2023

A Problem Best Put Off Until Tomorrow

Evan Albers
Bowdoin College

A Problem Best Put Off Until Tomorrow

An Honors Paper for the Department of Philosophy

By Evan Albers

Bowdoin College, 2023

I

The Project

That life might go better. Or rather, that it is *good* for life to go better, and that this ought to guide our decision making. This is the crux of utilitarianism. There are a series of particular iterations and considerations that flavor this fundamental principle – whether we ought to consider total utility, average utility, or so on – but the principle remains: achieving more of what makes life go better ought to guide our decision making. Effective Altruism (EA) shares a similar sentiment: we ought to do the most good that we can. There is some overlap between the two – if one adopts a utilitarian view of morality, then EA becomes in effect an extension of utilitarianism. When acting altruistically, we ought to maximize the utility we generate by expending our charitable resources. Because utilitarianism lends itself to quantifying, measuring, and maximizing "good," or rather, utility, it is particularly well suited to a view that prizes being "effective" in its aim. Although EA thinkers typically argue that it is distinct from utilitarianism, it is fairly clear that utilitarianism serves as significant inspiration for EA. Insofar as effective altruists adopt utilitarianism as their criterion of moral good, they might expose EA to many of the problems that plague utilitarianism.

The goal of this thesis is twofold: to establish the impact of a significant objection to utilitarianism on EA, and to determine how effective altruists might respond to said objections. This objection is the Nozickian utility monster. The primary concern is that the unborn millions of future mankind might present a feasible utility monster that could pose a problem to utilitarianism. In order to understand how such an objection might arise, we must first understand EA, and the thinking that has brought it about. We must also gain an understanding of the utility

monster and the moral intuition motivating its force. Having understood EA and the potential for servitude to the monster, we must then consider a specific scenario in which the future might present an undue burden upon the present. Finally, we will consider a number of ways to resolve the objection and examine whether these responses succeed or fail.

Determining what constitutes effective is tricky. Altruism is fairly straightforward: the act of concern for others with no expectation of return. Determining how to be an *effective* altruist, and *why* someone ought to pursue such a goal, is an important question. It seems that effective is entirely contingent on our notion of good, and that even the *aim* of being effective is not necessarily a given. Is *effective* good? What constitutes effective? Answering these questions is the aim of EA. Understanding the evolution of EA, and the problems lurking within the theory, requires a thorough exploration of the ideas that it builds upon. Peter Singer is considered a pioneer of the EA movement, and his essay, "Famine, Affluence, and Morality", introduces key principles that guide the movement. Understanding his work also demonstrates the extent to which EA is inspired by and relies upon utilitarian conceptions of good.

Singer's opening thought experiment in "Famine, Affluence, and Morality" will eventually ground EA in utilitarian principles. He asks us to consider the following scenario: suppose we are walking in the park on our way to work. We witness a child drowning in a shallow pond. After looking around, we realize that there is no one else in any position to help the child; if we do not save them, they will die. The only cost is being late to work, and perhaps mud on our clothes. Should we save the child? It seems fair to say that we should. Singer argues that this is evidence of a more fundamental moral principle. He writes that "if it is in our power to prevent something bad from happening, without thereby sacrificing anything of comparable moral importance, we ought, morally, to do it" (Singer, "Famine, Affluence, and Morality"). This

principle is commonly referred to as the *Principle of Sacrifice.* Singer also clarifies that by "sacrificing anything of comparable moral importance," he means to refer to any moral good, or moral failure, that might occur if we acted to prevent the original moral crisis (Singer, "Famine, Affluence, and Morality"). This clarification explicitly grounds his view in utilitarianism. Singer is justifying judgements of morality based on the utility they generate. In essence, if the marginal utility of helping some individual is greater than the utility that some individual might enjoy by taking another course of action, then the individual has a moral obligation to help. Singer then generalizes his view to justify support for international victims of crises.

Singer extends his view to those suffering internationally with an appeal to general ideas of impartiality. He begins "Famine, Affluence, and Morality" by discussing the impoverished state of the refugees living in Bengal in the early 1970's. Singer argues that his geographic distance plays no role in deciding whether to help the proverbial drowning child. If we were to learn of a child drowning halfway around the world, whose life we could save at little cost to our own, then we ought to save them, much as we would the drowning child immediately before us (Singer, "Famine, Affluence, and Morality"). A life saved is a life saved, and we have no right to discriminate by valuing some lives over others. This also seems fairly reasonable. Furthermore, the cost of providing such aid is typically very little. Given these two scenarios, we construct the following moral prescriptions:

(1) If we can do good without sacrificing equal or greater good, then we ought to.

(2) Simple spatial separation does not obviate this obligation.

The cost of providing such aid, however, is also of moral import. We can see this in (1); we ought to do good, but only if it does not cost something of similar moral import. Sacrifice is

rather straightforward when comparing a human life to a set of muddy clothes, but the moral choice becomes less clear when the costs of acting are greater.

Singer also argues that we have a moral obligation to help strangers, even at great cost to ourselves. In *The Life You Can Save,* a book written to bring effective altruism to a wider audience, Singer offers another thought experiment (originally developed by Peter Unger) to illustrate the moral intuition of saving lives at great personal cost. He asks us to consider Bob, a man who has invested his life's savings into a vintage car that he hopes to sell when he wishes to retire. Bob's ability to care for himself later in life is wrapped up in this car. However, because of its value, he in unable to insure it. One day Bob parks his car at the end up a railroad track. He then notices a train barreling down a parallel segment of track towards a child, who is playing on the tracks, oblivious to the train. Bob is unable to stop the train, and out of earshot of the child. However, he realizes he can throw a switch to divert the train onto the track at the end of which his beloved car is parked. The train will fly off the tracks and render his car a piece of vintage scrap metal, but the child will be saved. Bob considers the cost to himself and decides not to switch the track (Singer, *The Life You Can Save*). Singer briefly discusses the implications of Bob's choice.

Singer acknowledges that the typical moral intuition concerning Bob's choice raises a curious problem for his conception of altruism. He writes that many are repulsed by Bob's decision; how could one possibly value their car over another human life (Singer, *The Life You Can Save*)? The implication of such a view is wider than it may initially seem. Bob's car represents his retirement. Presumably we could have substituted any form of wealth for his car, and the intuition would have remained unchanged. Perhaps Bob has purchased a handful of porcelain pieces dated to the Ming dynasty, and he has stacked them at the end of the tracks. Or

perhaps even a pallet of cash. Regardless, his ability to retire will be destroyed by saving the child's life. Singer argues that disgust with Bob's decision implies that saving for retirement is a morally bankrupt decision. This disgust seems to relate to our valuation of human life; we find Bob's action repulsive because we value human life more than we value his car (Singer, *The Life You Can Save*). However, this intuition requires some clarification.

Utilitarians like Singer might argue that Bob should let the train plow over the child to preserve his car. If we were to determine that the cost of saving a life were less than $100,000, then we would conclude that Bob should save the car, sell it, and save the other lives that might be saved with the value of the car. But this argument misses the intuition that Singer is attempting to demonstrate. Bob need not be a utilitarian. Recall that he has already set this car aside for retirement. Singer is instead trying to demonstrate that even when the costs are extremely significant to individuals, we still would likely prioritize saving the child's life. This intuition also holds for the utilitarians who would leave the child as the mercy of the tracks. In their case, the threshold for saving a life is simply lowered. They would hold the same intuition if, for whatever reason, saving a life would otherwise cost more than $100,000. In fact, if it were more expensive to save a life. they would consider letting the child die to be a relative bargain. More interesting is what this scenario says about the notion about saving for retirement.

We can begin to see the contours of an objection to Singer's conception of morality in Bob's choice to save his car. Bob has purchased his car with the money he set aside throughout his life, with the eventual hope of achieving a comfortable retirement. If we grant the intuition that Bob ought to save the child, then it seems as though the burden of doing so is overridden by the benefit of saving the child's life. But why then, should Bob have ever saved such money in the first place? There are people dying all the time, not merely when Bob has finally purchased his

car. If Bob is obligated to save the child on the tracks, then similar thinking would obligate him to save the people suffering while he was saving for retirement, if his retirement dollars could achieve such an aim.

Singer's principle of sacrifice seems to suggest that individuals might be morally obligated to sacrifice far more than muddy clothes or even their retirement. If we consider that fundamentally, the criterion of morality is maximizing utility, as Singer believes it to be, then we ought to give until giving is no longer the utility maximizing course of action. If only a handful of persons practice effective altruism, however, this will most likely involve reducing their own standard of living to that of those whom they seek to help. Consider the case in which saving a life across the globe cost a penny. If I were to commit myself to effective altruism, then for every penny I spend I ought to consider: will this penny benefit me as much as much as it might benefit someone whose life it could save? The answer is no. And it cannot until the most utility that that penny could possibly provide is the difference between life and death for myself. This, in turn, leads to a more disturbing scenario regarding the value of life.

Singer's principles might compel one to save lives at the cost of their own. Consider the following scenario: Alice can choose to either donate two dollars of her own wage to purchase mosquito nets for children in Africa, or she can spend it on calories to sustain her existence. Alice is a day laborer who for whatever reason, be it her age, or perhaps mental disability, cannot learn any skills to increase her wage. She will at most be able to donate what she earns. And so she has; the cost of saving lives in the developing world is so low that Alice has starved herself for weeks. She lives in a city, where food is quite expensive. Food is so expensive, in fact, that Alice can save more lives by donating every penny of her earnings, rather than sustaining herself. Additionally, the volume of suffering in the developing world is so great that this utility

comparison is static; far more lives can always be saved by donating than by eating. As a result, Alice dies. Furthermore, it seems that by the principle of sacrifice. Alice has done the right thing. So long as Alice's death saves more than one life, she ought to give everything.

It might be argued that considering a longer-term view of Alice's possible contributions to saving lives might obligate her to at least survive. If she might be able to save more lives in the long term by staying alive and continuing to contribute, then she wouldn't be obliged to starve to death; rather, she should simply continue to live at a level that allows her to maximize her possible earnings, and by extension, the amount of good she can do. Unfortunately, this response fails to completely allay the worry regarding Alice. Perhaps in a world in which Alice can better herself, she might devote resources to doing so. But Alice cannot. We might even consider a scenario that destroys Alice's ability to do good for the developing world. Consider a rash of inflation that hits the city in which Alice lives; perhaps there is an influx of extremely wealthy migrants that drive up housing prices. Alice's wages don't increase, but her costs of living do. As a result, Alice can no longer afford to both live and save lives; if she were to live out the rest of her predictable life, she could save no more lives. However, if she were to starve herself, she might be able to prevent the immediate death of at least two people in the developing world. Because preventing two deaths is greater than the one death she would otherwise prevent (her own) by not working, Alice seems obligated to starve yet again. This example is fairly contrived. However, I only really wish to demonstrate that the Principle of Sacrifice *could* place excessive demands on its practitioners. Singer recognizes something similar to Alice's scenario in "Famine, Affluence, and Morality", and weakens his fundamental principle to address this objection.

