). This case is based on informal discussions by various people, including Eliezer Yudkowsky (2007).

that the mugger will use for helping very many orphans in the Seventh Dimension.

Pascal thinks that the stranger is almost certainly lying. However, the possible payoff is so enormous that he is forced to conclude that the expected utility of paying the mugger is positive.[8] Importantly, the mugger points out that as long as Pascal gives a non-zero probability to the mugger being able to reward him with any finite amount of utility, the mugger can increase the payoff until the offer has positive expected utility.[9] Consequently, expected utility maximization (with no bound on utilities) recommends that Pascal pay the mugger—and thus, it gives the intuitively wrong recommendation.

Another version of this case is relevant to the topic of this chapter. In this case, the mugger exploits Pascal's expected-utility-maximizing descendant by utilizing research on existential risk and the long-term potential of humanity:[10]

> **Pascal's Mugger Strikes Again:** A stranger in a pub tells Pascal that a secretive organization is preparing a deadly disease that will make Earth uninhabitable within the next two years. However, the brewery that makes a particular ale sold at the pub also develops cutting-edge vaccines, and they need another £2 to pay for their electricity bills, or else their supplier will shut the factory off. The stranger forgot his

[8] 'Utility' here can be interpreted as moral value or as a decision-theoretic construct representing the betterness of prospects. Moral value, in turn, should be understood as reflecting the importance or significance of an act or outcome from a moral perspective.

[9] This may not be true if utility is bounded as standard axiomatizations of expected utility maximization require. See for example Kreps (1988, p. 63) and §1 and §2.1 in Chapter 1 of this thesis.

[10] This case is from Balfour (2021).

> wallet at home but—he claims—Pascal can save humanity from this deadly disease by buying him a pint of this ale.

Again, Pascal thinks that the mugger is almost certainly lying. However, given that the future of humanity is at stake, buying a pint might be the right course of action.[11] The mugger also warns that Pascal will be mugged every waking moment for the rest of his life, not by the mugger, but by the future of humanity itself. The mugger argues that, as an expected utility maximizer, Pascal must always perform the action which seems least likely to condemn humanity to extinction: "[Y]ou'll need to maintain constant vigilance, thinking constantly about which of your actions is least likely to destroy humanity."[12]

These cases are silly. If one were confronted with claims such as these muggers', one would consider them outlandish. However, there are reasons to think that even outlandish propositions should be assigned a non-zero probability. For example, according to Bayesianism, conditionalization is the right way to respond to new evidence. So, on this view, if one assigns some proposition zero (subjective) probability, one will always continue to do so no matter the evidence one obtains. However, sufficiently strong evidence should convince one of the truths of even outlandish propositions. If the mugger takes Pascal for a visit in the Seventh Dimension, Pascal should consider the mugger's original offer more probable than before, and in

[11] One could object that, instead of buying a pint for the stranger, Pascal should donate that money to some organization that works to mitigate existential risk, as this is a more effective way of securing humanity's future. That seems right. However, if Pascal knows that he will not do so, then actualism would advise Pascal to buy a pint for the stranger.

[12] Balfour (2021, p. 123).

particular, not consider it impossible.[13] Therefore, even outlandish propositions should be assigned non-zero probabilities, albeit tiny ones.[14] However, provided that the probabilities and the utilities work out the right way, expected utility maximization (with no bound on utilities) implies that Pascal should pay the mugger.[15] More generally, it leads to

> **Probability Fanaticism:** For any tiny probability $p > 0$, and for any finite utility u, there is some large enough utility U such that probability p of U (and otherwise nothing) is better than certainty of u.[16]

In response to cases that involve tiny probabilities of huge payoffs, some have argued that we ought to discount very small probabilities down to zero—let's call this *Probability Discounting*.[17] If we are indeed rationally required or permitted to discount small probabilities, then we may have an argument against Longtermism provided that its truth depends on tiny probabilities of huge value. In fact, this may

[13] Pascal might still think that he was probably, for example, hallucinating rather than visiting the Seventh Dimension. However, if the mugger gave him the ability to visit the Seventh Dimension repeatedly, he should not consider the mugger's original proposition impossible, even if hallucinating is still the most likely explanation.

[14] For a related discussion, see Francis and Kosonen (n.d.).

[15] Why not just bound utilities? This seems implausible, at least when it comes to ethical decisions. For example, this theory would imply that it is better to save some (very large) number n of lives for sure than to save *any number* of lives with a probability of almost one. See §4 in the introduction of this thesis.

[16] Wilkinson (2022, p.449). For discussions related to Probability Fanaticism, see Beckstead (2013, ch. 6), Beckstead and Thomas (2020), Goodsell (2021), Russell and Isaacs (2021) and Russell (2021).

[17] Monton (2019) argues that very small probabilities should be discounted down to zero, while Smith (2014) argues that one is rationally permitted—but not required—to do so. Smith argues that discounting very small probabilities allows one to get a reasonable expected utility for the Pasadena game (see [Nover and Hájek 2004]). See Hájek (2014), Isaacs (2016) and Lundgren and Stefánsson (2020) for criticisms of discounting small probabilities.

be one of the most plausible ways in which the argument for Longtermism might fail.[18] As mentioned above, one possible longtermist cause area is the mitigation of existential risk. However, the actions of a single individual are very unlikely to affect whether an existential catastrophe occurs.[19] The argument for prioritizing such actions is that if they make a difference, they might make an enormous one, such as delay human extinction by centuries, millennia, or more.[20]

This chapter argues that Probability Discounting does not undermine Longtermism. Even if one accepts a view on which small probabilities should be discounted down to zero, one should still consider the far future to be of utmost importance (or reject Longtermism for some other reason). I will discuss three arguments against Longtermism from discounting small probabilities. §2 discusses the argument that the probabilities of existential catastrophes are so low that one ought to ignore them. §3 discusses the argument that once we ignore very-small-probability scenarios, such as space settlement and digital minds, the expected number of lives in the far future is too small for Longtermism to be true. Lastly, §4 and §5 discuss the argument that the probability that an agent makes a difference to whether an existential catastrophe occurs or not is so small that it should be ignored. This chapter concludes that none of these arguments undermine Longtermism. Before going into these arguments, I will first say more about Probability Discounting. This

[18]Greaves and MacAskill (2021, p. 25). Besides discounting small probabilities, one could avoid letting tiny probabilities of huge value dictate one's course of action by having a bounded utility function. See for example Beckstead and Thomas (2020).

[19]In contrast, for some suitably capacious 'we,' we together might be likely to make a difference to net existential risk. I will discuss this in §4 and §5 of this Chapter.

[20]Greaves and MacAskill (2021).

chapter focuses on three versions of Probability Discounting: Naive Discounting, Tail Discounting and State Discounting.[21] Next, I will introduce Naive Discounting.

1 Discounting small probabilities

This section introduces one of the simplest versions of Probability Discounting. It also discusses choosing the threshold below which probabilities are small enough to be ignored.

Probability Discounting was originally proposed by Nicolaus Bernoulli.[22] He writes: "[T]he cases which have a very small probability must be neglected and counted for nulls, although they can give a very great expectation."[23] But when are probabilities small enough to be discounted? Or, as Buffon writes, "one can feel that it is a certain number of probabilities that equals the moral certainty, but what number is it?"[24] Some have suggested possible discounting thresholds. For Buffon and Condorcet, the discounting thresholds were 1 in 10,000 and 1 in 144,768 (respectively). Buffon chose his threshold because it is the probability of a 56-year-

[21] For a discussion of the different versions of Probability Discounting, see Chapter 4 of this thesis.

[22] Monton (2019) calls discounting small probabilities 'Nicolausian discounting' after Nicolaus Bernoulli.

[23] Pulskamp (n.d., p. 2). Discounting small probabilities is Bernoulli's solution to the St. Petersburg paradox.

[24] Hey et al. (2010, p. 256). Nicolaus Bernoulli also raised this problem: "It is necessary […] to determine as far as where the quantity of a probability must diminish, so that it be able to be counted null." See Pulskamp (n.d., p. 5).

old man dying in one day—an outcome reasonable people usually ignore.[25] Condorcet's justification for his threshold is that 1 in 144,768 is the difference between the probability that a 47-year-old man would die within 24 hours and the probability that a 37-year-old man would, and that difference would not keep anyone awake at night.[26]

It seems implausible that agents are rationally required to use some particular discounting threshold. Monton, who defends Probability Discounting, agrees. He argues that the discounting threshold is subjective within reason.[27] He would consider a threshold of $1/2$ irrational and some astronomically small threshold unreasonable. Nevertheless, there is no particular discounting threshold that all agents are rationally required to use. For Monton, the discounting threshold is approximately 1 in 2 quadrillion.[28] His justification for this threshold is that 1 in 2 quadrillion is between $1/2^{50}$ and $1/2^{51}$, and he treats the probability of getting tails at least 50 times in a row (with a fair coin) as a probability-zero event.

So, Probability Discounting is the idea that one should ignore sufficiently small probabilities—but small probabilities of *what*? On one version of this view, we should ignore *outcomes* associated with tiny probabilities. There is some threshold probability t such that outcomes whose probabilities are below this threshold are ignored.[29] Ignoring such outcomes can be done by conditionalizing on

[25] Hey et al. (2010, p. 257). See Monton (2019, pp. 8–9) for a discussion of Buffon's view.

[26] See Monton (2019, pp. 16–17).

[27] Monton (2019, §6.1) Note that this threshold may also be vague. See Lundgren and Stefánsson (2020, p. 911).

[28] Monton (2019, p. 17).

[29] Alternatively, one might have a threshold probability t such that outcomes whose probabilities

the supposition that an outcome of non-negligible probability occurs, where an 'outcome of non-negligible probability' is one whose associated probability is at least as great as the discounting threshold.[30] After conditionalization, options are compared by their 'probability-discounted expected utilities'. Let $X \succsim Y$ mean that X is at least as preferred as Y, and let $EU(X)_{pd}$ mean the expected utility of prospect X when tiny probabilities have been discounted down to zero (read as 'the probability-discounted expected utility of X'). Then, this version of Probability Discounting—let's call it *Naive Discounting*—states the following:[31]

> **Naive Discounting:** First, conditionalize on obtaining some outcome of non-negligible probability. Then, for all prospects X and Y, $X \succsim Y$ if and only if $EU(X)_{pd} \geq EU(Y)_{pd}$.

To summarize, Probability Discounting is the idea that very small probabilities should be ignored in practical decision-making. One of the simplest versions of this view is Naive Discounting, on which one should conditionalize on not obtaining outcomes associated with negligible probabilities. Next, I will consider an argument against Longtermism that someone with this view might give.

---

are at most as great as this threshold are ignored, but outcomes whose probabilities are greater than the threshold are not ignored.

[30] Smith (2014, p. 478).

[31] See §1 in Chapter 4 of this thesis.

2 Probability of an existential catastrophe

This section discusses the argument that the probabilities of existential catastrophes are so low that we should ignore them. However, it seems that even in the next century, existential risks have probabilities that are above any reasonable discounting thresholds. Naive Discounting faces a problem with individuating outcomes, so it is unclear what it says. Naive Discounting also violates dominance. Tail Discounting is a more plausible view, as it solves the outcome individuation problem and does not violate dominance. However, Tail Discounting does not ignore near-term extinction risks, so it does not undermine Longtermism in this way.

2.1 Existential risks in this century

It might be argued that existential catastrophes are so unlikely that we should ignore them—let's call this the *Low Risks Argument*.

> **Low Risks Argument:** The probabilities of existential risks are so tiny that we should ignore existential risks; we should evaluate options as though those risks are guaranteed not to eventuate.

This argument requires a reference to some time period: What is the relevant time period during which existential risks are unlikely to occur? After all, eventually, humanity will (almost certainly) go extinct. However, even in the next century, the net existential risk seems non-negligible. Ord (2020, p. 167) estimates that the probability of an existential catastrophe within the next 100 years is 1/6—way above any reasonable discounting threshold. The British Astronomer Royal Sir Martin

Rees has an even more pessimistic view. Rees (2003, p. 8) writes: "I think the odds are no better than fifty-fifty that our present civilization on Earth will survive to the end of the present century." Ord (2020, p. 167) gives the following estimates for existential catastrophes from specific causes within the next 100 years: 1 in 1,000,000 from asteroid or comet impact, 1 in 30 from engineered pandemics and 1 in 10 from unaligned AI (see table 1). Other estimates for *extinction* risks in the next 100 years are, for example, 1 in 15 billion from a 10 km+ asteroid colliding with the Earth,[32] between 1 in 600,000 and 1 in 50 from an extinction-level pandemic,[33] and a very conservative assessment would assign at least a 1 in 1000 chance to an AI-driven catastrophe that is as bad or worse than human extinction.[34]

Table 1
Existential and Extinction Risks in the Next 100 Years

	Existential risk (Ord, 2020)	Extinction risk (Others)
Asteroids	1 in 1,000,000*	1 in 15 billion
Pandemics	1 in 30**	1 in 600,000 to 1 in 50
AI	1 in 10	$\geq$ 1 in 1000

*=including comets, **=engineered pandemics.