Singer defends the principle of sacrifice by arguing that excessive sacrifice would only occur if individuals gave alone or believed themselves to be giving alone. In other words, if everyone in some society gave a small proportion of their income, no single person would be obligated to give everything – the fractional portion would satisfy the needs of the developing world. There would be no need to give beyond that which satisfies the needs of the developing world. However, this amendment alone is not sufficient to resolve the demanding nature of the principle of sacrifice. It might still be argued that so long as the developed world enjoys a standard of living greater than that of the developing world, the developed world ought to donate more. Recognizing this, Singer tempers the principle of sacrifice somewhat. Rather than demand that we be willing to sacrifice anything of comparable moral import, we only need be willing to sacrifice something of moral insignificance (Singer, "Famine, Affluence, and Morality"). This is a good thing for Alice; her death is morally significant. Therefore, she ought not bring it about by starving herself to death in the service of an unseen multitude. Muddy clothes, however, are morally insignificant, so we still ought to save the child. This new principle is rightfully referred to as the *Weak Principle of Sacrifice* (Singer, "Famine, Affluence, and Morality").

The Weak Principle of Sacrifice has had significant influence on the effective altruist movement. Singer's ideas inspire an early article written by significant authors within the Effective Altruist movement. His influence is derived from the Principle of Sacrifice (PS) and the less demanding Weak Principle of sacrifice (WPS), which I paraphrase here:

(PS) If some action *A* accomplishes some moral good without sacrificing something of comparable moral good, then we ought to do *A*

(WPS) If some action *A* accomplishes some moral good without sacrificing something of moral significance, then we ought to do *A*

These principles are accompanied by a set of principles related to the valuation of good. Let them be the time and space principles (TP, SP), respectively:

(TP) Temporal position has no bearing on the value of utility; a util now is as worth as much as a util tomorrow, or yesterday.

(SP) Spatial position has no bearing on the value of utility; a util here is the same as a util across the ocean, or even across the universe.

These principles conspire to provide guidelines for moral action. Most importantly, they characterize an effective lower-bound on the obligations of morality; in conceding the strength of altruism's moral obligation, those trying to do good are settling for a middle ground that they know will improve the quality of lives around the world without incurring an unpalatable cost to the present. This central idea – that we ought to temper the demands of morality to arrive at a more effective, actionable theory of morality – inspires early EA literature.

"Giving Isn't Demanding," an essay on altruism, presents early arguments that strongly resemble tenets of effective altruism. The authors take WPS and modify it to provide an even less demanding view, with the goal of providing an realistically feasible standard of altruism that still maximizes good. Their effort is motivated by the worry that even WPS is too demanding, and that its demanding nature outstrips the obligatory purview of morality. Their remedy is the *Very Weak Principle of Sacrifice* (VWPS): generally speaking, middle-class members of affluent countries should contribute 10 percent of their income to *effectively* improve the lives of others (MacAskill et al.). The authors justify VWPS on based on two empirical premises: that giving financial resources is not very demanding, and in fact less demanding than many believe, and that financial resources do significant amounts of good, typically more than many believe. The authors' advance empirical evidence to support the VWPS. However, the strength of this

evidence's support of VWPS relies on the author's conceptions of effective and demanding, words which prescribe the strength of the principle.

There are two significant definitions underpin the author's definition of VWPS. The first is that *effective* is stipulated to refer to improvements that benefit people's lives far more than the typical improvement. The authors stop short of demanding the *most* effective possible use of resources; however, they do emphasize that not just any contribution will do (MacAskill et al.). The second important definition relates to how the authors define demanding. Whether some moral theory is demanding is determined by the ratio of how much one benefits others, relative to the cost to oneself (MacAskill et al.). It should also be noted that the authors are specifically defending the Very Weak Principle, not utilitarianism, or any form thereof. One's obligation to give their proportion of income is not contingent on maximizing the good, but rather the related observations that giving is typically easy, and that giving does significant good.

Giving is easy because the cost of doing so, in terms of well-being, is low. Well-being and income appear to have a logarithmic relationship – significant decreases in income correspond to less drastic decreases in well-being. Furthermore, giving financial resources is not akin to simply setting money on fire. There seems to be some evidence that donating money to others increases one's own welfare more than simply spending it on oneself. In short, in the worst-case scenario, in which donating money brings one no benefit whatsoever, the costs seem to be small. In the best-case scenario, we offset some of the loss in utility from donating via an increase from altruistic fuzzies of some kind. In either case, the authors conclude that VWPS is hardly demanding in practice (MacAskill et al.). Establishing that VWPS isn't demanding resolves much of the concern that one might have had with regards to the demands of WPS. The authors also argue that the VWPS still fulfills much of the moral intent of WPS.

Giving is also strikingly beneficial to the recipients. Although the VWPS only mandates donation of 10% of a middle-class individual's income, because of the skewed distribution of global wealth, this 10% can result in drastic improvements for many. Recall the logarithmic relationship between wealth and happiness. A doubling of wealth is required to achieve some standard increase in one's level of happiness. This relationship becomes much more profound within the context of global wealth inequality. A person earning the median US wage (at the time of writing of the article) of $28,850 a year would enjoy some about of increase in their happiness by having their salary doubled. A person earning the average Somali GNP per capita (a figure typically higher than a country's median wage), would need merely $133 to achieve a comparable increase in happiness (MacAskill et al.). By donating a fraction of their income, a middle-class citizen in the developed world could vastly increase the happiness of someone in the developing world.

This is also not an unrealistic view to take in the face of the transactional costs involved in donating to people across the world. Some charities can transfer donations at a rate of 90% - i.e., for every $100 donated, $90 are given *directly* to an individual in need (MacAskill et al.). So, it seems that even in the absence of a crisis of the sort that Singer details in "Famine, Affluence, and Morality," giving can still do incredible amounts of good. This further justifies the authors' revision to the WPS; if we can still affect significant good, and doing so isn't particularly demanding, then we ought to do so. It is important to note that this is a normative claim. VWPS is presented not merely as advice, but as a moral obligation.

VPWS constitutes an attempt at retaining the moral obligations of WPS, and to a lesser extent, PS. It achieves a similar, if somewhat watered down, aim of maximizing the utility one would generate via charitable decisions. However, the VWPS is not yet EA. EA adopts much of

the fundamental reasoning underpinning PS, WPS, and VWPS; namely that we ought to try and maximize utility in our charitable giving. In the quest for broader applicability, however, EA abandons specific characterizations of good, and in doing so loses some fidelity regarding the moral aims that motivate WPS and VWPS.

EA bears a strong resemblance to the Very Weak Principle yet seems to be somewhat distinct. It is difficult to pin down exactly what Effective Altruism is. There are a series of definitions and principles offered by a handful of authors, although few purport to give a "true" definition of Effective Altruism. One of the most definitive definitions stems from William MacAskill's forthcoming submission to the International Encyclopedia of Ethics. MacAskill defines Effective Altruism as "the *project* of using evidence and reason to try and find out how to do the most good, and on this basis, trying to do the most good" (*Effective Altruism - MacAskill - Major Reference Works - Wiley Online Library*). He also specifically notes that Effective Altruism is *not* a normative theory or claim. He argues that this distinguishes it from utilitarianism, and similar principles like PS, or VWPS.

While MacAskill's use of the term *project* to describe EA makes it difficult to describe concrete attributes of EA, he does mention some important notions regarding the project. According to MacAskill, EA is *welfarist,* and *impartial*. Human welfare is the sole object of concern for EA. However, because EA is a "project," and not a normative claim, MacAskill argues that EA is compatible with non-welfarist conceptions of the good. EA's impartiality is straightforward; it does not discriminate based on class, race, gender, and so on. MacAskill notes that this makes it compatible with Prioritarianism (*Effective Altruism - MacAskill - Major Reference Works - Wiley Online Library*). These principles strongly resemble characteristics of Singer's Principle of Sacrifice, and for good reason – MacAskill helped introduce the Very Weak

Principle of Giving as one of the authors of *Giving Isn't Demanding.* His efforts to exclude Effective Altruism from normative theories generally demands a more thorough examination.

MacAskill's designation of Effective Altruism as a *project*, rather than a normative claim, seems designed to preempt many objections that typically plague utilitarianism. By calling Effective Altruism a project rather than a normative claim, he can immediately wave away any sort of objection that deems Effective Altruism too demanding. This seems reasonable at first; MacAskill is simply trying to prevent perfect from being the enemy of the good. Rather than worry about what would happen if we were all perfectly obeying Effective Altruism to its logical extreme, we simply ought to try it a reasonable amount. Clearly there is at least some good that can be done. If it leads to uncomfortable choices, then one can simply abandon the project and return to whatever higher moral code they seem to be otherwise following. However, this approach seems somewhat odd. At points it feels as though MacAskill is trying to argue with the force of a moral theory, while avoiding the vulnerability that doing so entails.

One can practice EA (or rather, take part in the project thereof), without actually being the most effective altruist possible. MacAskill notes that there are significant philosophical and empirical barriers to being the most effective altruist, and that because of these barriers, simply trying to be effective is engagement enough. Such barriers include the valuation of future lives, or perhaps animal life. Without answers to such questions, the criterion of effectivity will be unclear. He also allows that effective altruists might be reasonably permitted to pursue other important projects in their lives. Even those who doggedly pursue being effective altruists as the central project of their life do not necessarily qualify as perfect effective altruists(*Effective Altruism - MacAskill - Major Reference Works - Wiley Online Library*). Given MacAskill's

rather specific definitions and allowances, VWPS almost sounds like a more compelling account of EA than his specific definition.

Despite attempting to solidify a concrete conception of EA, its definition remains elusive. EA almost sounds as though it is simply utilitarianism rigorously enforced; rather than blindly donating to charity, we ought to be selective in our giving to maximize utility. This is strikingly similar to Singer's original justification of the Principle of Sacrifice. Yet Effective Altruism is not normative – it does not prescribe a criterion of morality. But then why *should* anyone be an effective altruist? If it is for moral reasons that someone wishes to take part in the project, this seemingly would, by extension, render EA normative in nature. But even this attempt at squaring the seeming moral nature of EA with its supposed status as a project is weak at best.

From a philosophical perspective, it is difficult to attribute much to a project. In this sense, there is no real proposition or premises to be discussed, nor objected to. EA appears to be a vessel for the fairly vague aim of doing well, along whatever criterion one chooses. In the interest of exploring the potential consequences of adopting EA, I plan on investigating what might be said *if EA were practiced with utilitarianism as a strict conception of the good.* This conception of EA will be defined as follows:

(EA) We ought to give to effect the most good we can with the resources we have available to us

This phrasing is deliberately unclear. It might reasonably be interpreted as either PS or WPS/VWPS. I have selected such a phrasing to remain true to the seeming intent of the EA movement. I will investigate the implications of interpreting EA as either PS or WPS/VWPS, and attempt to determine what exactly each view implies. At first glance PS appears to commit EA to a strictly utilitarian viewpoint, while WPS/VWPS offer more leeway in terms of normative

strength. EA is particularly vulnerable to concerns regarding the Nozickian utility monster. Proceeding with a strict conception of the utilitarian obligation in EA facilitates a more straightforward demonstration of the problems that the monster entails and potential responses. Therefore, while these problems generalize to weaker conceptions of EA (those closer to WPS/VWPS), I will spend the majority of the proceeding discussion presuming EA to be closer to PS, and will generalize the objection to apply to the weaker version of EA after establishing it. I will proceed now by enumerating further implications of the strict EA (hitherto simply EA).

II

The Worries

EA gives us a strikingly simple criterion of morality. If some act allows us to do the most good that we can with what we have, then it is the right action to take. But this definition glosses over significant presumptions. The first is measuring exactly *what* good is. The second is consider *whose* good matters. The moral force of the utility monster objection stems from the ambiguities of the answers to these questions. These will be discussed at length, but for the moment, it is worth establishing how the utility monster represents an objection to utilitarianism, and how its frustrations for utilitarianism extend to EA. To do so I will begin by considering Longtermism, a view prioritizes helping the future over helping the present. Longtermism has evolved from EA and provides a framework that makes clear EA's vulnerability to the monster. I will then offer a more robust discussion of the monster to enumerate exactly *why* the objection applies to utilitarianism, and by extension EA. Finally, I will consider a series of objections to Longtermism as offered by Longtermist thinkers as potential objections to the view, in hopes that one of these objections will prove significant enough to reject Longtermism, thereby saving EA.

Measuring exactly what the *good* is, is difficult. It is possible that many can agree on maximizing the *good*, but if everyone has a different criterion of good, then this isn't necessarily effective, or in some cases, desirable. If right-wing nationalists were to be taken with EA, it is unlikely that many would find their actions particularly altruistic. If one simply strips a portion of humanity of their human status, then they lose moral significance. However, the question of *who counts* has other significant implications for Effective Altruism beyond the scope of right-wing altruists.

Utilitarianism, and by extension, EA must wrangle with the issue of future generations. EA gives us the aim to do as much good as we can with our lives, without sacrificing anything of comparable moral import. It seems plausible, then, that the good we can do can stretch far into the future. Because our decisions today have significant ramifications for the future, we must include future generations in the calculation of the welfare-maximizing course of action. But how should we go about accounting for the unborn? Wrestling with the role of future generations has in part given rise to the study of Longtermism – the view that the most important effects of our decisions are those in the far future, because the unborn population of the future is far greater than that of the present, and this unborn population will bear the costs of our decisions in perpetuity (Greaves and MacAskill). Longtermism is a natural extension of the motivation behind EA. If we wish to maximize the good we do, then a reasonable start is likely the unborn billions that might benefit (or suffer) by our actions today. However, it is distinct from EA, and has its own series of potentially extreme implications.

Longtermism is fundamentally a way of assessing the state of the world. Hilary Greaves and William MacAskill present both an axiological version of Longtermism and a deontic one in "The Case for Strong Longtermism". The axiological version does a reasonable job of

illustrating the fundamental thinking of Longtermism. Greaves and MacAskill define ASL as follows:

(ASL) In the most important decisions facing agents today,

(i) Every option that is near-best overall is near-best for the far future.

(ii) Every option that is near best overall delivers much larger benefits in the far future than in the near future (Greaves and MacAskill).

The far future is taken to mean everything more than 100 years from the point of decision. Near-best is taken to be the *ex-ante* value of some course of action, given the information available to agents at the point of decision (Greaves and MacAskill). In short, Greaves and MacAskill argue that any choice that is optimal overall, will be optimal for the future, and that the majority of any benefits stemming from such a choice lie in the far future rather than in the near future. They justify this position by appealing to the *benefit ratio* of potential courses of action.

The benefit ratio of future benefit to past benefit justifies the claims made by ASL. More specifically, Greaves and MacAskill define the benefit ratio (BR) as the claim that

(BR) The highest far-future *ex-ante* benefits that are attainable without net near-future harm are many times greater than the highest attainable near-future *ex-ante* benefits (Greaves and MacAskill).

They also prove that if BR holds for a given scenario, then so too does claim two of ASL, as well as claim one of all options involving no net expected near-future harm. Evaluating BR is a matter of quantitative analysis, and the authors make a series of important assumptions in their evaluative process. They define the *ex ante* value of some choice to be the *Expected Value* of that

choice. They also presume a total utilitarian axiology – total welfare is the chosen criterion of value (Greaves and MacAskill).

BR relies significantly upon the size of the future. If the number of humans that exist in the future is small, then the total amount of utility they can possibly experience is limited by their population size. Greaves and MacAskill examine a series of potential scenarios for future population sizes. They first consider the UN's estimate that Earth's population will plateau around 11 billion people within a century or so. They also consider some who argue that in the far future, technological innovations might enable a population in the area of 1 trillion. However, what is most important for defending BR is the *expected* number of future humans. This number captures the uncertainty inherent in making estimations about future populations. The authors point out that the math of expected values lends an extreme asymmetry to discussions of expected population. Even if one were 50% confident that the future human population would be zero, this would only divide the expected population of the future in half. By contrast, a one percent increase in the chance that the future human population consists of 1 trillion people per century, per 100 years, for 100 million years, will increase the expected population by a factor of 100 (when compared with a population of 10 billion, for 1 million years) (Greaves and MacAskill). Small changes in the confidence of extreme population numbers have an outsize effect on the expected future population.

Such spectacular numbers can seem like fantastical thinking. However, should they be true, the potential for utility is immense. While these changes may seem extremely unlikely, Greaves and MacAskill point to possible future developments such as space travel, and resulting interstellar settlement, or perhaps the rise of conscious AI as massive potential sources of future moral agents. They examine the settlement of the solar system to arrive at what they believe to

be a reasonable estimate of the number of future humans. On their view, a reasonable number for the expected total future population of humanity (i.e., how many humans will live before we die out) is 10^{24}. Such a number certainly gives the future significant representation in the moral congress of humanity. The current population of the world, around 8 billion, or 8 times 10^9; we of the present are hardly worth peanuts. Even if the probability of achieving such a volume of humanity were astonishingly small, perhaps one millionth of a percent, we would still be outweighed by the future, 10 billion to one (Greaves and MacAskill). Having established that the future mass of humanity is likely to be very, very large, Greaves and MacAskill use the expected mass of humanity to argue in favor of ASL.

I intend to use ASL as a mechanism for demonstrating a potential objection to the utility monster objection. Therefore, I only need provide one instance in which the benefits to the far-future are far greater than any comparable benefit to the near-future, using the same resources. If there is some course of action whose expected value is greater than all other actions, and the greater part of its expected value lies in the future, then ASL holds for this action. More importantly, this example will serve to demonstrate the objection I wish to make. Greaves and MacAskill offer two possible empirical realities whose prevention could potentially justify BR, and by extension, ASL. I present both of these situations for the purpose of finding an empirical example of ASL.

The future Is large. So, we should worry about ensuring that It will In fact, occur. If a mass extinction event were to occur between now, and the existence of the untold future masses, then we ought to worry about preventing such an event. One straightforward example includes the detection and deflection of asteroids. Such work is fairly cheap in today's terms – perhaps only just over $1 billion to ensure near certainty of prevention – and would reap massive rewards

in preventing a possible mass extinction that would eliminate the possibility of future humans. Greaves and MacAskill conclude that for every additional $100 spent, given their main estimate of future humanity (10^{24}), 300,000 lives are saved (Greaves and MacAskill). Another possible area of prevention relates to the development of artificial intelligence (AI).

There are a handful of ways in which the development of a super-intelligent AI system might negatively affect future utility. Such a system might "go rogue" and attempt to either eliminate or enslave humanity. It might also enable nefarious human actors to achieve total and permanent control. It is unclear on what grounds exactly Greaves and MacAskill can really call such a scenario a disaster. Earlier in the paper, they mention a possible source of future agents to be conscious AI systems that have moral significance (Greaves and MacAskill). If an AI were to wipe out humanity, or even simply prevent it from reproducing, and repopulate the future with more AI agents, then this would seem to be just fine. In fact, it might even be considered better than allowing humanity to continue; digital beings can be created nearly ad infinitum, while humans are a touch more resource intensive, with our pesky hunger and thirst. However, it will simply suffice to recognize the prevention of malicious AI as a pursuit that supports ASL.

It might be noted that a simple way to resist ASL within the context of these examples is to adopt average utilitarianism rather than total utilitarianism. This is reasonable, to a certain extent, but fails for reasons worth discussing at a later point. There are other avenues of resisting ASL that might succeed in both total-utilitarian and average-utilitarian terms, and I wish to explore these avenues before potentially ruling out total-utilitarianism entirely.

Longtermism, generally speaking, adopts the view that, because of the likely massive size of the future, we ought to prioritize the far future as a general rule in our utility maximizing existence. While there is a caveat pertinent to avoiding near-term harm, this functions as a fairly

limited constraint on the import of the future. Already one might conceive of worlds in which we forgo reasonable expenditures that pay off handsomely in the present and pay off less for the far-future, but which are forgone for the sake of prioritizing the future and its corresponding utility. Consider a world in which we must decide between purchasing antimalarial bug nets and investing in asteroid prevention. Greaves and MacAskill offer the purchase of bug nets as a plausible candidate for an expenditure that maximizes near-future utility on a per-dollar basis; it is the quickest, cheapest way to do the most near-term good. They argue that even this highly effective way of increasing welfare pales in comparison to the gains from something like asteroid deflection (Greaves and MacAskill).

I find this tradeoff between present and future utility somewhat disturbing. To a certain extent, it feels as though we are trading a slight decrease in the probability of some future lives perishing for the near certain death or horrible suffering of humans alive today. This is only possible due to the *potential* size of the future. But if prioritizing the unborn future is how we ought to act, then where does it stop? In some ways, it feels as though the concerns of the present are completely drowned out by the potential moral import of the future. I want to be able to say that we should save the millions suffering from malaria immediately alive today rather than devoting ourselves to a wildly uncertain future. I find that the following example demonstrates this moral intuition nicely:

> (Uncertain Disease) Consider the possible mutation of a disease. This disease will become endemic despite our efforts to resist it and affect the remainder of the history of humanity following its appearance. It does not result in significant disutility on a per-person basis; perhaps one merely has a head-cold for a day, and then makes a complete recovery. Thankfully, we can prevent its occurrence with a varying degree of uncertainty,

> depending on how many resources we dedicate to its prevention. Unfortunately, we cannot be completely sure of its prevention; we can be *extremely confident*, with probability of prevention approaching one, but never certain.

The present might be compelled to pay an unreasonable price to prevent the uncertain disease. Suppose that present-day scientists are aware of the possibility of the uncertain disease. They already believe that they can prevent the occurrence of the disease with 95% certainty. However, achieving additional increases in certainty is wildly expensive; it would require the defunding of programs that save lives in the developing world, and likely result in millions of preventable deaths in the present. Despite the uncertain disease's low costs on a per-person basis (a day-long cold), the simple volume of the future outweighs any consideration of the present. This is even considering that the uncertain disease is already extremely unlikely to occur. However, even a small additional increase in the likelihood of its prevention would still outweigh the costs to the present; only so many will die today, while untold trillions could suffer a cold for a single additional day over the course of their lives. This hardly seems morally excusable.