If we individuate outcomes as 'human extinction from an asteroid impact in the

[32]The risk of a 10 km+ asteroid colliding with the Earth is estimated to be 1 in 150 million. See Ord (2020, p. 71). It is estimated that an asteroid with a 10 km+ diameter has at least a 1% chance of causing human extinction. See Newberry (2021, p. 3).

[33]Millett and Snyder-Beattie (2017).

[34]Greaves and MacAskill (2021, pp. 14–15). The expert median estimate for an AI-driven catastrophe is 5%. See Grace et al. (2018, p. 733).

next 100 years,' 'extinction-level pandemic in the next 100 years' and so on, then some extinction (and existential) risks are plausibly non-negligible. One should not ignore, for example, a 1 in 1000 chance of an AI-driven catastrophe in the next 100 years. However, if we individuate outcomes as 'extinction due an asteroid impact on the 4th of January 2055 at 13:00–14:00', 'extinction due to an asteroid impact on the 4th of January 2055 at 14:00–15:00' and so on, then extinction (and existential) risks might be negligible. It is difficult to see what the privileged way of individuating outcomes would be, and choosing one way over the others seems arbitrary. More generally, Naive Discounting faces the following problem:[35]

> **Outcome Individuation Problem:** If we individuate outcomes with too much detail, all outcomes have negligible probabilities. Is there a privileged way of individuating outcomes that avoids this?

If there is a plausible solution to the Outcome Individuation Problem, this solution should not tell one to ignore a net existential risk of 1/6 or a 1/10 risk of an AI-driven catastrophe.[36] Consequently, Naive Discounting does not undermine Longtermism, at least in this way. However, these relatively high estimates of existential risks have also been questioned.[37] Might we, after all, have a challenge to

[35] See also Beckstead and Thomas (2020, p. 13).

[36] One possible solution is to individuate outcomes by their utilities. See §1 in Chapter 4 of this thesis. However, this solution would imply that a human extinction on the 15th of February 2022 and one on the 16th February 2022 are distinct outcomes, given that their values are slightly different. Consequently, all possible extinction outcomes might have negligible probabilities, even if net extinction risk is high. This would secure the result that Probability Discounting undermines Longtermism. However, first individuating outcomes in this way and then applying Probability Discounting is absurd because net extinction risk could be arbitrarily high.

[37] See Luisa Rodriguez (2021) on the 80,000 Hours podcast for an informal discussion on this

Longtermism?

2.2 Tail Discounting

In addition to the Outcome Individuation Problem, Naive Discounting also faces other problems. For example, it violates dominance.[38] Instead, one might accept *Tail Discounting*, which states that one ought to ignore both the left and the right 'tails' of the distribution of possible outcomes when these outcomes are ordered by one's preference.[39] Tail Discounting is a more plausible version of Probability Discounting than Naive Discounting. However, it does not undermine Longtermism, even if the probability of an existential catastrophe is tiny.

Call the outcomes that fall in the middle of the distribution of possible outcomes 'normal outcomes'. Then, Tail Discounting states the following:[40]

---

topic. Rodriguez argues that humanity has a high probability of recovery from a non-extinction catastrophe and that for many of the threats, it is difficult to imagine a single sudden cataclysm that kills literally everyone.

[38]For example, Naive Discounting judges a prospect that saves a life with a negligible probability (and otherwise nothing happens) as equally good as a prospect that certainly saves no one. Using very-small-probability outcomes as tiebreakers (as 'Lexical Discounting' does) still violates Statewise Dominance in a more complicated case. See §2 in Chapter 4 of this thesis. Also see Isaacs (2016), Smith (2016), Monton (2019, pp. 20–21), Lundgren and Stefánsson (2020, pp. 912–914) and Beckstead and Thomas (2020, §2.3) on discounting small probabilities and dominance violations.

[39]Beckstead and Thomas (2020, 2.3). Unless one considers very-small-probability outcomes in cases of ties (as the definition of Tail Discounting given in this chapter does), Tail Discounting violates dominance reasoning.

[40]More formally, this view states the following:

> **Tail Discounting:** To determine $EU(X)_{pd}$, first order the possible outcomes of some prospect X from the least to the most preferred. Then, conditionalize on obtaining some outcome in the middle part of the distribution such that the following necessary conditions hold for all outcomes o that are not ignored:
>
> i The probability of obtaining an outcome that is at least as good as o is above the discounting threshold and

Tail Discounting: For all prospects X and Y, $X \succsim Y$ if and only if

- $EU(X)_{pd} > EU(Y)_{pd}$ or
- $EU(X)_{pd} = EU(Y)_{pd}$ and $EU(X) \geq EU(Y)$,

where $EU(X)_{pd}$ and $EU(Y)_{pd}$ are obtained by conditionalizing on the supposition that a normal outcome occurs.

Tail Discounting solves the Outcome Individuation Problem because, on this view, it does not matter how finely outcomes are individuated; one always ignores the tails of the distribution of possible final values. When the possible outcomes of a prospect are ordered by one's preference, the order of these outcomes will not change by individuating these outcomes more finely.

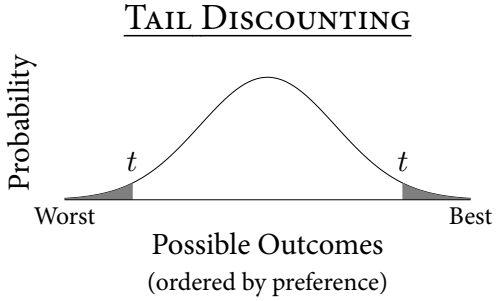
Next, suppose the possible outcomes of some prospect are normally distributed when they are ordered from the least to the most preferred. Then, Tail Discounting tells us to ignore the grey areas under the curve (the discounting threshold is denoted by t):

---

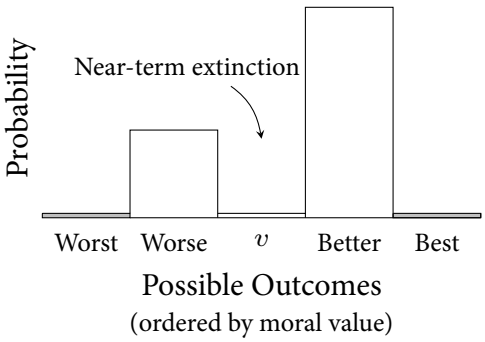
ii the probability of obtaining an outcome that is at most as good as o is above the discounting threshold.

If some outcome o fulfills the above necessary conditions, and

- the probability of obtaining an outcome that is better than o is below the discounting threshold, then decrease the probability of obtaining o until the total discounted probability of outcomes that are at least as good as o equals the discounting threshold (and conditionalize to make sure the remaining probabilities add up to 1), and
- if the probability of obtaining an outcome that is worse than o is below the discounting threshold, then decrease the probability of obtaining o until the total discounted probability of outcomes that are at most as good as o equals the discounting threshold (and conditionalize to make sure the remaining probabilities add up to 1).



What does Tail Discounting say about extinction risks? Suppose that the moral value of a near-term extinction is v. As long as v falls in the middle of the distribution of possible outcomes' values, Tail Discounting will not ignore the possibility of a near-term extinction. If there are non-negligible probabilities of worse and better outcomes than a near-term extinction, then near-term extinction scenarios may fall somewhere in the middle of the distribution of possible outcomes. Consider for example the following prospect:



In this case, the probability of a near-term extinction is tiny. However, the probability of obtaining an outcome that is at least as good as a near-term extinction is above the discounting threshold. Similarly, the probability of obtaining an

outcome that is at most as good as a near-term extinction is also above the discounting threshold. Consequently, Tail Discounting recommends against ignoring the possibility of a near-term extinction. Even if there is just a small probability of a near-term extinction and one can decrease this probability by just a small amount, Tail Discounting advises one to mitigate this risk (as long as the probabilities and the utilities work out the right way).

It seems plausible that the probabilities of both better and worse futures than a near-term extinction are above reasonable discounting thresholds. For example, the value of the world might be negative due to human and non-human animal suffering and continue to be negative in the future. Thus, there is a non-negligible probability that the future is worse than a near-term extinction. On the other hand, the value of the world might be net positive and continue to be so in the future. Alternatively, technological progress might increase well-being and create an overall positive future. Thus, there is a non-negligible probability that the future is better than a near-term extinction. Both better and worse possibilities seem non-negligible; neither is very unlikely. Consequently, someone who accepts Tail Discounting will not ignore the possibility of a near-term extinction. Tail Discounting only ignores outcomes with extreme values, and a near-term extinction event—plausibly—is not one.[41]

[41] One might object: So much the worse for Tail Discounting! By not advising one to ignore very-small-probability outcomes, such as (possibly) a human extinction, it fails to adequately capture the intuition behind Probability Discounting. Instead, one might accept a version of Tail Discounting on which one compares every prospect to some baseline prospect in the following way: First, calculate the differences in utilities between a given prospect and the baseline prospect in each state of nature. Next, order these differences from the largest loss to the largest gain. Then, ignore the right and left tails of this distribution. In effect, one is ignoring the possibility of a given prospect chang-

To summarize, I have discussed the Low Risks Argument: The probabilities of existential catastrophes are so low that we ought to ignore them. However, it seems that, even in the next century, the net existential risk and some specific existential risks have probabilities above any reasonable discounting thresholds. Naive Discounting faces the Outcome Individuation Problem, so it is unclear what it says; one can individuate existential catastrophes arbitrarily finely, and depending on how they are individuated, their associated probabilities may fall above or below the discounting threshold. However, an acceptable solution to this problem should not imply that one ought to ignore a net existential risk of 1/6 in the next century. Tail Discounting is more plausible than Naive Discounting, as it solves the Outcome Individuation Problem and does not violate dominance. However, as long as there are non-negligible probabilities of better and worse outcomes than a near-term extinction, Tail Discounting will not ignore near-term extinction risks, even if their associated probabilities are negligible.

To conclude, the Low Risks Argument does not undermine Longtermism. The next section discusses a second argument against Longtermism.

---

ing the value of the world by much. A prospect that lowers the probability of a near-term extinction will have a much higher value than the baseline prospect in some state of nature (namely, the state in which an extinction would have happened had the agent done nothing). This view—called *Baseline Tail Discounting*—will then ignore this large difference in value, assuming that it falls in the tail of the distribution of value differences. See Chapter 4 of this thesis on Baseline Tail Discounting (and a related view called Baseline Stochastic Discounting). However, the argument in §5 of the current chapter also shows (changing what needs to be changed) that Baseline Tail Discounting does not undermine Longtermism.

3 Size of the future

This section discusses the argument that once we ignore very-small-probability scenarios, such as space settlement and digital minds, the expected number of individuals in the far future is too small for Longtermism to be true. Contrary to this, I will argue that there are enough individuals in the far future in expectation for Longtermism to be true even if one accepts Probability Discounting.

3.1 Expected population sizes required for Longtermism

For Longtermism to hold, it also needs to be true that there is in expectation a sufficient number of individuals in the far future.[42] If in expectation the number of individuals is small no matter what we do, then it will not be true that even relatively small changes in the probability of an existential risk have great expected value. So, the argument goes, once we ignore very-small-probability scenarios, such as space settlement and digital minds, the expected number of future people becomes too small—let's call this the *Small Future Argument*.

> **Small Future Argument:** Once we ignore unlikely scenarios, the expected number of individuals in the far future is too small for Longtermism to be true.

Next, I will discuss whether or not there are enough individuals in the far future for existential risk mitigation to have a higher expected value than the neartermist

[42] More precisely, it is not the number of individuals but the amount of value that matters. There might be a great quantity of value in the far future even if the number of individuals is relatively small if these individuals live very long lives. See for example Gustafsson and Kosonen (n.d.).

causes. The cost-effectiveness of antimalarial bednet distribution may be used as an upper bound to attainable near-term benefits per unit of spending.[43] The distribution of insecticide-treated bednets in malarial regions saves a life on average for a little over $4000.[44] Suppose Shivani is thinking how to improve the world the most with her $10,000.[45] By donating to the Against Malaria Foundation, she can save in expectation 2.5 lives. Suppose that Longtermism is true in Shivani's situation if and only if, in expectation, more than 2.5 additional lives exist in the far future if she donates to some longtermist cause.[46]

An example of existential risk mitigation that longtermists might focus on is the detection and potential deflection of asteroids.[47] It is estimated that NASA's Spaceguard Survey, which tracks near-Earth objects in order to identify any on impact trajectories, reduced extinction risk by at least 1 in 2000 trillion per $100 spent.[48,49]

[43] Greaves and MacAskill (2021, p. 2).