We can avoid such a scenario by prioritizing the present over the future. I worry that if we don't do this, the utility of the future will negate the interests of the present. Nozick's utility monster is an older objection to utilitarianism that gives voice to this very worry.

The utility monster is an ugly result of adherence to the utilitarian principle. Nozick proposes a character who is immune to the tyranny of diminishing marginal utility. Many of us experience less utility from successive consumption of some good; a few pastries are good, but each additional pastry is rarely as good as the last, and eventually one just wishes to stop eating all together. This is no issue for the utility monster. Each additional pastry is equally as delicious as the last. In the proverbial race for utility then, the monster will outlast us all; he simply never

gets tired of pastries, and so long as he enjoys it more than we might even the first, then our moral import will never outweigh his. For a society that is observing utilitarian principles, the utility maximizing choice always involves serving the monster. Any pastry that we chance upon by virtue of luck or manufacture must be given to the monster, as this would maximize utility. This might hold for either the consideration of total utility or average utility – in the latter case the monster merely needs to enjoy significantly more utility from any given resource to outpace the average increase, but it is still logically possible to conceive of such a thing. Nozick charges that the potential existence of the monster represents an objection to utilitarianism (Nozick). It seems as though utilitarianism would bind society in service to the monster by simple virtue of its spectacular disposition, and Nozick finds this disturbing.

Perhaps the monster is simply too outlandish to conceive. After all, it seems to require an individual capable of experiencing nearly infinite utility, and this seems fairly unrealistic. Is it so fair to discard a moral theory on the grounds of an impossible being? Perhaps not. However, if one were to propose an alternate form of monster, the objection might recover its force. Consider instead, all the people that might potentially live. Depending on our presumptions about the probability of their existence and the magnitude of their population, our decisions today might influence untold quantities of utility. Perhaps, as long as enough people benefit, we may find ourselves bound to serve the future by the principle of utility. There will always be those humans yet unborn. Because their utility hinges on our present decisions, we may find ourselves obligated to subordinate our own interests and desires, even those we find reasonable or feel intuitively entitled to, to the interests of the future. This mass of future humanity might act, in aggregate, as a utility monster.

The leap to servitude involves a series of generous assumptions regarding the nature of morality and our decision-making process. Perhaps the future cannot represent such a monster, because to serve others is a concept inherent to morality. It would be odd to envision a moral code that simply faltered because there were too many people who would benefit by its adoption. Morality frequently compels us to assist others, and considering the future merely represents an extension of this obligation. However, if we presume for the moment that enslaving the present to the future is in fact an undesirable outcome, then utilitarianism (and by extension, EA) must muster some kind of response. There are also many who would agree that enslavement to the future is an undesirable result. The matter at hand, then, is to find a way that enslavement need not stem from the principle of utility.

Longtermism is the mechanism by which the unborn millions might enslave the present. Recall that rather than a single entity, the utility monster might instead be the corporate interest of millions of unborn future humans. We must therefore find some reason to resist Longtermism, and its claims regarding the primacy of the future in the hopes that this would provide a more general reason for reasserting the import of the presence. By weakening the importance of the future, the present might not be so beholden to it, thereby avoiding the utility monster characterization of the future, and saving EA from the utility monster objection. MacAskill and Greaves outline a series of objections to Longtermism that might accomplish this. I will consider these objections throughout the rest of this thesis. I will list the objections here, in addition to providing a brief overview of each, beginning with the arbitrariness objection.

The first objection is that much of the supposed empirical support for Longtermist thinking is grounded on generous presumptions about the state of the world, or even outlandish ones. Rejecting these presumptions might weaken Longtermism (Greaves and MacAskill). The

second accuses the construct of expected value of leading to decision making that results in utility-monster like devotion to the future. We might be able to find some method of *ex ante* assessment that does not result in prioritizing the future, therefore ASL would fail(Greaves and MacAskill). The third potential objection relates to the observation that, in their calculations, Longtermists typically adopt a discount rate of zero; utility in the future is equally as valuable as utility today. If even a slightly non-zero discount rate were adopted, arguments about the import of the far-future might collapse (Greaves and MacAskill). I consider each of these objections in the order so detailed. I consider each with the hope that it will succeed, and no further consideration will be necessary. I also consider Derek Parfit's Repugnant Conclusion, its corresponding implications due it its largely similar structure, and adopting average-utilitarianism in response to The Repugnant Conclusion to mitigate the import of the future. However, I find that each objection to Longtermism, and the adoption of average-utilitarianism, fails to provide a compelling resolution to the problem of the utility monster, and am forced to consider whether we must reject utilitarianism, and by extension, EA, altogether.

Asteroids and killer AI. 10^{24} future humans. It is fair to say at this point, that evaluating quantitative support for ASL has ventured so far into the weeds that it feels somewhat arbitrary. Any number of wild assumptions could produce similarly shocking results to support just about any quantitative argument. If this were the case, then ASL hold little value to those who disagree with the assumptions mustered to support the scenarios in which it holds. Greaves and MacAskill recognize this and acknowledge it as a contentious objection. They argue, however, that there is a difference between some set of assumptions being reasonable, and there being no logical argument against such assumptions. If I were to assume that in 100 years a series of climate catastrophes would annihilate humanity, there is no indisputable way to disprove my assumption.

One might, however, argue that such an assumption is unreasonable. Greaves and MacAskill argue that the assumptions that would not support ASL in any way are exactly such unreasonable assumptions (Greaves and MacAskill).

Independent of how reasonable or unreasonable statements about the future are, it seems that Longtermism might present a certain level of concern regardless. We might still reasonably be troubled by the *potential* for extreme demands on the future, even if it is impossible to evaluate possible states of the far future given modern methods for doing so. Evaluating the potential possible states of the future in a systematic and meaningful way is also a task that could fill not just one thesis, but many. Insofar as the scope of responding to the objection is beyond that of this paper, and that we still have reasonable concern for the *possibility* of a Longtermist type-world (one in which the future far outweighs the present), I will proceed as though this objection has been resolved and will presume for the sake of argument that the Longtermist predictions of the future are in fact, beyond reasonable doubt. More specifically, this entails conceding to their predictions of the size of future humanity, and their calculations regarding the probabilities of events far in the future. In short, I am presuming that the examples so far given are in fact the optimal course for maximizing far-future utility, and that the utility of these actions has been appropriately calculated. Objections to the use of expected value as a criterion of welfare, however, are potentially much more significant.

ASL lends itself to a certain decision-making process that could give pause to any sort of rational risk averse agent. Consider a scenario in which we are offered one of two choices. One can save 1000 lives with complete certainty. Alternatively, one could accept an infinitesimally small chance of saving many, many more lives, but this chance comes at the direct cost of the 1000 lives we could have saved otherwise. We shall stipulate further that the number of lives is

so great that even for an extremely small probability of success, the expected value of the gamble is greater than 1000 lives. On the *ex ante* assessment of utility practiced by ASL, one ought to take the gamble. However, this seems intuitively odd. The most likely scenario is one in which 1000 people die for no gain whatsoever. This might go against the moral intuition of wishing to maximize utility; the use of expected value as a function of utility has resulted in an extreme likelihood of zero payoff, rather than a certain significant payoff. I think that the tradeoff between purchasing bug nets, and investing in asteroid deflecting or preventing killer-AI presents a similar dilemma. Greaves and MacAskill refer to such thinking as "fanatical" decision making, so-called for the seemingly fanatical pursuit of expected utility. Given that one might prefer a "non-fanatical" decision theory to a fanatical one, the authors worry that the existence of such a non-fanatical theory would undermine ASL (Greaves and MacAskill).

Support for ASL might be weakened by the existence of a non-fanatical decision theory. However, if it could be shown that the costs of avoiding fanaticism are too high, then the existence of a non-fanatical theory might not be as problematic as it initially seems. Greaves and MacAskill appeal to a thought experiment offered by Nick Beckstead and Teruji Thomas to demonstrate such a scenario. Consider a series of gambles. The first yields a modest benefit, certainly. The last gamble in the series yields an incredibly high benefit, with an exceedingly small probability. Each intervening gamble represents a significant increase in the reward of the new gamble compared with the previous one, and a slight decrease in the associated probability of such a reward occurring. At what point in the series then, ought we stop accepting additional risk? If we do not accept the final gamble, then at some point within the series, we have concluded that accepting *slight* additional risk, in return for *massive* potential additional reward, is not worth the tradeoff (Greaves and MacAskill). Beckstead and Thomas refer to such a

decision as “timidity” (Beckstead and Thomas). Such timidity might reveal, in some scenarios, extreme and counterintuitive forms of risk avoidance.

It is unclear whether Beckstead and Thomas’ series of gambles resolves the fanatical objection. It should be noted that even if there are “timid” agents, and their timidity presents serious costs, it does not alter the original scenario in which fanatical decision-making results in 1000 lives lost, for effectively no gain. Appealing to the irrationality of intermediate judgements involved in arriving at the moral intuition unfortunately does little to alter confidence in the intuition. One might acknowledge that somewhere along the line of ascending gambles, their reasoning fails. However, one will still likely save 1000 lives with certainty over a miniscule chance at many more. The issue of fanaticism stems largely from the use of expected welfare as the method of evaluating future welfare. The St. Petersburg paradox provides another counterintuitive conclusion stemming from the use of expected value as a method for evaluating *ex ante* welfare.

The St. Petersburg Paradox illustrates a key pitfall in relying on expected value as a metric of future welfare. I provide here a somewhat stylized conception of the paradox to demonstrate the nature of the objection it presents to expected value. Consider Peter, an ordinary fellow with a comprehensive grasp of probability theory. He walks into a café in Washington D.C. As he walks in, he holds the door for a stranger behind him. The stranger happens to be Jerome Powell, chairman of the Federal Reserve. Out of gratitude to Peter, Powell offers him a potential deal. Powell might, for the right price, be willing to play the following game with Peter: Powell will flip a coin until a head shows. He will then pay Peter $\$2^n$, where n is the number of total flips. How much should Peter be willing to pay for such a game? Powell can clearly back it up; he does, after all, control the money supply. Peter, dwelling on this question himself, recalls

his comprehensive understanding of probability theory and calculate the expected value of the game. It is infinite. The decrease in probability of achieving $n + 1$ heads is matched by a proportional increase in the payoff of the game. The value then sums to infinity.

Powell's game results in a curious dilemma for Peter. His understanding of probability tells him that the expected value of the game is infinite. Therefore, he should be willing to pay nearly any price. Powell is cutting him in on the deal of a lifetime. And yet, common sense screams that one ought to pay no more than a handful of dollars for the opportunity. Anecdotally, I have never observed a run of tails more than four or five in my life. The probability of winning eight or fewer dollars (two tails, followed by a head) is 87.5%. And yet, according to the expected value of the game, Peter could fork over every resource available to him, and he would still be getting a deal. While such a game seems far removed from reality, the example is still illustrative of the potential problems wrought by expected value.