[44] GiveWell (2020).

[45] This case is modified from Greaves and MacAskill (2021).

[46] Note that longtermist causes typically also create near-term benefits, and these near-term benefits might be great enough for existential risk mitigation to pass a cost-effectiveness analysis even if one ignores the far future effects of one's acts. Moreover, even if the near-term benefits are not sufficient on their own, the far future effects might add just enough expected value to make existential risk mitigation the best course of action (even though most of the expected influence of existential risk mitigation comes from near-term effects). This is irrelevant to whether or not one should mitigate existential risks, but it matters to whether Longtermism is true. This point is important. Even if Longtermism turns out to be false, existential risk mitigation might still be the right course of action. It is also worth noting that paradigmatic neartermist causes, such as distributing anti-malarial bednets, can also have foreseeable long-term effects.

[47] Greaves and MacAskill (2021, p. 11).

[48] Greaves and MacAskill (2021, p. 11). Coincidentally, this is Monton's threshold for discounting small probabilities ($5 \cdot 10^{-16}$). See Monton (2019, p. 17).

[49] Interestingly, in 2022 NASA will redirect an asteroid for the first time in human history (by slamming a spacecraft into it) for testing technologies that we may need in the future. Their target is a 500-foot-wide moon orbiting a half-mile-wide asteroid called Didymos. This moon is roughly the size of an asteroid that can obliterate cities. See Drake (2020).

But further work on asteroids is expected to have lower cost-effectiveness.[50] It is estimated that a 10 km+ asteroid has at least a 1% chance of causing human extinction if it collides with the Earth.[51] While the probability of a 10 km+ asteroid colliding with the Earth is on average 1 in 1.5 million per century, astronomers are confident that they have found all 10 km+ asteroids in at least 99% of the sky.[52] The remaining risk of a 10 km+ asteroid colliding with the Earth in the next 100 years is estimated to be 1 in 150 million.[53] Consequently, the probability of human extinction from an asteroid impact in the next 100 years is 1 in 15 billion.

The cost of detecting (with almost certainty) any remaining 10 km+ asteroids is estimated to be at most $1.2 billion, and we might assume that we can reduce extinction by 5% (relatively) if we detect one on a collision course.[54] Shivani's proportion of the $1.2 billion required to reduce the risk to (near) zero is 1/120,000. It is plausible that she would reduce the risk by the same proportion, that is, by 1 in 2.4 million.[55] Consequently, by donating $10,000 to asteroid detection, Shivani can provide a 1 in 33,000 trillion absolute reduction in the probability of extinction from an asteroid collision in the next 100 years.[56]

Another possible longtermist cause area is the prevention of extinction-level

[50] Greaves and MacAskill (2021, p. 11).

[51] Newberry (2021, p. 3).

[52] Ord (2020, p. 71).

[53] Ord (2020, p. 71).

[54] Newberry (2021, pp. 5–6). Inspired by the movie Don't Look Up, Lubin and Cohen (n.d.) estimate that humanity could, in theory, defend itself against a comet of a 10km diameter using existing technology even in the extreme case where it is detected just six months before impact.

[55] Greaves and MacAskill (2021, p. 16). $0.05 \cdot 10000/(1.2 \cdot 10^9) \approx 4 \cdot 10^{-7}$.

[56] $1/(15 \cdot 10^9) \cdot 0.05 \cdot 10000/(1.2 \cdot 10^9) \approx 3 \cdot 10^{-17}$ (1 in 33,000 trillion).

pandemics.[57] The risk of an extinction-level pandemic in the next 100 years is estimated to be between 1 in 600,000 and 1 in 50.[58] Taking the geometric mean of the two methods that generate the lower estimates for extinction risk gives a probability of about 1 in 22,000 for extinction from a pandemic over the next 100 years.[59] It is estimated that $250 billion spent on strengthening healthcare systems would reduce the chance of an extinction-level pandemic in the next 100 years by at least a proportional 1%.[60] Consequently, by donating $10,000 to pandemic prevention, Shivani can provide a 1 in 2.5 billion relative reduction and a 1 in 50 trillion absolute reduction in the probability of an extinction-level pandemic in the next 100 years.[61]

Lastly, another possible longtermist cause area is the prevention of an existential catastrophe due to artificial general intelligence.[62] In the most comprehensive study of its kind, AI experts estimated that the probability of an extremely bad outcome, such as human extinction, due to high-level machine intelligence (at any point in time) is 5%.[63] The same experts gave a 50% chance for high-level machine

[57] Greaves and MacAskill (2021, p. 12).

[58] Millett and Snyder-Beattie (2017).

[59] Greaves and MacAskill (2021, p. 12).

[60] Millett and Snyder-Beattie (2017, p. 379).

[61] $0.01 \cdot 10000/(250 \cdot 10^9) \approx 4 \cdot 10^{-10}$ (1 in 2.5 billion). $1/22000 \cdot 0.01 \cdot 10000/(250 \cdot 10^9) \approx 2 \cdot 10^{-14}$ (1 in 50 trillion).

[62] See for example Greaves and MacAskill (2021, pp. 14–15). GPT-3 (n.d.) disagrees: "There is no evidence that artificial general intelligence (AGI) is an existential threat. AGI has the potential to cause a lot of harm, but so far there is no evidence that it will be able to achieve a level of intelligence that would allow it to cause existential harm."

[63] Grace et al. (2018, p. 733). "High-level machine intelligence" is achieved when unaided machines can accomplish every task better and more cheaply than human workers. See Grace et al. (2018, p. 731).

intelligence occurring by 2061.[64] Given these survey results, even a very conservative estimate would assign at least a 0.1% chance to an AI-driven catastrophe as bad or worse than human extinction in the next 100 years.[65] Furthermore, it is plausible that $1 billion spent on AI safety would decrease the probability of such an outcome by at least 1%.[66] Consequently, $1 billion would provide at least a 0.001% absolute reduction in existential risk.[67] Thus, by donating $10,000 to AI safety, Shivani can provide a 1 in 10 million relative reduction and a 1 in 10 billion absolute reduction in the probability of an AI-driven catastrophe in the next 100 years.[68]

Shivani's options are as follows:

Shivani: Shivani has $10,000 to donate and she has four options:

i *Against Malaria Foundation* She saves in expectation 2.5 lives.

ii *Asteroid detection* She can provide a 1 in 33,000 trillion absolute reduction in the probability of extinction from an asteroid collision in the next 100 years.

iii *Pandemic prevention* She can provide a 1 in 50 trillion absolute reduction in the probability of an extinction-level pandemic in the next 100 years.

[64] Grace et al. (2018, p. 731).

[65] Greaves and MacAskill (2021, pp. 14–15).

[66] Greaves and MacAskill (2021, p. 15).

[67] Greaves and MacAskill (2021, p. 15).

[68] $0.01 \cdot 10000/10^9 = 10^{-7}$ (1 in 10 million). $0.001 \cdot 0.01 \cdot 10000/10^9 = 10^{-10}$ (1 in 10 billion).

iv *AI safety* She can provide a 1 in 10 billion absolute reduction in the probability of an AI-driven catastrophe in the next 100 years.

As mentioned earlier, we have assumed that Longtermism is true in Shivani's situation if and only if, in expectation, more than 2.5 additional lives exist in the far future if she donates to one of the longtermist causes. For it to be the case that over 2.5 additional lives exist in the far future if she donates to asteroid detection, the expected number of beings in the far future must be over 83,000 trillion.[69] Similarly, for it to be the case that over 2.5 additional lives exist in the far future if she donates to pandemic prevention, the expected number of beings in the far future must be over 125 trillion.[70] Finally, for it to be the case that over 2.5 additional lives exist in the far future if she donates to AI safety, the expected number of beings in the far future must be over 25 billion.[71] Is the expected number of lives in the far future large enough for Longtermism to be true in Shivani's situation?

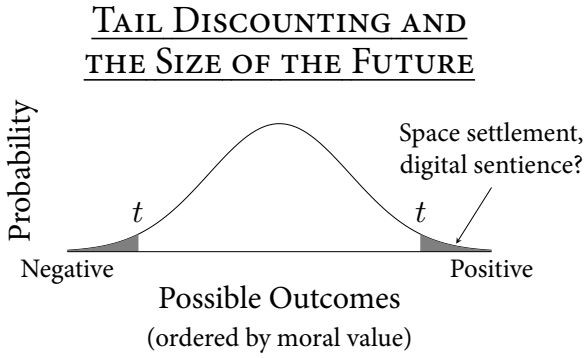
TABLE 2
EXPECTED POPULATION SIZES REQUIRED FOR LONGTERMISM

Asteroid detection	83,000 trillion
Pandemic prevention	125 trillion
AI safety	25 billion

[69] $8.3 \cdot 10^{16} \cdot 3 \cdot 10^{-17} \approx 2.5$.
[70] $1.25 \cdot 10^{14} \cdot 2 \cdot 10^{-14} = 2.5$.
[71] $2.5 \cdot 10^{10} \cdot 10^{-10} = 2.5$.

3.2 Is the size of the future large enough?

Longtermism might depend on the possibilities of space settlement or the creation of digital minds because these possibilities inflate the value of the future; given these possibilities, the stakes are so high that even small reductions in existential risks have enormous expected value. If Longtermism depends on these possibilities, Tail Discounting undermines Longtermism if obtaining an outcome at least as good as these is very unlikely. In that case, Tail Discounting would ignore these possibilities, and the size of the future would not be large enough for Longtermism to be true (see the graph below).



Space settlement and the creation of digital minds might be the kind of unlikely best-case scenarios Tail Discounting ignores. However, it seems that the number of expected lives in the far future is sufficiently large for the argument for Longtermism to go through, even if we ignore these very-small-probability scenarios.[72] This is because there might be in expectation a sufficient number of individuals in

[72] Greaves and MacAskill (2021, §3).

the future if humanity survives for a long time on Earth. Based on the estimate of extinction risk due to natural causes, the expected future lifespan of humanity is at least 87,000 years.[73] On the other hand, the average lifespan of hominins is around one million years. Assuming a constant population size of 11 billion and an average lifespan of 80 years, this would mean that the expected number of humans is 12 trillion if humanity lives for a further 87,000 years and 140 trillion if humanity lives for a further million years.[74]

So, if humanity lives for 87,000 years in expectation, then AI safety leads to Longtermism (given that 12 trillion is greater than the required 25 billion expected future lives). This means that if Shivani donates to AI safety, more than 2.5 additional individuals live in the far future in expectation—so Longtermism is true in her situation. However, asteroid detection and pandemic prevention do not lead to Longtermism, as the expected number of individuals is not large enough (conditional on ignoring the very-small-probability scenarios). However, if humanity lives for one million years in expectation, then pandemic prevention also leads to Longtermism (given that 140 trillion is greater than the required 125 trillion expected future lives). In that case, more than 2.5 additional individuals live in the far future in expectation if Shivani donates to pandemic prevention.

However, humans are an atypical species, so extinction risk due to natural

[73] Snyder-Beattie et al. (2019).

[74] $11 \cdot 10^9 \cdot 87000/80 \approx 1.2 \cdot 10^{13}$ and $11 \cdot 10^9 \cdot 1000000/80 \approx 1.4 \cdot 10^{14}$. The UN Department of Economic and Social Affairs projects the world population to plateau at 11 billion. See United Nations and Social Affairs (2019). However, there are also signs of population decline. See Bricker and Ibbitson (2019). Note that the higher the world population is, the easier it is for Longtermism to be true; one antimalarial bednet will always save just one (or at most a few) people, but asteroid detection, pandemic prevention and AI safety will affect everyone.

causes and the lifespan of a typical hominin species may not be suitable bases for estimates of humanity's lifespan. How long might humanity survive? Even if we only stay on Earth, we have around one billion years until the Earth becomes uninhabitable.[75] If humanity survives for a billion years with a constant population size of 11 billion and an average lifespan of 80 years, then the number of humans would be 140,000 trillion.[76] In that case, asteroid detection, pandemic prevention and AI safety all would lead to Longtermism. But of course, humanity may become extinct well before the Earth becomes uninhabitable. How long must humanity's future be for asteroid detection, pandemic prevention and AI safety to lead to Longtermism?

For asteroid detection to lead to Longtermism, humanity's expected lifespan (ignoring the tail outcomes) must be at least 600 million years (given a constant population size of 11 billion and a human lifespan of 80 years).[77] Then, the expected number of humans in the far future is above the required 83,000 trillion. Pandemic prevention, in turn, leads to Longtermism if humanity's expected lifespan is at least 900,000 years (again, given a constant population size of 11 billion and a human lifespan of 80 years). Then, the expected number of future beings is above the required 125 trillion.[78]

Lastly, how long must humanity's future be for AI safety to lead to Longtermism? Suppose that the far future starts after 100 years. The expected number of

[75] Adams (2008). In principle, it might be possible to stay on Earth and keep the planet habitable for longer by changing its orbit or through stellar engineering projects that increase the sun's lifespan.

[76] $11 \cdot 10^9 \cdot 10^9/80 \approx 1.4 \cdot 10^{17}$.

[77] $11 \cdot 10^9 \cdot 604 \cdot 10^6/80 > 8.3 \cdot 10^{16}$.

[78] $11 \cdot 10^9 \cdot 909091/80 > 1.25 \cdot 10^{14}$.

beings in the far future is sufficiently large (above 25 billion) if humanity's expected lifespan *in the far future* is at least 182 years (given a constant population size of 11 billion and a human lifespan of 80 years).[79] Assuming a constant risk of extinction per year, this will be the case if humanity's expected lifespan is 265 years (this includes humanity's expected lifespan in the near and the far future). So, for AI safety to lead to Longtermism, it would have to be the case that humanity's expected lifespan is at least 265 years.

It seems plausible that humanity's expected lifespan is at least 265 years. This would be true if the risk of extinction per year is at most 0.38%.[80] Assuming a constant risk throughout the next 100 years, Ord's (2020, p. 167) estimate for existential risk is below this.[81] So, even if the probability of human extinction were 1/6 in the next 100 years, this would still be low enough for AI safety to lead to

[79] $11 \cdot 10^9 \cdot 182/80 > 2.5 \cdot 10^{10}$.

[80] $1/0.00377 \approx 265$. With a 0.00377 risk of extinction per year, humanity's expected number of years in the far future (after the next 100 years) is

$$1/0.00377 - \sum_{n=1}^{100} (1 - 0.00377)^n \approx 182.$$

This includes the possibility that humanity survives for a very long time, even when unlikely. However, these outcomes do not contribute much to the expectation. For example, the probability that humanity survives at least 2000 years is $(1 - 0.00377)^{2000} \approx 0.0005$—a probability that is plausibly above the discounting threshold. The contribution of the next 2000 years to humanity's expected lifespan is

$$\sum_{n=1}^{2000} (1 - 0.00377)^n \approx 264.$$

This is close to the expected lifespan of humanity (265 years).

[81] Existential risk in the next 100 years is 1/6 if the risk per year is 0.18% ($(1 - 0.0018)^{100} \approx 0.835$.) This is lower than the maximum 0.38% probability of human extinction per year with which AI safety leads to Longtermism. Ord (2020) does not give an estimate for extinction risk in the next 100 years. However, he believes this to be significantly lower than 1/6 (personal correspondence).

Longtermism. However, the probability of human extinction is lower than 1/6, as human extinction is just one type of existential catastrophe. Thus, the case for Longtermism from AI safety is even stronger.

Furthermore, there are many factors that we have not taken into account. First, it seems plausible that the risk of extinction per year is not constant.[82] For example, there may be a few particularly dangerous moments expected to happen within the next couple of centuries, such as the development of artificial general intelligence, after which the yearly risk of extinction is significantly lower.[83] If we now live in a 'time of perils' after which the yearly risk of extinction is significantly lower, existen-

[82] According to the "Simple Model" of existential risk mitigation, the expected value of the future is

$$EU[F] = v \sum_{n=1}^{\infty} (1-r)^i = \frac{v(1-r)}{r},$$

where v is the value of human existence each century (assumed to be constant) and r is a per-century existential risk (also assumed constant). In this model, the value of the future is the value of a single century divided by the per-century risk. See Ord (2020, appendix E). This model implies that the value of reducing existential risk this century by some fraction f is $EU(X) = fv$. This result is surprising because the value of existential risk reduction is capped at the value v (an additional century of human existence)—it is not astronomical. See Thorstad (n.d.) for a discussion of the Simple Model. If human population stays at a constant 11 billion, each person living for 80 years, then the value of an additional century of human existence (measured in lives) is approximately 14 billion ($1.25 \cdot 11,000,000,000 = 13,750,000,000$). It was assumed that Shivani could save in expectation 2.5 lives by donating to Against Malaria Foundation. So, if the Simple Model is right, Shivani should donate to a longtermist cause if she can decrease (relatively) the probability of extinction by at least 1 in 5 billion ($2.5/13,750,000,000 \approx 2 \cdot 10^{-10}$).

[83] Thorstad (n.d.) argues that the belief that existential risks are high is unlikely to ground the overwhelming importance of existential risk mitigation unless coupled with the time of perils hypothesis. This is so because the higher the probability of existential risk per century, the shorter the expected lifespan of humanity is. Therefore, a high level of risk means the size of the future is correspondingly smaller. However, if we now live in a particularly dangerous period after which existential risk is much lower, then the size of the future can be considerable. However, Thorstad (n.d.) also argues that the time of perils hypothesis is probably false. Therefore, pessimism about existential risks does not justify the overwhelming importance of existential risk mitigation.

tial risk mitigation more easily leads to Longtermism.[84] Ord (2020, pp. 189–191) argues that humanity's first task is to reach existential security—a place where existential risk is low and stays low.

The size of the future seems large enough for Longtermism to be true—even if we ignore very-small-probability scenarios such as space settlement and digital minds. Asteroid detection leads to Longtermism if humanity's expected lifespan is at least 600 million years. With pandemic prevention and AI safety, the required expected lifespans are 900,000 and 265 years, respectively. Finally, it would be overconfident to be near-certain that space settlement or digital sentience will not occur, given that there is no known reason why they should be physically impossible.[85] If one gives a non-negligible probability for at least one of these scenarios, then the expected number of lives in the far future will be much greater. To conclude, even if we ignore very-small-probability scenarios, such as space settlement and digital minds, the expected number of lives in the far future seems large enough for Longtermism to be true. Thus, the Small Future Argument does not undermine Longtermism.

[84]The astronomer Sagan (1997, p. 173) writes about the time of perils: "Some planetary civilizations see their way through, place limits on what may and what must not be done, and safely pass through the time of perils. Others are not so lucky or so prudent, perish." Rees (2003, pp. 7-8) echoes this by writing that "the most crucial location in space and time (apart from the big bang itself) could be here and now. […] What happens here on Earth, in this century, could conceivably make the difference between a near eternity filled with ever more complex and subtle forms of life and one filled with nothing but base matter."

[85]The entrepreneur Elon Musk (n.d.) wants humanity to be a spacefaring civilization: "You want to wake up in the morning and think the future is going to be great—and that's what being a spacefaring civilization is all about. It's about believing in the future and thinking that the future will be better than the past. And I can't think of anything more exciting than going out there and being among the stars."

4 Probability of making a difference

This section discusses the argument that the probability of making a difference to whether or not an existential catastrophe occurs is tiny, and thus, we should ignore the possibility of influencing the occurrence of existential catastrophes. One type of Probability Discounting naturally captures this idea. However, I will show that the different versions of this view violate Statewise Dominance or Acyclicity, which makes them less plausible as theories of instrumental rationality.

4.1 State Discounting

The final objection to Longtermism from discounting small probabilities is that the probability of making a difference to whether or not an existential catastrophe occurs is so tiny that it should be discounted down to zero—let's call this the *No Difference Argument*.

> **No Difference Argument:** The probability of making a difference to whether or not an existential catastrophe occurs is so small that we should ignore the possibility of making a difference.

If it is indeed the case that Shivani has only a negligible probability of having an impact with all of the possible longtermist causes, and such small probabilities should be discounted down to zero, then she should instead donate to the Against Malaria Foundation. Consequently, Longtermism would be false in her situation.

Recall that the absolute reductions in the probability of extinction that Shivani can provide are 1 in 33,000 trillion with asteroid detection, 1 in 50 trillion with pan-

demic prevention and 1 in 10 billion with AI safety (see table 3). If Shivani plans to donate less than $10,000, her probability of impact is even smaller.[86] As these numbers are tiny, it may not be unreasonable to ignore the possibility of Shivani making a difference to existential risks with her donation to the longtermist causes.[87] But which version of Probability Discounting allows her to do this?

Table 3
Absolute Reductions of Extinction Risks with $10,000

Asteroid detection	1 in 33,000 trillion
Pandemic prevention	1 in 50 trillion
AI safety	1 in 10 billion

One version of Probability Discounting captures the No Difference Argument naturally. Recall that Naive and Tail Discounting ignore outcomes associated with small probabilities. However, one might ignore *states* associated with small probabilities instead—let's call this *State Discounting*.[88]

State Discounting For all prospects X and Y, $X \succsim Y$ if and only if

[86] Conversely, if she plans to donate more than $10,000, her probability of impact is higher. It is plausible that at least some individuals are in a position to have a non-negligible impact on existential and extinction risks via donations. Sam Bankman-Fried, the founder and CEO of FTX and a member of Giving What We Can, set out to make as much money as he could in order to give away everything he earned to charity. He is now the primary funder of the FTX Foundation's Future Fund, which works to improve humanity's odds of surviving and flourishing for thousands of years or longer. See FTX Future Fund (2022).

[87] Note that some might have a non-negligible impact on existential risks by doing direct work instead of donating money. For them, Longtermism may be true in the context of choosing which career to pursue or how to spend one's free time.

[88] Note that the definition of State Discounting given here considers very-small-probability states in cases where the prospects would otherwise have equal probability-discounted expected utility.

- $EU(X)_{pd} > EU(Y)_{pd}$ or
- $EU(X)_{pd} = EU(Y)_{pd}$ and $EU(X) \geq EU(Y)$,

where $EU(X)_{pd}$ and $EU(Y)_{pd}$ are obtained by conditionalizing on the supposition that no state of negligible probability occurs.

In order to use State Discounting to argue against Longtermism, we need a way of individuating states that guarantees that states in which Shivani makes a difference to existential risks are negligible. This can be done by individuating states in terms of whether some act makes a difference to existential catastrophes as follows: In one state, an existential catastrophe happens no matter what one does; in another state, one's actions make a difference to whether or not the catastrophe happens; and in the final state, an existential catastrophe does not happen no matter what one does. Let's call the second state a *difference-making state*. If the difference-making state is associated with a tiny probability, then one should ignore it. In effect, one would then ignore the possibility of making a difference to whether or not an existential catastrophe happens.

There are different ways of partitioning states, and thus, many versions of State Discounting. The focus of this section will be a version of State Discounting on which states are partitioned by comparing prospects to some status quo prospect, which corresponds to doing nothing.[89] Let's call this view *Baseline State Discounting*.

[89] On another version of State Discounting, prospects are always compared two at a time, and the possible states of the world are partitioned for every pairwise comparison separately. On a third version, states are partitioned by comparing all available options at once. See the appendix for a discussion of these two views.

Baseline State Discounting: States are partitioned by comparing every prospect to a status quo prospect (each separately).

How might Baseline State Discounting undermine Longtermism? Recall that by donating $10,000 to the Against Malaria Foundation, Shivani can save 2.5 lives in expectation—let's round that to 2. By donating the same money to AI safety, she can provide a 1 in 10 billion absolute reduction in the probability of an AI-driven catastrophe in the next 100 years. Baseline State Discounting compares the Against Malaria Foundation and AI safety to a status quo prospect (i.e., 'do nothing'), each separately.

Let's start by comparing AI safety to doing nothing. In order to capture the idea of the No Difference Argument, states must be individuated based on whether Shivani makes a difference to an AI-driven catastrophe as follows (see table 4): In state 1, an AI causes an existential catastrophe no matter what Shivani does. In state 2, an AI does not cause an existential catastrophe if she donates to AI safety, but it will cause an existential catastrophe if she does nothing. Lastly, in state 3, an AI does not cause an existential catastrophe no matter what she does. If Shivani's discounting threshold is higher than 1 in 10 billion, then she should ignore the possibility of state 2 obtaining. Consequently, the probability-discounted expected utility of AI safety equals (or is marginally better than) that of doing nothing. In effect, Shivani would then ignore the possibility of making a difference to whether or not an AI-driven existential catastrophe happens.