It might be the case that fanaticism is not nearly as counterintuitive as it may seem. Greaves and MacAskill point out that the risk of dying from either cycling 35 miles, or driving 500, is one-in-a-million. Yet many are still willing to wear a helmet or seatbelt to further reduce this risk (Greaves and MacAskill). It might be reasonably observed that this is not the thinking that typically leads many to don their helmet or seatbelt. There is a wide range of states between healthy and dead that may occur as a result of a driving or cycling related accident. I wear a helmet while riding to prevent death in some cases sure, but also to hopefully avoid being rendered catatonic by the front bumper of a two-ton pickup on a country road in Maine. The given example notwithstanding, it is still plausible that there are other empirical scenarios in which we find fanaticism reasonable. Providing said scenarios would weaken the strength of the fanatic objection to ASL.

ASL must also contend with whether we should care about the future. All of Greaves and MacAskill's calculations proceed with the assumption of a time discount of *zero*. This would indicate that future life is worth exactly as much as life in the present. The authors recognize the role that such an assumption plays in supporting their argument. Even a modestly positive discount rate would quickly drive the value of future life to zero. They argue that such a presumption is justified because it is widely accepted among moral philosophers, and somewhat so amongst influential economists (Greaves and MacAskill).

I also consider whether the adoption of average-utilitarianism in place of total-utilitarianism might reduce the import of the future. Consider why the future exerts such influence over the present: there are many trillions of future humans, each of whose utility matters. If one were to consider the *average* utility at any one point, then there would be no overriding influence due to size alone; if some course of action that resulted in an increase in the average of the future, and was accompanied by significant decreases to the average of the present – which is easily possible given a scenario like the uncertain disease – then such courses of action would not be granted the status of moral good that they otherwise might have been under total-utilitarianism.

The principle of utility holds that we ought to maximize utility. Effective Altruists simply wish to carefully apply such a principle, generally speaking. They argue that charity is not a special case; we ought to bring the same scrutiny to bear in our charitable decision making process as in any other. Singer argues that we ought to consider those distant spatially, and typically these people that those in the developed world can help most. However, there is another group that, by virtue of its size, might benefit even more, in total, from our present consideration: the future. In fact, the size of the future might be so great that it swamps any consideration of the

present. We wish to avoid such a conclusion. It seems that there are at least a handful of plausible avenues of investigation: we might reasonably argue that using expected value as a decision-making process is counter-intuitive, or we might simply argue that future lives ought to be discounted in some way. If either of these objections holds in sufficient force, we might argue that the interests of future humans do not in fact outweigh those of the present, if either due to the failure of present methods of evaluating the value of decisions, or due to the inherently lesser value of future utility.

III

A Question of Fanaticism

The future, by virtue of its mass, seems to weigh on the present. In this chapter I consider the possibility that this phenomenon stems from using expected value as a method of *ex ante* welfare assessment. We typically adopt expected value as a reasonable way to evaluate future decisions. Because we face an uncertain future, we incorporate our estimate of the probabilities of certain occurrences into our valuation of courses of action. However, we need not necessarily adopt expected value as our method of evaluation. I consider the possibility that the purported influence of the future is merely a mathematical fiction stemming from the use of expected value. If it were to be the case that expected value is in fact, a counter-intuitive method of ascribing value to uncertain decisions, I could argue that the future might not constitute a utility monster under another system of valuation. However, if expected value is a reasonable metric by which to value uncertain decisions, then I cannot argue that the extreme influence that the future exerts on present-day utility calculations is inappropriate as a result of using expected value to

assess the value of uncertain decision. Such a conclusion would eliminate a promising avenue of resolving the utility monster objection to Longtermism, and by extension, EA.

I consider two examples in which using expected value for the purpose of decision-making leads to counter-intuitive behavior. The first example relates to pairwise comparisons, and a more general obsession with extreme payoff commonly referred to as "fanatic" decision making. The second objection centers around the St. Petersburg paradox and relates more significantly to mathematical oddities that stem from the expected value calculation. I conclude that in both cases, the use of expected value as a means of evaluating the *ex ante* welfare payoffs of decisions leads to counterintuitive decision making. I conclude that these examples provide sufficient grounds to reject the use of expected value alone, and propose the additional consideration of the variance of payoffs as an alternative method of assessing the *ex ante* value of decisions that still has the initial intuitive appeal of expected value, while also avoiding its failures in the examples I provide. Additionally, I explore the possibility that considering variance in addition to expected value also resolves the utility monster objection by reducing the appeal of certain decisions in which according to expected value alone, Longtermism would prioritize the future. Unfortunately I conclude that while some decisions do appear less appealing when taking variance into account, considering variance in addition to expected value only assists with a limited range of scenarios, and does nothing to ameliorate those scenarios in which payoffs have low-variance *and* a high expected value for the future.

Expected value has a straightforward mathematical form. For the purpose of estimating utility, it can be thought of as a sum of the utility resulting from all possible outcomes of a course of action, weighted by the probability of the given outcome occurring. The nature of probability introduces an asymmetry, of sorts. Probability is constrained between the values of zero and one.

Value, by contrast, is not. It can be infinitely large, or even infinitely negative. This asymmetry can result in seemingly odd decision making. One might forego a near-certain payoff of reasonable size in favor of an astronomical payoff that will almost certainly not payoff. This is the seemingly "fanatical" decision making that Greaves and MacAskill worry about in making the case for Longtermism. Beckstead and Thomas offer a particularly devilish case illustrating the intuitions involved in fanaticism.

The devil can strike quite the deal. Imagine that John is on his deathbed, slowly approaching death. God materializes before him in an ethereal glow. He offers him a card, redeemable for an additional happy year of life. Not to be outdone, the devil appears in a cloud of fiery, yet intriguing, brimstone. He offers John an alternative; rather than a guaranteed year, John could have a 99.9% chance at 10 happy years, and a 0.1% chance of nothing. John, reasonably, accepts the offer. The devil then ups the stakes. He offers John a new deal; a 99.9^{2}% chance of *100* happy years, otherwise, he receives nothing. John also accepts this deal – it still seems like an incredible tradeoff. Thousands of similar deals later, John ends up accepting a deal for a $99.9^{50,000}$% (a bit less than one in 10^{21}) chance of 10^{50000} years of happy life. As one might guess, John's gamble does not payoff, and he dies (Beckstead and Thomas).

John's seemingly preventable demise presents a curious dilemma. 1 in 10^{21} are impressively terrible odds. John might have better luck raiding a medicine cabinet and hoping for the best. Yet via the series of deal comparisons, the devil convinces John to take his deal. Nick Beckstead and Teruji Thomas offer this dilemma to demonstrate seemingly fanatical decision-making on John's part. John is fanatical because he ends up choosing an infinitesimally small chance at a ridiculous amount of life, over a guaranteed additional decade, which would reasonably satisfy most people (Beckstead and Thomas). Beckstead and Thomas characterize

three potential ways to resolve John's apparent fanaticism: timidity, recklessness, and non-transitivity.

Timidity is a potential way to avoid John's fanatical decision making. At some point in the series of pairwise comparisons, John decides that any increase in reward cannot outweigh a marginal decrease in the probability of such a reward occurring. In essence, John decides that some level of probability is simply too low, no matter the reward. Beckstead and Thomas offer a more rigorous definition of timidity (T):

> (T) By any possible standard of closeness, there's a finite payoff *x*, and close-together, positive probabilities $p > q$, such that for every finite payoff *y*, no matter how high, getting *x* with the slightly higher probability *p* is no worse than getting *y* with the slightly lower probability *q* (Beckstead and Thomas).

In short, the decrease in the probability of *y's* occurrence as a result of moving from probability *p* to *q* outweighs any increase in payoff to be had by receiving *y* rather than *x*. Recklessness, by contrast, is fanatical decision making – no level of probability is too low, so long as the reward is big enough. Beckstead and Thomas also formalize recklessness (R):

> (R) For any finite payoff *x*, no matter how good, and any positive probability *p*, no matter how tiny, there's a finite payoff *y*, such that getting *y* with probability *p* is better than getting *x* for sure (Beckstead and Thomas).

I will use fanatical and reckless interchangeably, as they refer to the same decision-making process. Finally, one can avoid the dilemma altogether by rejecting transitivity between the deals. Beckstead and Thomas offer the following definition of non-transitivity:

> (NT) There are prospects *A*, *B*, and *C*, such that *A* is better *B*, *B* is better than *C*, but *A* is not better than *C* (Beckstead and Thomas).

Recklessness is concerning because of its implications for Longtermism. If individuals ought to act recklessly, then they might make seemingly outlandish decisions. If faced with a choice of either quintupling the number of future humans (recall, 10^{24} is the reasonable estimate) with a probability of 0.0001%, and saving 1 billion lives today with certainty, the reckless individual is morally obligated to try to quintuple future humanity. However, in all likelihood, 1 billion people will die for no gain whatsoever. So, reckless behavior should be avoided. Beckstead and Thomas argue that this can be done either by adopting timidity or rejecting transitivity. Timidity, however, involves seemingly significant costs.

Timidity might plausibly arise from multiple sources. If there were an upper limit on the quantity of utility an individual could experience, then it would be reasonable to believe that some arbitrarily large increase in reward might not be worth any decrease in the probability of its occurrence. If the increase were to exceed the maximum utility one could experience, it wouldn't be worth anything. Another potential form is Nicolausian discounting. Nicolausian discounting involves treating probabilities below a certain threshold as zero; if something is extremely unlikely, we ought not consider is for the purpose of decision making (Beckstead and Thomas). Beckstead and Thomas also describe tail discounting. Tail discounting is similar to Nicolausian discounting, except that values are discounted based on their extremity, rather than their probability of occurrence. This allows for the grouping of extreme values into a cumulative probability, to avoid cases of redescription that make Nicolausian discounting unpalatable to some (Beckstead and Thomas). In essence, tail discounting remains true to the general notion that, if the probability of something occurring is quite small, we ought to ignore it.

Payoff discounting can lead to counterintuitive choices. Choosing an arbitrary probability value, under which we are unconcerned with any changes in payoff, can be odd. By ignoring all

payoffs below a certain probability threshold, we are in effect asserting that *no* increase in payoff can possibly entice us to take the deal. Similarly, we are also asserting that even the *slightest* decrease in probability below the threshold might force us to reject a deal, even decreases that might be so little as to be meaningless to the non-timid. Consider an extremely timid individual. For whatever reason, their threshold of timidity is extremely high. They are offered a deal whose probability of failure (i.e., a zero payoff), is one in one trillion, and the positive payoff is $1,000. This deal happens to sit right at their threshold of timidity; if the probability of failure were any higher, they would reject the deal in favor of any certain positive payoff. Even if they were offered $1,000,000, with a two in one trillion probability of failure, they would not accept it. The cost of timidity, in such a case, seems quite high.