Table 4

AI Safety vs. Baseline

	State 1 $p \approx 0.001$	**State 2** $p = 10^{-10}$	**State 3** $p \approx 0.999$
AI safety	AI doom	No AI doom	No AI doom
Do nothing	AI doom	AI doom	No AI doom

Donating to the Against Malaria Foundation involves no uncertainty, as (we have assumed) it certainly saves two lives. As the Against Malaria Foundation certainly results in a better outcome than doing nothing, its probability-discounted expected utility is greater than that of doing nothing (see table 5).

Table 5

AMF vs. Baseline

	State 1
AMF	Two additional lives saved
Do nothing	No additional lives saved

So, the probability-discounted expected utility of AI safety equals that of doing nothing, while the probability-discounted expected utility of the Against Malaria Foundation is greater than that. Therefore, Shivani should donate to the Against Malaria Foundation, and Longtermism is false in her situation. Thus, Baseline State Discounting provides a prima facie case against Longtermism. If states are partitioned as in table 4, and the difference-making state (i.e., state 2) has neg-

ligible probability with all of the possible longtermist causes, then Baseline State Discounting undermines Longtermism.

4.2 State Individuation Problem

However, State Discounting faces an analogous problem to the Outcome Individuation Problem but with states instead of outcomes:[90]

> **State Individuation Problem:** If one individuates states with too much detail, all states have negligible probabilities. Is there a privileged way of individuating states that avoids this?

Earlier, states were individuated in terms of whether or not Shivani could make a difference to the occurrence of an AI-driven catastrophe. However, there are many ways in which such a catastrophe might happen. The occurrence of an AI-driven catastrophe was treated as a privileged basis for individuating states. We were interested in whether Shivani can affect the occurrence of an AI-driven catastrophe with no regard to how it might happen or how much utility is at stake. However, this seems arbitrary. Why should states be individuated in this way rather than some other way?

Apart from individuating states as finely as possible, it seems the only non-arbitrary way of individuating states is by the utilities of their outcomes. But in the case of existential risk from an AI, individuating states by the utilities of their outcomes would most likely result in many different states instead of just three,

[90] See §3 in Chapter 4 of this thesis.

as in the earlier example (table 4). This is so because these catastrophic scenarios would most likely differ in value. As a result, individuating states by the utilities of their outcomes does not guarantee that one will ignore the possibility of influencing the occurrence of an AI-driven catastrophe if and only if its probability is tiny. For example, one might ignore the possibility of making a difference even when the probability of doing so is high. This can happen if the different scenarios in which one makes a difference differ in value, and all these scenarios have tiny probabilities (even though their total probability is high). So, individuating states by the utilities of their outcomes does not capture the idea of the No Difference Argument.

Furthermore, individuating states by the utilities of their outcomes results in a violation of dominance.[91] Let $X \succ Y$ mean that X is strictly preferred to Y. Then, Baseline State Discounting violates the following dominance principle if states are individuated by utilities:[92]

> **Statewise Dominance:** If the outcome of prospect X is at least as preferred as the outcome of prospect Y in all states, then $X \succsim Y$. Furthermore, if in addition the outcome of X is strictly preferred to the outcome of Y in some possible state, then $X \succ Y$.

To see how Baseline State Discounting violates Statewise Dominance if states are individuated by utilities, consider the following prospects:

> **Space Settlement:** The Earth has a billion years left until the Sun expands and makes the Earth uninhabitable. However, a space settle-

[91] See §3 in Chapter 4 of this thesis.

[92] Savage (1951, p. 58) and Luce and Raiffa (1957, p. 287).

ment program might expand humanity's lifespan, with some cost ϵ. There are two alternative programs whose successes depend on some mutually exclusive events E_1, E_2, E_3 and E_4 as follows:[93]

Space program 1 Gives a 3% chance of humanity surviving for two billion years (if event E_1 happens) and a 2% chance of humanity surviving for five billion years (if event E_2 or E_3 happens). Otherwise, humanity will survive for a billion years on Earth (if event E_4 happens).

Space program 2 Gives a 4% chance of humanity surviving for two billion years (if event E_1 or E_2 happens) and a 1% chance of humanity surviving for five billion years (if event E_3 happens). Otherwise, humanity will survive for a billion years on Earth (if event E_4 happens).

Suppose the discounting threshold is (implausibly) just above 2%, and also suppose that the utility of humanity's lifespan equals its duration (in billions of years). Let's first compare *Space program 1* to the baseline, which again is 'do nothing.' Individuating states by the utilities of their outcomes results in the following states (see table 6): In state 1, humanity lives for two billion years if *Space program 1* is chosen, and otherwise, humanity lives for one billion years; in state 2, humanity lives for five billion years if *Space program 1* is chosen, and otherwise, humanity

[93]Note that usually in decision theory an event is defined as a set of states, which is not the case here. For example, state 2 in table 6 is composed of two mutually exclusive events. Here 'event' is used in its common meaning outside of decision theory. 'State', in turn, refers to a collection of maximally fine-grained possible states of the world. The reason for understanding states in this less fine-grained way is that maximally fine-grained states would all have probabilities below the discounting threshold.

lives for one billion years; and in state 3, both *Space program 1* and doing nothing result in humanity living for one billion years. The probability of state 2 is below the discounting threshold, so one should conditionalize on state 2 not happening. Then, the probability-discounted expected utility of *Space program 1* is $1.03-\epsilon$.[94]

TABLE 6
SPACE PROGRAM 1 VS. DOING NOTHING

	State 1	**State 2**	**State 3**
Event	E_1	E_2 or E_3	E_4
p	0.03	0.02	0.95
Space program 1	$2-\epsilon$	$5-\epsilon$	$1-\epsilon$
Do nothing	1	1	1

Next, let's compare *Space program 2* to the baseline. In this case, states are individuated similarly as before, except that the states have slightly different probabilities, as now E_2 results in state 1* (see table 7). As before, state 2* has negligible probability, so the possibility of state 2* is ignored. Consequently, the probability-discounted expected utility of *Space program 2* is $1.04-\epsilon$.[95]

[94] $0.03/0.98 \cdot (2-\epsilon) + 0.95/0.98 \cdot (1-\epsilon) \approx 1.03 - \epsilon$.

[95] $0.04/0.99 \cdot (2-\epsilon) + 0.95/0.99 \cdot (1-\epsilon) \approx 1.04 - \epsilon$.

TABLE 7

SPACE PROGRAM 2 VS. DOING NOTHING

	State 1*	**State 2***	**State 3***
Event	E_1 or E_2	E_3	E_4
p	0.04	0.01	0.95
Space program 2	$2-\epsilon$	$5-\epsilon$	$1-\epsilon$
Do nothing	1	1	1

The probability-discounted expected utility of *Space program 2* is greater than that of *Space program 1* ($1.04-\epsilon$ vs. $1.03-\epsilon$), so *Space program 2* is better than *Space program 1*. However, the only difference between these alternatives is that the former results in a lifespan of two billion years for humanity if event E_2 happens, while the latter results in a lifespan of five billion years in that case. So, when states are individuated in the usual way (see table 8), the two space programs give the same outcomes in states 1**, 3** and 4**, but *Space program 1* gives a better outcome in state 2**. This is a violation of Statewise Dominance. This Statewise Dominance violation happens because the partition of states is different for each option, leading to a situation where the states in which Space Program 1 beats Space Program 2 are ignored for Space Program 1 but not for Space Program 2. So, the most plausible way of individuating states (i.e., by utilities) leads to a violation of Statewise Dominance—which makes Baseline State Discounting less plausible as a theory of instrumental rationality.

Table 8

Space Program 1 vs. Space Program 2

	State 1**	**State 2****	**State 3****	**State 4****
Event	E_1	E_2	E_3	E_4
p	0.03	0.01	0.01	0.95
Space program 1	$2-\epsilon$	$5-\epsilon$	$5-\epsilon$	$1-\epsilon$
Space program 2	$2-\epsilon$	$2-\epsilon$	$5-\epsilon$	$1-\epsilon$

To summarize, the No Difference Argument states that the probability of making a difference to whether or not an existential catastrophe happens is so tiny that the possibility of making a difference should be ignored. Baseline State Discounting captures this idea naturally. And, it presents a prima facie challenge to Longtermism, as there is only a tiny probability that Shivani can make a difference to whether or not an existential catastrophe occurs. However, Baseline State Discounting faces the State Individuation Problem. As before, one might solve this by individuating states by the utilities of their outcomes. But if states are individuated by utilities, then it is not guaranteed that Baseline State Discounting ignores the possibility of making a difference if and only if the probability of doing so is tiny. So, Baseline State Discounting does not capture the idea behind the No Difference Argument if states are individuated by utilities. Furthermore, individuating states by the utilities of their outcomes also results in a violation of Statewise Dominance, which makes Baseline State Discounting less plausible as a theory of instrumental rationality. Nevertheless, one might still insist that there is some *other* privileged way of individuating states that avoids the violation of Statewise Dominance. Al-

ternatively, one might reject Baseline State Discounting and cash out the No Difference Argument in some other way. So, the No Difference Argument might still challenge Longtermism. However, the next section presents a more general response to the No Difference Argument.

5 Probability Discounting and Each-We Dilemmas

This section argues that Probability Discounting faces Each-We Dilemmas. These can be solved by accepting *Collective Difference-Making*. However, doing so also blocks the No Difference Argument. Some possible justifications for Collective Difference-Making will be discussed.

5.1 Collective Difference-Making

According to Parfit (1984, p. 91), a theory faces Each-We Dilemmas if "there might be cases where, if each does better in this theory's terms, we do worse, and vice versa."[96] To see how Probability Discounting faces Each-We Dilemmas, consider the following case (see table 9 depicting the decision-situation faced by a single agent):[97]

[96] Each-We Dilemmas differ from Prisoner's Dilemmas because in the former even impartial and altruistic agents who accept the same moral theory can end up choosing worse options by the lights of that theory when those choices are evaluated together.

[97] Versions of Probability Discounting that ignore very-small-probability outcomes face the following Each-We Dilemma:

> **Asteroid:** Multiple asteroids are heading toward the Earth, and for each of them, there is a tiny probability that it will hit unless it is stopped. However, the probability that at least one of them will hit the Earth is high if none of the asteroids are stopped.

Asteroid: An asteroid is heading toward the Earth and will almost certainly hit unless stopped. There are multiple asteroid defense systems, and (unrealistically) each has a tiny probability of hitting the asteroid and preventing a catastrophe. However, the probability that one of them succeeds is high if enough of them try. Attempting to stop the asteroid involves some small cost ϵ.[98]

TABLE 9
ASTEROID

	State 1	**State 2**	**State 3**
Attempt	Collision $-\epsilon$	No collision $-\epsilon$	No collision $-\epsilon$
Do nothing	Collision	Collision	No collision

In this case, the probability of state 2 happening is below the discounting threshold, so the possibility of state 2 should be ignored. However, then doing nothing is better than attempting to stop the asteroid because it gives a better outcome in states 1 and 3. So, Probability Discounting recommends against attempting to stop the asteroid because the probability of making a difference is below the discounting threshold, and trying to stop the asteroid incurs a small cost. Consequently,

There are multiple asteroid defense systems, and each can only target one asteroid. Attempting to stop an asteroid involves some small cost ϵ.

As the agents can only attempt to stop one asteroid, and the probability of this asteroid hitting the Earth is tiny, versions of Probability Discounting that ignore very-small-probability outcomes recommend against attempting to stop the asteroid. Consequently, an asteroid will almost certainly hit the Earth—which could have been prevented had enough agents attempted to do so.

[98] This cost is so small that the asteroid hitting the Earth is worse than a cost of ϵ to all the relevant people.

the asteroid will almost certainly hit the Earth—which could have been prevented almost certainly had enough agents attempted to do so.

Many have appealed to expected benefits in order to solve collective action problems.[99] For example, it is sometimes argued that one cannot justify voting by merely appealing to the consequences of one's act because there is only a minuscule probability that one vote makes a difference.[100] The expected benefits of voting can nonetheless be great because if one's vote makes a difference, it will impact millions of people.[101] However, if one ought to discount very small probabilities, then appealing to expected benefits cannot solve collective action problems in which it is almost certain of each person that they make no difference. If one vote is extremely unlikely to make a difference, and one should ignore tiny probabilities, then the expected benefits of voting are negligible.

If Probability Discounting is to avoid Each-We Dilemmas, agents must somehow take into account the choices of other people. They must accept

> **Collective Difference-Making:** One ought to take into account the choices of other people and consider whether the collective has a non-negligible probability of making a difference.[102]

[99] See Parfit (1984, pp. 73–75), Parfit (1988) and Kagan (2011). For a criticism of this solution, see Nefsky (2011).

[100] Parfit (1984, p. 73).

[101] Parfit (1984, pp. 73–75).

[102] Note that, on Collective Difference-Making, it matters whether the small probabilities are independent for the different agents. Suppose that a googolplex agents face *Pascal's Mugging*. The probability that at least one of them gets a thousand quadrillion happy days in the Seventh Dimension is still small even if they all pay the mugger because the probability of obtaining the great outcome is not independent for the different agents: Either the mugger has magical powers, or he does not. So, Collective Difference-Making recommends that the agents ignore the small prob-

There are several different ways to interpret Collective Difference-Making. On one interpretation, agents should choose a small enough discounting threshold so that Each-We Dilemmas do not arise to begin with (and adjust the threshold lower if they anyway do arise). This interpretation is 'collective' because agents ought to take into account the choices of others when choosing the discounting threshold. On another interpretation, all the choices faced by different agents should be evaluated collectively, and if the total probability of some event or outcome is above the discounting threshold, then no one should discount. This latter view is similar to what Monton (2019) and Smith (2016) say in diachronic cases, where we consider different choices made by the same agent over time. They argue that relevantly similar choices faced by one individual must be evaluated collectively, and one should not discount if the total probability of some event or outcome is above the discounting threshold.[103] So, on this interpretation, Collective Difference-Making implies that one should reason as if one was facing sequentially all the choices faced by different agents.

The probability that Shivani and all the other agents together can make a difference to existential risks seems non-negligible. For example, if we spend $1 billion on AI safety, it is plausible that we can provide at least a 1 in 100,000 absolute reduction in the probability of an AI-driven catastrophe.[104] This estimate is

ability. However, if the probabilities were independent, then Collective Difference-Making would recommend against discounting, provided that the total probability of at least one person obtaining the great outcome is sufficiently high.

[103]The approach advocated by Monton (2019) and Smith (2016) assumes that there are no intrapersonal Each-We Dilemmas because rational agents have the power to commit to making some choices in the future.

[104]$0.001 \cdot 0.01 = 0.00001$. Greaves and MacAskill (2021, pp. 14–15) estimate that there is at

conservative. As mentioned earlier, the median expert estimate for an AI-driven catastrophe at any point in time is 5%, while the calculation assumed a 0.1% risk in the next 100 years. Also, $1 billion spent on AI safety might decrease the probability of an AI-driven catastrophe by more than 1%. So, if one ought to accept Collective Difference-Making, then—plausibly—Probability Discounting does not undermine Longtermism. Shivani should not ignore the possibility of making a difference because she and the other agents have a non-negligible chance of preventing an existential catastrophe.

The details of Collective Difference-Making do not matter for the purposes of this chapter, so I will only briefly mention some possible justifications for and problems with Collective Difference-Making. The details do not matter because, if Collective Difference-Making is plausible, then Probability Discounting does not undermine Longtermism, as Shivani and all the other agents have a non-negligible chance of making a difference. But if Collective Difference-Making is implausible, then Probability Discounting faces Each-We Dilemmas, making it implausible as well. Either way, Probability Discounting does not undermine Longtermism.

5.2 Justifications for Collective Difference-Making

How can Collective Difference-Making be justified? In response to collective action problems, some argue that we have reasons for action coming from the participatory nature of one's act. On these views, the reason for action is that by doing

---

least a 0.1% chance of an AI-driven catastrophe in the next 100 years, and that $1 billion of spending would decrease this probability by at least 1%. See Greaves and MacAskill (2021, p. 15).

so, one could be part of a group of people who together could make a difference.[105] For example, some argue that we have collective reasons for action.[106] On this view, groups, like individuals, have reasons to make outcomes better, benefit other people, avoid harming other people and benefit themselves. We have reasons as a group to carry out some action because we would together be making things better.[107] Furthermore, there might be things that some groups ought to do, even if they have never coordinated in the past nor will ever coordinate in the future.[108] This view can solve collective action problems if the reasons of groups bear on the reasons of individuals. In that case, the agents in *Asteroid* may have a collective reason to attempt to stop the asteroid and an individual reason to do their part. Similarly, Shivani and the other agents may have a collective reason to prevent an existential catastrophe (if they have a non-negligible probability of having an impact) and an individual reason to do their part.

Others, in turn, argue that one's act can be part of causing some outcome without making a difference.[109] This can happen when the outcome not happening would be at least partly a result of there not having been enough similar acts.[110]

[105] Nefsky (2017, p. 2756). For a criticism of these views, see Nefsky (2015).

[106] See for example Dietz (2016). Consider also this view from Parfit (1984, p. 70):

> "Even if an act harms no one, this act may be wrong because it is one of a set of acts that together harm other people. Similarly, even if some act benefits no one, it can be what someone ought to do, because it is one of a set of acts that together benefit other people."

See also Parfit (1984, pp. 31–31).

[107] Dietz (2016, p. 960).

[108] Dietz (2016, p. 957).

[109] Nefsky (2017).

[110] Nefsky (2017, p. 2753).

The idea is that one has a reason to act in a certain way because one could be making a causal contribution toward bringing about some outcome (even though one would not make a difference in expectation). The conditions for making a causal contribution without making a difference are that it is up in the air whether or not the outcome in question will occur; that part of what could determine whether it occurs is whether enough people act in the relevant way going forward; and that it is up in the air whether or not enough people will act in that way going forward.[111]

On this view, the agents in *Asteroid* should attempt to stop the asteroid because doing so might be making a causal contribution toward stopping it, even though in expectation they would not be making a difference.[112] Similarly, Shivani should mitigate existential risks because she might thereby be making a causal contribution toward preventing an existential catastrophe (even though in expectation she would not be making a difference). In both Shivani's case and *Asteroid*, it is up in the air whether or not the existential catastrophe will occur; part of what could determine whether it occurs is whether enough people mitigate existential risks; and

[111]Nefsky (2017, p. 2758). On reasons to vote, Nefsky (2017, pp. 2756–2757) writes: "But, contrary to the expected utility approach, the main reason to vote does not come from this minuscule chance of making a difference—from extremely remote chance of the election turning on your vote. Rather, it comes from the fact that your vote could help to elect Mr. Powers [the better candidate] regardless of whether the election turns on it (which it almost certainly will not). Your vote could help because, at the time at which you vote, more votes for Mr. Powers are needed to prevent the disastrous outcome, and there is no guarantee that there will be enough such votes. So, by voting, you are making a causal contribution toward preventing the bad outcome, when there is a real risk that this outcome will not be prevented due to a lack of exactly that sort of contribution. Making such a contribution in those circumstances makes progress toward preventing the bad outcome, even if what happens will not turn on your having done so."

[112]It is unclear whether Nefsky would apply this theory to cases such as *Asteroid*. On cases in which each person has a tiny chance of triggering some result regardless of what others do (such as *Asteroid*), Nefsky (2011, p. 367n11) writes: "It seems to me, though, that such a case would not be a collective harm case."

it is up in the air whether or not enough people will mitigate existential risks.

Alternatively, one can also justify Collective Difference-Making with, for example, rule-consequentialism. Rule-consequentialism states that agents should decide what to do by applying rules whose acceptance will produce the best consequences. Rule-consequentialism would (presumably) advise that the agents attempt to stop the asteroid because doing so conforms to a rule whose acceptance produces the best consequences. Similarly, Rule-consequentialism would (presumably) advise Shivani to mitigate existential risks because 'mitigate existential risks' is a rule whose acceptance produces the best consequences in the long run.

Another way of justifying something close to Collective Difference-Making comes from Evidential Decision Theory. According to Evidential Decision Theory, the best act is the one that gives the best expectations for the outcomes, conditional on one choosing it. Evidential Decision Theory is often contrasted with Causal Decision Theory. According to Causal Decision Theory, agents ought to maximize the best expected causal consequences. On this view, causality plays an important role in instrumental rationality: Only those consequences that have a causal link with one's act count. In contrast, evidentialists do not require a belief in a causal link between one's act and the consequences.[113]

Evidential Decision Theory favors something akin to Collective Difference-Making because it implies that an agent ought to reason as if they were choosing on behalf of all relevantly similar agents.[114] Evidential Decision Theory recom-

[113]See Nozick (1969).

[114]MacAskill et al. (2021) argue that an altruistic and morally motivated agent who is uncertain between Evidential and Causal Decision Theory should generally act following the former, even if

mends not discounting the probability of making a difference in *Asteroid* if doing so provides sufficient evidence of others also not discounting. And it may, if others are similar to the agent in relevant ways. The idea is that under certain conditions, conditional on some agent acting, it is likely that enough people act to deflect the asteroid, so the probability of stopping the asteroid is non-negligible. Similarly, Evidential Decision Theory recommends Shivani to mitigate existential risks if doing so provides sufficient evidence of others mitigating these risks as well. And it may, if others are similar to Shivani in relevant ways. However, Evidential Decision Theory does not solve Each-We Dilemmas in cases where one's actions do not provide suitably strong evidence of how other agents will act. If ignoring the small chance of stopping the asteroid does not provide sufficiently strong evidence of other agents doing so as well, then Evidential Decision Theory recommends doing nothing instead of attempting to stop the asteroid.

5.3 Problems with Collective Difference-Making

I have discussed some ways of justifying Collective Difference-Making. However, Collective Difference-Making faces some problems as well. First, to even start estimating the number of very-small-probability choices all agents make, one needs to know who counts as an agent. Do small children count? What about animals? Or possible intelligent aliens or AI? Evidential Decision Theory can solve this: All agents who are relevantly similar to oneself count (in proportion to how similar

she has a higher credence in the latter. They argue that the existence of correlated decision-makers will affect the stakes for Evidential Decision Theory but not for Causal Decision Theory and that it is rational to hedge if one faces decision-theoretic uncertainty.

they are to oneself) because then one's actions are evidence of how they will act. Another possible solution is that those on a collective endeavor with oneself count.[115] On this view, for example causally disconnected intelligent aliens do not count.

Another problem for Collective Difference-Making is the violation of Separability. Let X be a prospect that concerns what is going on in the part of the world we might make any difference to, and let Y be a prospect that concerns what happens somewhere far away, such as a distant galaxy. Also, let $X \oplus Y$ be the combined prospect of the near prospect X and the far prospect Y. Then, Separability states the following:[116]

Separability:

i For all near prospects X and Y, and any far prospect Z, $X \succ Y$ if and only if $X \oplus Z \succ Y \oplus Z$.

ii For all far prospects X and Y, and any near prospect Z, $X \succ Y$ if and only if $Z \oplus X \succ Z \oplus Y$.

Collective Difference-Making violates Separability because what one ought to do

[115] For example, Kutz (2000, p. 89) writes:"Jointly acting groups consist of individuals who intend to contribute to a collective end."

[116] Russell (2021, p. 15). Contrast Separability with *Background Independence*:

> **Background Independence:** For all prospects X and Y, and any far outcome z, $X \succ Y$ if and only if $X \oplus z \succ Y \oplus z$ (Russell, 2021, p. 18).

Background Independence is related to the Egyptology objection to the Average View in population ethics. See McMahan (1981, p. 115) and Parfit (1984, p. 420). The background outcome z does not add any uncertainty, so it will not interact with X and Y in different ways in different states. Thus, unlike Separability, Background Independence is consistent with (first-order) Stochastic Dominance. See Russell (2021, p. 18n13).

depends on what choices other distant agents face.[117] For example, Collective Difference-Making implies that the agents in *Asteroid* should not attempt to stop the asteroid if no other agents were facing the same choice; but given that enough others are also facing this choice, they should attempt to stop the asteroid. So, what agents should do depends on what choices others face.

Furthermore, there is a trade-off between maintaining Separability and avoiding Each-We Dilemmas. The fewer agents' choices one considers in one's decision-making, the more Each-We Dilemmas occur, and vice versa. For example, if one only takes into account the choices of other humans living on Earth right now, then one might end up in an Each-We Dilemma situation with future generations. Alternatively, if one only takes into account the choices of those who are on a collective endeavor with oneself, then one might end up in an Each-We Dilemma with those not on this collective endeavor.

Suppose that possible intelligent aliens would not be on a collective endeavor with us. We might then end up in the following kind of Each-We Dilemma with them:

[117] Wilkinson (2022, §6) shows that denying Probability Fanaticism leads to violations of Separability (or first-order Stochastic Dominance), even in cases where the choices of different individuals are probabilistically independent. See also Beckstead and Thomas (2020). However, Russell (2021) shows that (first-order) Stochastic Dominance and Separability are inconsistent (assuming *Positive Compensation*: One can always compensate for making things worse nearby by making things sufficiently better far away, and vice versa). Also see Goodsell (2021). Russell (2021, p. 14) writes: "[W]hat is better than what really does depend in strange ways on what is going on in distant space and time… it matters whether you think there is another St. Petersburg population lottery going on in a distant galaxy. This is bizarre—but Stochastic Dominance tells us that it is true." Stochastic Dominance is consistent with the separability of simple prospects, that is, prospects that have finitely many possible outcomes (Russell, 2021, p. 14). However, as Russell points out, whatever justifies the separability of simple prospects will probably also justify full Separability.