Timidity's cost might be attributable to extreme stipulation or avoided by bounded utility. The cost worry persists, however, despite such attempts to avoid it. Perhaps the cost of timidity seems quite high because I have simply offered an extremely timid person. It is not the nature of their timidity that offends intuition, but rather its extremity. If we were to examine a "reasonably" timid person, then timidity might not appear so extreme. Perhaps our timid individual faces a choice of either certain receipt of an additional year of life, or a 1 in 1000 chance at 10000 years of life. 1 in 1000 is their timidity threshold; any lower, and they will reject the deal. Even if the devil were to offer one *million* years, albeit with the probability of one in 1001, this deal would be too risky, and rejected as a result. This timidity does seem somewhat strange. However, rejecting it leads one back down the slippery slope of pairwise comparisons. Adopting an upper bound for utility could also justify timidity. However, there are also ways to circumvent such a bound. Consider a potential bound on the number of enjoyable lives a person can experience. After 10^{10} years, life loses most of its zest; one has been everywhere, done

everything, etc. and so forth; living another year just wouldn't do much for them. Therefore, if the devil were to offer someone 10^{20} years of life, with a slightly lower probability than receiving 10^{10} years, they would not accept the trade. They would appear timid; no increase in the potential utility of their trade would justify a decrease the payoff probability. This seems somewhat reasonable; diminishing marginal utility is a widely accepted phenomenon that would result in such a scenario. However, let's say the devil has anticipated the boredom of longevity. Instead of offering so many more lives for an individual, he simply promises to create as many years of happiness as one would have received otherwise, spread out among many individuals such that the marginal utility of each year is maximized. It seems that more people living longer, happier lives is insulated from the effects of diminished marginal utility. More is always better by the same amount. Timidity would then still be seen as an extremely costly disposition, in terms of happy lives forgone.

Discussion of timidity thus far has maintained a detached air. It feels almost as though one is playing a game of linguistic whack-a-mole with the devil. First, we decide at some point, that enough is enough, and we will not accept any lower probability of success. But then, he ups the stakes until our decision theory begs us to take his deal; it is simply too good to pass up. We still want to avoid taking a ludicrous deal, so we set a bound on our utility. We slam our hammer down, arguing that at a certain point, we simply cannot get any happier. The devil pops up from another hole, seemingly unscathed. He then offers to extend his generosity to everyone, removing the potential bound on enjoyable utility. While these attempts to quash the devil's temptation are reasonable, they have left us no better in terms of avoiding his deals. Yet we ought not quit the table; however significant the surface level costs of timidity appear, they are not quite as extreme as the fanaticism that constitutes the alternative.

Reckless decision making also harbors significant costs. John's deal with the devil leads him to sacrifice his guaranteed additional year of life. Perhaps it could have been an extremely happy one. Recklessness might generally lead us to abandon pursuits that have significant positive utility with great certainty, in favor of those that have spectacular payoffs. Consider the potential tradeoffs inherent in Longtermism. One might be faced with the choice of spending money on either asteroid deflection, or saving 1000 lives from near-certain, painful death (perhaps inhabitants of a malaria riddled area, in which one is near certain to catch the disease). Reckless decision making might prompt one to choose to invest in deflection, regardless of how small the probability is, simply because avoiding potential extinction could save an incredible number of future lives. This is indicative of a more fundamental distinction between timidity and recklessness.

Timidity seems to be a product of excessive concern with regards to the probability of a given outcome, while recklessness appears to be a concern with the payoff. This binary characterization arises from an expected value conception of utility. By considering either recklessness or timidity, we are taking one component of the expected value equation to its mathematical extreme. We are varying the potential value of an outcome by one of the two fundamental characteristics we have conferred upon it. It is therefore worth considering whether the counterintuitive implications of timidity and recklessness stem from the expected value definition, rather than a more fundamental relationship between uncertainty and decision-making.

Neither timidity nor recklessness come across as decisively intuitive. It seems reasonable that at some point it is not worth pursuing additional payoff because of the unlikelihood of its occurrence. But it also seems reasonable that we might want to sacrifice some certainty in

service of additional reward. Yet these can't be the only dimensions pertinent to decision making. Expected value weights each value by the probability of its occurrence. If one is maximizing expected value, then it holds that they ought to choose courses of action that maximize expected value. But maximizing expected value results in recklessness and avoiding such behavior results in timidity.

The St. Petersburg Paradox offers another outlandish example that demonstrates the counterintuitive nature of using expected value as a method of evaluating the *ex ante* welfare payoff of decisions. Recall Peter's deal with Jerome Powell. It is worth reframing the deal in terms of additional years of happy life rather than dollars. Let's say that instead of Powell, Peter runs into the devil. The devil offers him a similar deal; he will flip a coin until the coin reads tails. He will then grant peter 2^n years of additional happy life (or endow so many happy lives spread out over some mass of humanity), where n in the number of heads that the devil observes before reading tails. This game shares the structure of the St. Petersburg Paradox (which is only a paradox in the colloquial sense), as described by Martin Peterson in the Stanford Encyclopedia of Philosophy, and originally described by Nicolaus Bernoulli (Peterson). Because the probability of observing an additional heads is 0.5, and the payoff doubles with each additional heads, the expected value of the game is infinite, as each number of heads adds one to the total expected value, and we might observe infinite heads in a row.

The St. Petersburg Paradox presents a curious dilemma for the purpose of evaluating the expected value of a decision. How should Peter value the devil's game? Typically, when assessing the value of random games, one turns to the expected value. But that seems outlandish in this case: the expected value of the game is infinite. It is outlandish for two reasons. The first is that while the expected value of the game is infinite, in all likelihood the average player will

walk away with only a few extra years of life: perhaps 32 additional years in exceptional cases, but rarely more. The magnitude of infinity hardly seems to capture the fact that most of the time, the game will not be worth many years at all. Half the time it won't be worth any additional years (the first flip being a tails). Yet expected value still judges the game to be worth infinite additional lives. The game also reveals a sort of logical barrier to assessing its value using the expected value equation.

Peter's ability to value the devil's game using expected value falls prey to infinity. Even if one were to concede that supposedly "infinite" utility is possible (i.e., an infinite number of future years of additional life), then we might simply be concerned that Peter will die before the coin ever reaches termination. In fact, strictly speaking, because the payoff is only ever meted out after the devil sees a tail, one could never actually enjoy an infinite payoff. The devil would either be flipping the coin forever, or upon the cessation of flipping, pay out a finite payoff, which is necessarily lesser in magnitude than infinity.

The St. Petersburg Paradox, while not necessarily directly related to Longtermism, still has important implications for the use of expected value as a method of *ex ante* assessment. The presumption that expected value is a reasonable way to evaluate uncertain decisions is fairly widespread. The St. Petersburg Paradox, and more specifically, Peter's game, demonstrates that there are significant problems with relying on expected value as a metric of the value of a decision. Not only does the expected value of the game *necessarily* overstate the payoff that we could ever achieve from it, but it also misrepresents it at the achievable level. It would be one thing if infinite value happened to represent something that was effectively infinite. It is entirely another to realize that a game that will typically pay off anywhere from 0-8 additional years of

happy life characterized as one that pays out an infinite number of years. With such a difference in mind, it is worth revisiting *why* agents are maximizing expected value in the first place.

Expected value is a method agents use to evaluate utility in the face of uncertainty. However, revisiting the fundamental aim of utilitarian agents helps demonstrate how expected value in particular obscures their original aim. Utilitarian agents ought to maximize utility. Faced with the uncertainty of the world, it is sometimes unclear how to do so. So, agents within the context of any singular decision, are attempting to determine how to maximize the utility of that *particular* decision. This is subtly distinct from the strict meaning of expected value.

Mathematical definitions of expected value refer to it as an aggregation of the potential outcomes of some system. An uncertain system might never output its expected value; if the potential outcomes of a game are either one or zero, the expected value of the game, presuming non-zero probabilities for both outcomes, will never be either one or zero. It will be something in between. But the agent facing such a game will only ever experience an outcome of either zero or one. So, perhaps agents should focus on what will be the *most likely* outcome of a decision and proceed from such a standpoint.

Considering the most likely outcome in place of the expected value of a system is a step towards avoiding some of the counter-intuitive implications of expected-value reasoning. One might reasonably use a form of tail-discounting, where one ignores particularly extreme possibilities, and then chooses based on the most likely outcome of the system. Consider Alice, who is deciding between two possible games to play. One involves a 75% chance of receiving $1000, and a 25% chance of receiving nothing. The other involves a 0.1% chance of receiving $760000, and a 99.99% chance of receiving nothing. Alice will only play the game she selects once. The expected value of the first game is then $750, while that of the second is $760. If Alice

were making her decisions based on expected value, then she ought to choose the second game, where she will, in all likelihood, earn nothing. It seems at least plausible that Alice ought to play the first game; it is very likely that she will win $1000, and she only gains a marginal increase in expected value by playing the second game, at the very high cost of most likely receiving nothing.

The intuitive appeal of expected value is rooted in part in the nature of repeated games. Alice is faced with the choice between the two games. Given that she can only play the game once, perhaps choosing the game with a more certain payoff is a more rational course of action. Expected value is the value we might expected to observe after many iterations of a game. However, when faced with a one-off decision, the expected value will differ wildly from the actual outcome. Alice will likely receive nothing if she chooses the second game. Given the difference between the value likely to occur, and the expected value of the decision, it makes less sense that Alice would rely on expected value as a guide for her decision-making.

The difference between expected value and the likely outcome of decisions matters significantly for the distinction between reckless and timid decision-making. Recklessness arises as a result of using expected value to weigh decisions. One would never accept the devil's ludicrous deals if one were weighing decisions by their most likely outcomes. Similarly, timidity appears far more rational. If one recognizes that the likelihood of receiving a spectacular payoff is impossibly low, then increases in the payoff carry little meaning. Furthermore, turning down new deals that offer greater payoff in return for lower uncertainty cease to be irrational.

Unfortunately, while focusing on the most likely outcome seems intuitively appealing, it merely constitutes a form of timidity. By focusing on the most likely outcome, we are in effect giving excess weight to the probability of an outcome at the cost of considering the potential

value of the outcome. If Alice were to decide to play the first game rather than the second, it is because the payoff of the second fails to outweigh the diminished likelihood of its occurrence. For the reasoning to hold, we would have to stipulate that any increase in the payoff of the lesser likely option would not be worth playing game two. Perhaps a 0.1% chance at $750000 isn't worth $1000, but a 0.1% chance at lifetime financial security for oneself and one's children is. It might be so even considering that Alice can only play the game once; even a one-off, slight chance at lifetime financial security is could potentially be worth more than $1000, although this seems far less likely. It also seems that this can be justified in non-expected-value terms. A 0.1% chance at never needing to worry about money seems valuable in its own right, not simply because it would be worth *x* dollars over many iterations. So, making decisions based on the most likely outcome will most likely result in the costs of timidity.

Picking the best, most likely outcome will likely have similar consequences to timidity across an individual's set of decisions. In any singular decision, picking the best, most likely outcome can be rationally justified. However, acting in such a manner over many decisions results in the expected value timidity that appears to be irrational. While expected value may misrepresent the value of any one specific game, it still appropriately represents the expected value of a timid decision-making process over many decisions. If, generally speaking, one consistently chooses the less-risky option, then one will miss out on potential value over time by not choosing risky options that have a chance of succeeding. Playing the St. Petersburg Paradox once is typically not a lucrative affair. But if one were to play thousands, or even millions of times, they would likely do quite well at some point. In this way, what seems to be irrational at the level of individual decisions might seem deeply irrational as a *system* of decision making.