> **Asteroid 2:** Asteroids are heading toward different planets (one for each planet), and they will almost certainly hit unless they are stopped. There is one asteroid defense system on every planet, and (unrealistically) each has a tiny probability of hitting the asteroid and preventing a catastrophe. However, the probability that at least one of them hits an asteroid is high if enough of them try. Again, trying to stop the asteroid incurs some small cost ϵ.

It would be better if everyone attempted to stop the asteroid heading toward their planet. Probably at least one of the planets would survive. However, if one should ignore what happens on faraway planets, then one should ignore the possibility of successfully stopping the asteroid heading toward one's planet. Consequently, no planets survive. So, if one ignores the choices of some group of agents, then one might end up in an Each-We Dilemma with this group. On the other hand, if one cares about the difference all agents can make, then violations of Separability will be more common. Also, if there is a large number of agents, one might not discount tiny probabilities very often, if ever.[118]

Another problem for Collective Difference-Making is cluelessness: It seems impossible to evaluate how many very-small-probability choices other agents face. So, Collective Difference-Making needs some way of handling situations where

[118]Wilkinson (2022) writes on the long-run argument for maximizing expected value: "How well the world as a whole goes is not determined by just a few decisions by a single agent, but instead by countless different agents making separate small-scale decisions. In this setting, having all of those agents maximize expected value seems to be quite a good policy, even when doing so produces fanatical verdicts. Repeated enough times, even fanatical choices will pay off eventually." However, note that this will only happen if the probabilities are sufficiently independent for the different agents.

one is clueless about what choices others face. However, many other theories also face the problem of cluelessness, so this problem need not disadvantage Collective Difference-Making over the alternatives.[119]

Finally, another task for the proponents of Collective Difference-Making is to spell out the details of when agents should refrain from discounting small probabilities. Does it only have to be the case that sufficiently many agents face sufficiently many very-small-probability choices, or do enough of those agents also need to refrain from discounting? Do their choices need to be relevantly similar (such as attempts to stop a particular asteroid heading toward the Earth), or is it enough that they involve similarly small probabilities but in very different contexts? What happens if different agents assign different probabilities to the same events?

I will not attempt to solve these problems in this chapter. Instead, as mentioned earlier, my argument is that if Collective Difference-Making is implausible, then Probability Discounting is also implausible because it leads to Each-We Dilemmas. On the other hand, if Collective Difference-Making is plausible, then Probability Discounting does not undermine Longtermism because Shivani and all the other agents together have a non-negligible probability of making a difference. Either way, Probability Discounting and the No Difference Argument do not undermine Longtermism. However, discounting small probabilities might still be relevant to

[119]However, this problem may be more serious for Collective Difference-Making. For example, an agent might think there is a tiny probability that countless agents face very-small-probability choices. Should the agent discount that probability down to zero and ignore this possibility? If the agent ignores this possibility, then the number of individuals is small, and they are right to ignore it. On the other hand, if the agent does not ignore this possibility, then the number of individuals is large, and the agent is right not to ignore it.

what longtermists should focus on, as there might be a class of existential risks that we cannot make a difference to, even together.

6 Conclusion

I have discussed three arguments against Longtermism from discounting small probabilities. First, I discussed the Low Risks Argument: The probabilities of existential catastrophes are so low that we ought to ignore them. However, even in the next century, the net existential risk and some specific existential risks are above any reasonable discounting thresholds. Naive Discounting faces the Outcome Individuation Problem, so it is unclear what it says. However, an acceptable solution to this problem should not imply that one ought to ignore a net existential risk of 1/6 in the next century. Tail Discounting is more plausible than Naive Discounting. However, as long as there are non-negligible probabilities of better and worse outcomes than a near-term extinction, Tail Discounting will not ignore near-term extinction events even if their associated probabilities are negligible.

The second argument against Longtermism I discussed is the Small Future Argument: Once we ignore very-small-probability scenarios, such as space settlement and digital minds, the expected number of lives in the far future is too small for Longtermism to be true. However, this does not seem true. For example, AI safety leads to Longtermism if humanity's expected lifespan is at least 265 years. Therefore, the Small Future Argument does not undermine Longtermism.

Finally, I discussed the No Difference Argument: The probability that an agent

can make a difference to whether or not an existential catastrophe occurs is so small that it should be discounted down to zero. Baseline State Discounting captures this idea naturally. It may also challenge Longtermism, as there is only a tiny probability that Shivani can make a difference to whether or not an existential catastrophe occurs. However, if states are individuated in the most plausible way (i.e., by utilities), Baseline State Discounting violates Statewise Dominance, which makes it less plausible as a theory of instrumental rationality.

Lastly, I argued that Probability Discounting faces Each-We Dilemmas. If Probability Discounting is to avoid Each-We Dilemmas, it needs Collective Difference-Making: Agents must take into account the choices of other people and consider whether the collective can make a difference. However, if we accept Collective Difference-Making, then Probability Discounting does not undermine Longtermism because Shivani and all the other agents together have a non-negligible probability of making a difference.

All in all, I have discussed three ways in which discounting small probabilities might undermine Longtermism. I have argued that these arguments do not succeed. Discounting small probabilities gives no reason to reject Longtermism.

Appendix

A State Discounting and Acyclicity

Earlier I discussed a version of State Discounting on which states are partitioned by comparing prospects to a status quo prospect. But there are different views about how states should be partitioned. On another version of State Discounting, prospects are always compared two at a time, and the possible states of the world are partitioned for every pairwise comparison separately. Alternatively, one could compare all available options at once and partition the states for every choice set separately. Let's call these views *Pairwise State Discounting* and *Set-Dependent State Discounting*, respectively.

> **Pairwise State Discounting:** States are partitioned by comparing two prospects at a time.

> **Set-Dependent State Discounting:** States are partitioned by comparing all prospects at once.

The argument against Longtermism from Pairwise and Set-Dependent State Discounting is similar to that from Baseline State Discounting. Recall that by donating $10,000 to the Against Malaria Foundation, Shivani can save two lives in expectation. By donating the same money to AI safety, she can provide a 1 in 10 billion absolute reduction in the probability of an AI-driven catastrophe in the next 100 years. Instead of partitioning the states by comparing AI safety and the Against

Malaria Foundation to a status quo prospect, Pairwise and Set-Dependent State Discounting partition the states by comparing these two options. Consequently, it ignores tiny differences between these prospects. As the whole choice set only includes two alternatives, both views treat the case similarly.

As before, in order to capture the idea of the No Difference Argument, states must be individuated based on whether or not Shivani makes a difference to an AI-driven catastrophe as follows: In state 1, an AI causes an existential catastrophe no matter what Shivani does. In state 2, an AI does not cause an existential catastrophe if she donates to AI safety, but it will cause an existential catastrophe if she donates to the Against Malaria Foundation. Lastly, in state 3, an AI does not cause an existential catastrophe no matter what she does. Donating to the Against Malaria Foundation saves two lives in all states. Shivani's choice situation is shown in table 10.

Table 10
AI Safety vs. AMF

	State 1 $p \approx 0.001$	**State 2** $p = 10^{-10}$	**State 3** $p \approx 0.999$
AI safety	AI doom	No AI doom	No AI doom
AMF	AI doom + 2 lives	AI doom + 2 lives	No AI doom + 2 lives

As before, if Shivani's discounting threshold is higher than 1 in 10 billion, she ought to ignore the possibility of state 2 obtaining. Consequently, donating to the Against Malaria Foundation is better because it gives a better outcome in states

1 and 3. So, like Baseline State Discounting, Pairwise and Set-Dependent State Discounting challenge Longtermism if they partition states as in table 10.

However, partitioning states as in table 10 leads to a violation of the following principle:

> **Acyclicity:** If $X_1 \succ X_2 \succ \cdots \succ X_n$, then it is not the case that $X_n \succ X_1$.

According to Pairwise and Set-Dependent State Discounting, states might be partitioned differently depending on what other options are available, and this can generate cycles. The former violates Acyclicity within choice sets, while the latter violates Acyclicity across choice sets when two options are compared at a time.

Suppose that Shivani gives a 5% probability for an AI-driven catastrophe and that (implausibly) her discounting threshold is 2%. Next, to see why Pairwise and Set-Dependent State Discounting violate Acyclicity, consider the following options:

> **Acyclicity Violation:**
>
> *Against Malaria Foundation* Saves two lives and gives a 5% probability of an AI-driven catastrophe.
>
> *Pure AI safety* Decreases the probability of an AI-driven catastrophe to 3%.
>
> *Mixed AI safety* Decreases the probability of an AI-driven catastrophe to 4% and, in addition, saves one life in the near-term future.

First, let's compare Pure AI safety to donating to the Against Malaria Foundation (see table 11). States are partitioned in the same way as in table 10. But, in this case, the probability of state 2 is not below the discounting threshold, so there is a non-negligible chance that Shivani can influence whether an AI-driven catastrophe occurs. Consequently, when state 2 is not ignored, donating to Pure AI safety is better than donating to the Against Malaria Foundation.

Table 11
Pure AI Safety Is Better than AMF

	State 1	**State 2**	**State 3**
p	0.03	0.02	0.95
Pure AI safety	Doom	No doom	No doom
AMF	Doom + 2 lives	Doom + 2 lives	No doom + 2 lives

Next, let's compare the Against Malaria Foundation to Mixed AI safety (see table 12). Again, states are partitioned in the same way as in table 10. This time, the probability of state 2* is below the discounting threshold, so Shivani should ignore the possibility of influencing an AI-driven catastrophe. Moreover, when state 2* is ignored, the Against Malaria Foundation is better than Mixed AI safety because it gives a better outcome in states 1* and 3* (two lives saved instead of one). So, now we have that Pure AI safety is better than the Against Malaria Foundation, which is better than Mixed AI safety. It follows by Acyclicity that Mixed AI safety is not better than Pure AI safety.

Table 12

AMF Is Better than Mixed AI Safety

	State 1*	**State 2***	**State 3***
p	0.04	0.01	0.95
Mixed AI safety	Doom + 1 life	No doom + 1 life	No doom + 1 life
AMF	Doom + 2 lives	Doom + 2 lives	No doom + 2 lives

However, when we compare Pure AI safety and Mixed AI safety pair-wise, we find the opposite: Mixed AI safety is better than Pure AI safety (see table 13). As before, states are partitioned the same way as in table 10. In this case, the probability of state 2** is below the discounting threshold, so one should consider Pure AI safety and Mixed AI safety equally effective at reducing the probability of an AI-driven catastrophe. Consequently, Mixed AI safety is better than Pure AI safety because it gives a better outcome in states 1** and 3**. So, now we have that Pure AI safety is better than the Against Malaria Foundation, which is better than Mixed AI safety, which is better than Pure AI safety. This is a violation of Acyclicity.[120]

Table 13

Mixed AI Safety Is Better than Pure AI Safety

	State 1**	**State 2****	**State 3****
p	0.03	0.01	0.96
Pure AI safety	Doom	No doom	No doom
Mixed AI safety	Doom + 1 life	Doom + 1 life	No doom + 1 life

[120] Pairwise and Set-Dependent State Discounting also violate (first-order) Stochastic Dominance. See §4 in Chapter 4 of this thesis.

Pairwise State Discounting violates Acyclicity even within a single choice set; when all three options are available, there is no clear winner. Therefore, it is unclear what Pairwise State Discounting implies and what one ought to choose. Set-Dependent State Discounting, in turn, violates Pair-Wise Acyclicity, that is, it violates Acyclicity when we only compare two options at a time (as in tables 11, 12 and 13). However, if we compare all three options at once, then Set-Dependent State Discounting avoids this violation of Acyclicity, at least if states are partitioned as in table 14. In this case, states 2*** and 3*** have probabilities below the discounting threshold, so Shivani should ignore the possibilities of states 2*** and 3***. Once she does that, the Against Malaria Foundation comes out as the best option because it gives the best outcome in states 1*** and 4***. Within this choice-set, the Against Malaria Foundation is better than Mixed AI safety, which is better than Pure AI safety, which is worse than the Against Malaria Foundation. So, there is no violation of Acyclicity.