The seemingly irrational nature of recklessness might simply be a consequence of how one chooses the set of decisions to be evaluated. If the set is restricted to a single decision, then expected value will potentially lead to fanatical decision-making. However, if one expands the set of decisions to all agents who may make the decision, or rather the set of decisions that one makes over the course of their life, then expected value seems more reasonable. Consider again John's deal with the devil. Would it still be irrational to pick the devil's extremely unlikely deal if one were to view it from the perspective of an arbitrarily large population? Perhaps every human on their deathbed is offered such a deal. For that matter, presume that every human that has ever lived and that ever will live, has been or will be offered such a deal. Eventually, someone will probably hit it. And when they do, their corresponding increase in utility ought to make up for the untold lost singular years of life foregone in the pursuit of such a result.

To a certain extent, reframing the lens in a societal manner lends itself to discussions of morality. Morality is agent neutral; it concerns what anyone ought to do. For the purpose of most types of utilitarianism, this means that any decision that will in aggregate increase utility is the right decision. Expected value is a framework for quantifying uncertainty that functions best when considering many, many actions. Because morality is the criterion by which agents judge the worth of all competing courses of action, reasoning by means of expected value will result in the application of such reasoning over many decisions.

Redeeming expected value within the context of moral decision-making redefines what might be considered fanatical decision-making. The criterion of fanaticism is no longer whether the probability of some payoff is extremely small, but rather whether one will be able to play the game long enough to achieve said payoff. The devil's ultimate deal is a poor choice for John. But from the perspective of morality, if everyone is offered this deal, we all ought to choose it.

Shifting the criterion of fanaticism, however, might simply kick the proverbial can down the fanatical road. If one allows that sacrificing a certain gain for a greater extremely unlikely gain is reasonable given a requisite number of iterations, then how does one decide this number of iterations? Perhaps we have merely introduced new forms of timidity and recklessness.

The devil, as we have established, is a wily fellow. Let's presume that instead of taking the devil's offers to the extreme, John instead draws the line at 80%. He has reasoned through the probabilities involved and concluded that he would much rather have more years of life with reasonable uncertainty, seeing as this is a one-off gamble. Once he becomes worried that he will receive *no* payoff rather than any payoff, he declines. The devil deduces his reasoning and sweetens the deal. Not only will he offer this deal to John, but he will also offer it to *x* number of other individuals on their death beds. These individuals will be in the same situation as John; they will be deciding between a certain year of life and some additional years of life with the same uncertainty. They will also be informed that *x* other people have been offered a similar choice. Suppose further that everyone offered the deal is familiar with probability and expected value. What *x* should rationally prompt John to take the deal?

Determining whether accepting the deal is a rational action now rests on how many iterations of the deal will occur. If one is confident that one million people will be offered the chance to a game with a 1 in 100 chance of a $100 payoff, or receive a fifty-cent reward with certainty, the moral choice is to play the game. The expected value of everyone playing the game is $1 million, while if everyone simply took the reward, the value would be half as much. However, if one were confident that they would be the only person to play the game, they should probably take the fifty cents. In other words, one ought to be concerned not only with the expected value of a decision, but additionally the probability with which the decision *will ever*

occur. If John were told that 40 other people were offered a 79% chance at 10 years of life, or a year for certain, the probability of that deal yielding equal or greater years of life when taken 40 times is 99.9971%. So, if everyone takes the deal, it is near certain that there will be a greater number of years than if everyone were to take the guaranteed year. The numbers are even more spectacular if we consider John's original deal.

John was originally presumed to be fanatic for giving up a certain additional year of life for 1 in 10^{21} chance at 10^{50000} years of happy life. However, let's presume that every person for the rest of humanity's history is offered such a choice on their deathbed. If we proceed with the reasonable estimate of 10^{24} future humans, then we arrive at a very different picture. There is approximately a 74% chance that *at least* one person will receive the 10^{50000} years, for every 10^{21} who play the game. Given that *100 times* this many people play the game, we might expect to receive a fair number of wildly long-living individuals. Additionally, their cumulative increase in years of life lived fair outweighs the 10^{24} certain years of life forgone by those who have not won. Furthermore, the potential boundedness of utility need not worry us; we might simply stipulate that if the game is won, 1.25×10^{49998} individuals will live happy lives that are 80 years long. That outcome is pretty alright.

It might be objected that as a decision-making process, considering the number of iterations involved in a given scenario has no bearing on whether the individual decision is fanatic in nature. However, from the perspective of moral theory, one need not consider whether a decision is fanatic on the individual level; we need only concern ourselves with what *everyone* ought to do. If everyone throwing away their guaranteed additional years of life results in a far greater number of happy lives lived, then we certainly ought to do it. There are a series of concerns, however, with adopting such a view of moral decision making.

It is possible that considering the number of iterations merely changes the form of recklessness and timidity. What is the minimum probability that we ought to accept? So far, I have proceeded by considering the probability that choosing the uncertain game results in a net increase in utility when compared with the certain payoff. However, it is unclear what threshold this probability should be set at, for the purpose of determining the moral choice within a set of actions. Furthermore, I have limited the potential pertinent information by merely considering *whether* an action might potentially increase utility. It might also be reasonable to attempt to consider the magnitude of increase. However, it still seems reasonable to allow that the probability of whether a payoff occurs within a given system is important for the purpose of the fanatic intuition.

Considering the probability of *whether* a payoff will occur, given a certain number of iterations, seems oddly similar to considering the expected value of a decision in the first place. It is possible that we have merely reformulated the notion of expected value; instead of focusing on the probability that an event occurs for any one person, we have substituted the probability that it occurs for *at least one person*. This is still assessing an expected value, of sorts. However, instead of altering the criterion of decision making, accounting for the number of possible iterations of some game is really examining a different system entirely. Perhaps the intuitions surrounding fanaticism stem from an issue of perspective, not from expected value. The expected value of one person taking a gamble is just that – an expected value. But if we consider that *everyone* might be presented with such a gamble, as we might in the case of morality, then we ought to consider the overall distribution of the underlying system, rather than its expected value alone. If some course of action *A* has a highly variable distribution of payoffs, then perhaps we

ought not take it over a comparable course of action *B* with a comparable expected value, but lower underlying variance.

It should be acknowledged that appeals to variance have little to offer for those who are comfortable with the nature of expected value. If one is comfortable with the implication of fanatic decision-making, then there is no reason to appeal to variance. Importantly, such a view eliminates a potential avenue of inquiry for rejecting Longtermism; we cannot appeal to potential deficits in our ability to assess the future if we stipulate that such a method is reasonable. If we accept expected value then the monster remains, and we must continue our search.

So, what might this mean for Longtermism? If we conclude that assessing the likelihood of payoffs occurring for *at least one* person, then we will have in effect abandoned expected value. Does this change avoid swamping the present via the future? Unfortunately, although it provides a more reasonable justification for previously supposed fanatical decision making, it does not appear to free the present from the future.

Reformulating examples offered in support of Longtermism illustrates how considering variance in conjunction with expected value fails to reduce the mathematical influence of the future on the decision-making process. Consider asteroid deflection. Asteroid deflection presents a burden on the present in the form of upfront costs for asteroid tracking, deflection, etc. Its expected value, however, is "astronomically" high. If asteroid deflection prevents extinction, then the payoff is as large as any population that would have otherwise never existed. The distribution of payoffs, however, is highly skewed. In many possible worlds, asteroid deflection will have a payoff of zero; the chance of an extinction level event is vanishingly small. Furthermore, it has little value outside of the extremes: the chances of a non-extinction level asteroid collision that is still worth deflecting are also quite small. The variance of the welfare

payoff of asteroid deflection, then, is quite high. Despite this high variance, which we might hope to serve as a barrier to fanatic decision making, we would probably still invest in asteroid deflection at the expense of the present.

There is great difficulty in pinning down the intuition that undergirds disgust with servitude. Asteroid deflection characterizes some of these difficulties. One worry is that such discussions simply degrade to a stipulative battle. If we merely stipulate that the benefit to the future is far greater than the benefit to the present, then of course we are obligated to serve the future; to do otherwise would be a rebuke of the principle of utility. A more careful assessment of courses of action, however, would suggest that passing judgement on expected value is more difficult than some would make it seem. Perhaps we can spend some amount of money on asteroid deflection. Even if there is no asteroid immediately bearing down on earth, we could acquire the capability to deter one if there were. Alternatively, we could spend the same dollars on purchasing bug nets for those at risk for malaria today. A cursory consideration of expected value favors investing in deflection over bug nets; untold billions might benefit from their assured protection from asteroid extinction, while only thousands might benefit from the same investment in bug nets. However, temporal considerations muddy the waters somewhat. Why should we invest in asteroid deflection if we know that such an event is not imminent? Could we not, reasonably, delay such a decision to when a collision is imminent, and then perhaps start spending on deflection? These are difficult empirical questions to answer, which is why we appeal to expected value as a metric of decision value; it captures the uncertainty inherent to such investigations. However, by arguing that probability can account for these difficulties, we have in effect again resorted to stipulation.

That comparisons of expected value rely so heavily on stipulation suggests something more fundamental about the moral relationship between the past and the future. Even if we were to remove any sort of considerations related to how we evaluate future actions, and simply stipulate that the future would benefit significantly from the sacrifice of the present, we would have a similar problem. We might simply argue that many millions of people, far in the future, will enjoy great benefit from some arbitrary action that we could take in the present. This action would result in a sacrifice of utility on our part. But the future benefit will far outweigh any present cost. The future would still represent a utility monster, independent of the mechanism that Longtermism uses to evaluate the payoff of various decisions.

If we cannot appeal to the mechanism by which we value decisions, then we must find another way to avoid the enslavement of the present. One plausible avenue is the simple devaluation of the future. We might concede that certain costly actions in the present could have spectacular benefits for the future. However, because these benefits occur in the future, for some reason yet to be found, they ought to be valued less than benefits that occur today. We might then leverage such a discount to reassert the moral worth of the present.

IV

Tomorrow Still Matters

The combined mass of future humanity poses a problem for Utilitarianism, and by extension, EA. If the concerns of the future will always outweigh those of the present, then present humanity might find itself constantly serving the future, to its own detriment. This predicament strongly resembles the Utility Monster mentioned by Nozick. I have already

attempted to avoid this objection by arguing that the future's apparent dominance of present-day utility calculations stemmed from the use of expected value to evaluate the *ex ante* welfare payoff of courses of action. Unfortunately, this did not provide the certainty that avoiding the future utility monster requires. I will now consider how to justify a temporal discount rate. Such a rate, if reasonable grounds were found to support it, would allow the present to devalue the far-future with ease, and in doing so reassert its own import. Derek Parfit discusses temporal discounting at length in *Reasons and Persons*. I consider here a series of his arguments to determine whether adopting a temporal discount rate is feasible.

Parfit examines a series of possible arguments in favor of temporal discounting (which he refers to as "Social discounting"). He judges them all to fail. However, they also offer reasonable promise to avenues of caring less about the future, even if not by means of discounting. The most relevant arguments relate to the probability of events, opportunity costs, the greater welfare of future generations, and worries of excessive sacrifice.