Table 14
No Violation of Acyclicity

	State 1* **	**State 2*** **	**State 3*** **	**State 4*** **
p	0.03	0.01	0.01	0.95
Pure AI safety	Doom	No doom	No doom	No doom
Mixed AI safety	Doom + 1 life	Doom + 1 life	No doom + 1 life	No doom + 1 life
AMF	Doom + 2 lives	Doom + 2 lives	Doom + 2 lives	No doom + 2 lives

Note that this case also shows that Set-Dependent State Discounting violates the following intuitively plausible principles:[121]

[121] Sen (1977, pp. 63–66). Contraction Consistency implies Acyclicity. See Sen (1977, p. 67).

Contraction Consistency: For all prospects X and Y, if it is permissible to choose X from the set $\{X, \ldots, Y\}$, then it is permissible to choose X from any subset of the set $\{X, \ldots, Y\}$.

Strong Expansion Consistency: For all prospects X, Y and Z, if it is permissible to choose X from the set $\{X, \ldots, Y\}$, then if it is permissible to choose Y from the set $\{X, \ldots, Y, \ldots, Z\}$, it is permissible to choose X from the set $\{X, \ldots, Y, \ldots, Z\}$.

Set-Dependent State Discounting violates Contraction Consistency because it is permissible to choose the Against Malaria Foundation when both AI safety options are available (table 14), but it is not permissible to choose it when only Pure AI safety is available (table 11). On the other hand, Set-Dependent State Discounting violates Strong Expansion Consistency because it is permissible to choose Pure AI safety when the Against Malaria Foundation is the only alternative (table 11). However, it is not permissible to choose Pure AI safety when all three options are available, but it is anyhow permissible to choose the Against Malaria Foundation (table 14).

Set-Dependent State Discounting is choice-set dependent. It implies that what Shivani ought to do depends on what other options are available to her, even if she will not choose them. Consequently, whether or not Longtermism is true in Shivani's situation may also be choice-set dependent. Longtermism may be true if Shivani only considers donating to AI safety and the Against Malaria Foundation. However, if she also considers donating to, for example, asteroid detection,

then Longtermism may no longer be true in her situation. In that case, there might be more states of nature because of a greater number of available options. Consequently, the difference-making state(s) might now have probabilities below the discounting threshold. This seems implausible. Having more longtermist options should not make Longtermism harder to achieve. However, one implication of Set-Dependent State Discounting is that adding more options can decrease the probabilities of the difference-making state(s) sufficiently to render Longtermism false.

To summarize, like Baseline State Discounting, the two alternative versions of State Discounting present a challenge to Longtermism. However, they give cyclic recommendations, which makes them less plausible as theories of instrumental rationality.

References

Adams, F. C. (2008), Long-term astrophysical processes, *in* N. Bostrom and M. Cirkovic, eds, 'Global Catastrophic Risks', Oxford University Press, Oxford.

Balfour, D. (2021), 'Pascal's Mugger strikes again', *Utilitas* **33**(1), 118–124.

Beckstead, N. (2013), On the overwhelming importance of shaping the far future, PhD thesis, Rutgers, the State University of New Jersey.

Beckstead, N. and Thomas, T. (2020), 'A paradox for tiny probabilities and enormous values'. Global Priorities Institute Working Paper No.10.

URL: *https://globalprioritiesinstitute.org/nick-beckstead-and-teruji-thomas-a-paradox-for-tiny-probabilities-and-enormous-values/*

Bostrom, N. (2003), 'Astronomical waste: The opportunity cost of delayed technological development', *Utilitas* **15**(3), 308–314.

Bostrom, N. (2009), 'Pascal's Mugging', *Analysis* **69**(3), 443–445.

Bostrom, N. (2013), 'Existential risk prevention as global priority', *Global Policy* **4**(1), 15–31.

Bricker, D. and Ibbitson, J. (2019), *Empty Planet: The Shock of Global Population Decline*, Crown, New York.

Dietz, A. (2016), 'What we together ought to do', *Ethics* **126**(4), 955–982.

Drake, N. (2020), 'Why NASA plans to slam a spacecraft into an asteroid'.

URL: *https://www.nationalgeographic.com/science/article/giant-asteroid-nasa-dart-deflection*

Francis, T. and Kosonen, P. (n.d.), 'Ignore outlandish possibilities'. Unpublished manuscript.

FTX Future Fund (2022), 'Principles'.

URL: *https://ftxfuturefund.org/principles/*

GiveWell (2020), 'GiveWell's cost-effectiveness analyses'.

URL: *https://www.givewell.org/how-we-work/our-criteria/cost-effectiveness/cost-effectiveness-models*

Goodsell, Z. (2021), 'A St Petersburg Paradox for risky welfare aggregation', *Analysis* **81**(3), 420–426.

GPT-3 (n.d.), 'Is artificial general intelligence an existential threat?'.
URL: *https://beta.openai.com/playground*

Grace, K., Salvatier, J., Dafoe, A., Zhang, B. and Evans, O. (2018), 'When will AI exceed human performance? Evidence from AI experts', *Journal of Artificial Intelligence Research* **62**, 729–754.

Greaves, H. and MacAskill, W. (2021), 'The case for strong longtermism'. Global Priorities Institute Working Paper 5–2021.
URL: *https://globalprioritiesinstitute.org/hilary-greaves-william-macaskill-the-case-for-strong-longtermism-2/*

Gustafsson, J. and Kosonen, P. (n.d.), 'Prudential Longtermism'. Unpublished manuscript.

Hájek, A. (2014), 'Unexpected expectations', *Mind* **123**(490), 533–567.

Hey, J. D., Neugebauer, T. M. and Pasca, C. M. (2010), Georges-Louis Leclerc de Buffon's 'Essays on moral arithmetic', *in* A. Sadrieh and A. Ockenfels, eds, 'The Selten School of Behavioral Economics: A Collection of Essays in Honor of Reinhard Selten', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 245–282.

Isaacs, Y. (2016), 'Probabilities cannot be rationally neglected', *Mind* **125**(499), 759–762.

Kagan, S. (2011), 'Do I make a difference?', *Philosophy and Public Affairs* **39**(2), 105–141.

Kreps, D. M. (1988), *Notes on the Theory of Choice*, Westview Press, Boulder.

Kutz, C. (2000), *Complicity: Ethics and Law for a Collective Age*, Cambridge University Press, Cambridge.

Lubin, P. and Cohen, A. N. (n.d.), 'Don't forget to look up'. Unpublished manuscript.
URL: *https://arxiv.org/pdf/2201.10663.pdf*

Luce, R. D. and Raiffa, H. (1957), *Games and Decisions: Introduction and Critical Survey*, Wiley, New York.

Lundgren, B. and Stefánsson, H. O. (2020), 'Against the De Minimis principle', *Risk Analysis* **40**(5), 908–914.

MacAskill, W. (2019), 'Longtermism', Effective Altruism Forum.
URL: *https://forum.effectivealtruism.org/posts/qZyshHCNkjs3TvSem/longtermism*

MacAskill, W., Vallinder, A., Shulman, C., Österheld, C. and Treutlein, J. (2021), 'The evidentialist's wager', *Journal of Philosophy* **118**(6), 320–342.

McMahan, J. (1981), 'Problems of population theory', *Ethics* **92**(1), 96–127.

Millett, P. and Snyder-Beattie, A. (2017), 'Existential risk and cost-effective biosecurity', *Health Security* **15**(4), 373–383.

Monton, B. (2019), 'How to avoid maximizing expected utility', *Philosophers' Imprint* **19**(18), 1–24.

Musk, E. (n.d.), 'Mars & beyond: The road to making humanity multiplanetary'.
URL: *http://www.https://www.spacex.com/human-spaceflight/mars/*

Nefsky, J. (2011), 'Consequentialism and the problem of collective harm: A reply to Kagan', *Philosophy and Public Affairs* **39**(4), 364–395.

Nefsky, J. (2015), Fairness, participation, and the real problem of collective harm, *in* M. Timmons, ed., 'Oxford Studies in Normative Ethics', Vol. 5, Oxford University Press, Oxford, pp. 245–271.

Nefsky, J. (2017), 'How you can help, without making a difference', *Philosophical Studies* **174**(11), 2743–2767.

Newberry, T. (2021), 'How cost-effective are efforts to detect near-Earth-objects?'. Global Priorities Institute Technical Report T1–2021.
URL: *https://globalprioritiesinstitute.org/how-cost-effective-are-efforts-to-detect-near-earth-objects-toby-newberry-future-of-humanity-institute-university-of-oxford/*

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

Nozick, R. (1969), Newcomb's problem and two principles of choice, *in* N. Rescher, ed., 'Essays in Honor of Carl G. Hempel: A Tribute on the Occasion of His Sixty-Fifth Birthday', Reidel, Dordrecht, pp. 114–146.

Ord, T. (2020), *The Precipice: Existential Risk and the Future of Humanity*, Bloomsbury, London.

Parfit, D. (1984), *Reasons and Persons*, Clarendon Press, Oxford.

Parfit, D. (1988), 'What we together do'. Unpublished manuscript.

Pulskamp, R. J. (n.d.), 'Correspondence of Nicolas Bernoulli concerning the St. Petersburg Game'. Unpublished manuscript. Accessed through: https://web.archive.org/.

URL: *http://cerebro.xu.edu/math/Sources/NBernoulli/correspondence_petersburg_ game.pdf*

Rees, M. (2003), *Our Final Hour: A Scientist's Warning: How Terror, Error, and Environmental Disaster Threaten Humankind's Future in This Century—on Earth and Beyond*, Basic Books, New York.

Rodriguez, L. (2021), 'Luisa Rodriguez on why global catastrophes seem unlikely to kill us all'.

URL: *https://80000hours.org/podcast/episodes/luisa-rodriguez-why-global-catastrophes-seem-unlikely-to-kill-us-all/*

Russell, J. S. (2021), 'On two arguments for fanaticism'. Global Priorities Institute Working Paper 17–2021.

URL: *https://globalprioritiesinstitute.org/on-two-arguments-for-fanaticism-jeff-sanford-russell-university-of-southern-california/*

Russell, J. S. and Isaacs, Y. (2021), 'Infinite prospects', *Philosophy and Phenomenological Research* **103**(1), 178–198.

Sagan, C. (1997), *Pale Blue Dot: A Vision of the Human Future in Space*, Ballantine Books, New York.

Savage, L. J. (1951), 'The theory of statistical decision', *Journal of the American Statistical Association* **46**(253), 55–67.

Sen, A. (1977), 'Social choice theory: A re-examination', *Econometrica* **45**(1), 53–88.

Smith, N. J. J. (2014), 'Is evaluative compositionality a requirement of rationality?', *Mind* **123**(490), 457–502.

Smith, N. J. J. (2016), 'Infinite decisions and rationally negligible probabilities', *Mind* **125**(500), 1199–1212.

Snyder-Beattie, A., Ord, T. and Bonsall, M. (2019), 'An upper bound for the background rate of human extinction', *Scientific Reports* **9**(1), 11054.

Thorstad, D. (n.d.), 'Existential risk pessimism and the time of perils'. Unpublished manuscript.

United Nations, D. o. E. and Social Affairs, P. D. (2019), 'World population prospects 2019: Highlights'.
URL: *https://population.un.org/wpp/Publications/Files/WPP2019_Highlights.pdf*

Wilkinson, H. (2022), 'In defence of fanaticism', *Ethics* **132**(2), 445–477.

Yudkowsky, E. (2007), 'Pascal's Mugging: Tiny probabilities of vast utilities'.
URL: *http://www.overcomingbias.com/2007/10/pascals-mugging.html*

Conclusion

This thesis has explored different approaches to cases that involve tiny probabilities of vast value. Chapter 1 discussed one possible approach: Boundedness. Chapter 1 showed that decision theories on which utilities are bounded, such as Expected Utility Theory, violate Ex Ante Pareto if combined with an additive axiology, such as Total Utilitarianism. Chapter 2, in turn, showed that standard axiomatizations of Expected Utility Theory violate Statewise Dominance in cases that involve possible states of zero probability. Chapters 3–6 discussed another approach: Probability Discounting. Chapter 3 argued that Probability Discounting, if plausible, solves the 'Intrapersonal Addition Paradox' and thus helps avoid the Repugnant Conclusion. Chapter 4 explored what the most plausible version of Probability Discounting might look like and what problems the different versions of this view have. Chapter 5 focused on one type of problem, namely, money pumps. The Independence Money Pump, in particular, presents a difficult challenge to Probability Discounting. Finally, Chapter 6 argued that Probability Discounting does not undermine Longtermism, namely, the view that morally speaking what matters the most is the far future.