Parfit considers that the probability of an event's occurrence might be grounds for temporal discounting. He considers the following: "it is often claimed that we should discount more remote effects because they are less likely to occur" (Parfit 481). He offers an example that mirrors some of those offered by Longtermists to demonstrate the gap between potential costs to the present, and massive benefits to the future. The particular example is unimportant. More pertinent to our considerations is his assessment of the more general facets of probability-related discounting:

> We ought to discount those predictions that are more likely to be false. Call this a *Probabilistic Discount Rate.* Predictions about the further future are more likely to be false. So the two kinds of discount rate, *Temporal* and *Probabilistic,* roughly correlate.

> But they are quite different. It is therefore a mistake to discount for time rather than for probability…If we discount for time rather than probability, we may thus be led to what, even on our own assumptions, are the wrong conclusions (Parfit 482).

In short, temporal and probabilistic discounting might overlap significantly, but they are not equivalent. If we look to justify temporal discounting on probabilistic grounds, we might be led astray in cases in which events in the future actually are quite certain, but by discounting them on temporal grounds, we give these future events an inappropriately low weighting in our moral calculation. Additionally, probabilistic discounting is quite familiar to Longtermists.

The Longtermists already discount on probabilistic grounds. Expected value weights outcomes based on their probability of occurrence, in effect, discounting events that have a lower likelihood of occurrence. Therefore, even if probabilistic discounting were a reasonable justification for temporal discounting, Longtermists could not appeal to it as salvation from servitude; we have already established that such servitude occurs in spite of probabilistic discounting. If temporal discounting is to devalue the monster, we must find another justification. Parfit also considers the opportunity costs of taking a given action as a potential justification for temporal discounting.

Opportunity

cost is also sometimes offered as a justification for temporal discounting. Parfit illustrates this justification nicely:

> It is sometimes better to receive a benefit earlier, since this benefit can then be used to produce further benefits. If an investment yields a return next year, this return will be worth more than the same return after ten years, if the earlier return can be reinvested properly over these ten years. When we have added in the extra benefits from this reinvestment, the total sum of benefits will be greater. A similar argument covers certain kinds of cost. The delaying of benefits thus involves *opportunity costs*, and vice versa.

There are costs to receiving benefits now, rather than later. Perhaps we ought to discount temporally to account for these costs. Parfit argues that this argument fails due to similar reasons as probabilistic justifications do for temporal discounting. Rather than discount temporally, we might rather simply add the cumulative opportunity costs to whatever assessment we conduct. Additionally, if we were to discount temporally, rather than explicitly on the basis of opportunity cost, we might arrive at improper conclusions. For opportunity cost as well as probabilistic discounting, temporal discounting represents a hammer, while either of the two more specific justifications are more of a scalpel. As moral operators, we ought to prefer the latter.

We cannot appeal to opportunity cost to justify temporal discounting. However, considering opportunity cost itself might offer a reasonable avenue of respite from the demands of the future. At first glance, it is opportunity cost that embodies much of the discomfort with the utility monster to begin with; the opportunity cost of serving the monster is our own present well-being. Perhaps the opportunity cost of investing in asteroid deflection technology is saving children in Africa from malaria by providing bug nets.

Unfortunately, it seems as though considering opportunity cost does little to shift the relative expected values of different courses of action. If opportunity cost were to provide a way out, it would need to somehow make courses of action that benefit the present have a greater expected value than those that benefit the future. However, opportunity cost suffers from a symmetry that prevents this; any course of action is an opportunity cost to another course of action. By making any one choice, we must necessarily sacrifice another. Consider two decisions, A and B. Decision A results in 100 total utils in the far future, while action B results in 99 utils in the present. Both events occur with certainty. If we consider the opportunity cost to detract from the value of a decision, then the expected value of decision A is 1. The expected value of B, by contrast, is -1; we have sacrificed a world in which we might have had 100 utils to gain 99. If opportunity cost does not give us grounds to devalue the future, then we need to continue our search. The wealth of future generations might provide one possible justification.

The future often represents an improvement on the present. If the future does exceptionally well, then perhaps we are justified in prioritizing our own utility somewhat. Perhaps this is a reason to discount temporally. Parfit considers this possibility, enumerating it within two possible scenarios:

> If we assume that our successors will be better off than we are now, there are two plausible arguments for discounting the benefits and costs that we give to and impose on them. If we measure the benefits and costs in monetary terms, adjusted for future inflation, we can appeal to the diminishing marginal utility of money. The same increase in wealth generally brings a smaller benefit to those who are better off. We may also appeal to a distributive principle. An equally great benefit given to those who are better off may be claimed to be morally less important (Parfit 484).

He argues that both of these lines of reasoning fail for similar aforementioned reasons. The first and foremost is that, for the purpose of justifying temporal discounting, we are again adopting a proxy in its place. Adjusting for diminishing marginal utility or appealing to distributive principles might be fair reasons to devalue future utility, but temporal discounting is more general, and will therefore lead us to discount in cases we ought not to. However, for our purposes, these might be reasonable methods of devaluing the future.

Appealing to the diminishing nature of marginal utility might offer a reason to prioritize the present over the future. This aspect of utility forms a significant component of the Effective Altruist argument for assisting the developing world; because doing good for those with little is comparatively cheap, we ought to direct our resources to those causes that require the least dollars per some standard increase in utility. If we consider temporal distance, rather than spatial distance, we might arrive at a similar relationship between the present and the future. In comparison with the future, we might enjoy relative levels of utility similar to the gap between the developed and developing world. We therefore might be justified in keeping some resources to help the present, rather than simply pouring everything into the future.

Unfortunately appealing to the wealth of the future as justification for prioritizing the present can only do so much. It is, at the end of the day, an empirical argument. It is largely not distinct from simply maximizing total utility; we have simply stipulated in this case that perhaps the present deserves more attention due to the diminishing marginal utility the future might face. Furthermore, this also fails in light of the future's massive number of people. Perhaps if the future is filled with a small number of wealthy people, the difference in marginal utility might allow us to favor the present. However, if the future is filled with *many more* people living lives of comparable personal happiness, then there is no reason to presume that the future will enjoy

any lesser marginal utility than the present. Additionally, because of its size, it will then still swamp the present.

Parfit also considers the possibility that perhaps the excessive demands of the future present grounds for discounting future utility. This is of course, the very concern of this thesis. He outlines The Argument from Excessive Sacrifice as follows:

> A typical statement runs: 'We clearly need a Discount Rate for theoretical reasons. Otherwise any small increase in benefits that extends far into the future might demand any amount of sacrifice in the present, because in time the benefits would outweigh the costs (Parfit 484).

He dismisses this argument for similar reasons as detailed above. If we were to discount in fear of the burden presented to the present, then we would be merely papering over another more fundamental principle; that the future ought not exert such moral sway over the present. This is a separate argument. Instead of appealing to discounting, we ought to let this more driving consideration inform our thinking.

Parfit's rebuttal of the argument from excessive sacrifice carries particular import for the investigation at hand. Not only does it fail to justify social discounting as a potential method of devaluing the future, but it also reveals more troubling aspects of our investigation. We have seen that it is at least quite difficult to justify a social discount rate. Often, we care more about other, more fundamental concerns that will not save us from servitude. It is unclear what further road remains for devaluing the future. We could simply stipulate that the future matters less. But this is hardly compelling, and for that matter, might conflict with the principle of utility. Seeing as we would rather preserve the principle of utility than reject the primacy of the future, we ought not stipulate its inferior worth.

At this point it is worth taking stock of where we have been. We have examined EA and Longtermism, and determined that some burgeoning mass of the future could play such a role in utilitarian calculations. We determined three potential routes of changing the relationship between the present and future: change how we assess expected utility, discount the future to a more significant extent, or potentially adopt average utilitarianism. Our assessment of utility turned out to matter less than originally believed; we can allay fears of fanaticism by considering variance, but it also happens that the future can lord over the present independent of how we assess expected utility. I then turned to a temporal discount rate, in the hopes that, for some reason, we might be able to simply value future utility less. But this too, has produced little. It turns out that in many instances in which we think we ought to discount temporally, we really are guided by another more specific moral principle. Unfortunately, these principles either provide no grounds for us to devalue the future, or they have already been accounted for in our calculation of utility, and therefore to discount temporally on their basis would in effect be double discounting. We arrive then, exactly where we began. The future still holds the present hostage.

This leaves me with the final option as detailed earlier. I will now investigate as to whether adopting average-utilitarianism in place of total-utilitarianism can help allay concerns of the future's dominance in moral considerations.

V

The Monster Lurks Today

Reconsidering the use of expected value failed to grant the present any respite from the future. So too did attempts at finding a temporal discounting rate. I will now consider the third route of avoiding the import of the future that I described earlier: appealing to average utilitarianism. I begin by considering Derek Parfit's Repugnant Conclusion and contrast the utility monster that he observes within the Repugnant Conclusion with that which I observe within the far-future. I then argue that utilitarians are not only vulnerable to a world in which the Repugnant Conclusion is optimal, but rather are obligated to bring about the Repugnant Conclusion regardless. I argue this based on observations regarding the nature of diminishing marginal utility and the utilitarian obligation to have children based on the utility that such additional life would engender. I also consider whether questions of identity influence the burden of the future and conclude that because utilitarianism concerns itself with utility alone, questions of identity have no effect on the import of the future. I then consider the utility monster inherent in the obligation to bringing about the Repugnant Conclusion. In response to the monster, I consider the benefits of adopting average utilitarianism, and evaluate Parfit's Mere Addition Paradox. I conclude that the Mere Addition Paradox renders average utilitarianism unacceptable and reject average utilitarianism as a mechanism for avoiding the moral obligation imposed by the size of the future.

Parfit realizes the problem posed by the burgeoning mass of future humanity. He offers a series of principles that lead to what he calls the *Repugnant Conclusion.* The most important of these is the Impersonal Total Principle (IP):

(IP) If other things are equal, the best outcome is the one in which there would be the greatest quantity of whatever makes life worth living (Parfit).

He then asks us to compare two worlds: world A, in which inhabitants enjoy a certain level of happiness, and world B, which is populated by twice as many individuals as world A, but who enjoy an average happiness that is greater than half that enjoyed in world A. In other worlds, there is, in total, more happiness in world A than in world B. If one accepts the Impersonal Total Principle, then one *must* prefer world B to world A; put in otherwise moral terms, world B is *better* than world A. However, it also seems that there might be a world with a still greater population than world B, whose population are living happy enough lives for one to deem this new world better than B. Parfit extends this reasoning to form the Repugnant Conclusion, which he states as follows:

(RC) For any possible population of at least ten billion, 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 (Parfit).

Parfit finds such a conclusion unpalatable, hence his declaration of its repugnance. He then couches the repugnant masses in terms of a Nozickian utility monster.

The combined mass of lives barely worth living, might, for Parfit, constitute a utility monster. Recall Nozick's supposed monster: an individual that, by stipulation, enjoys far greater utility returns by consuming resources than any other individual. Society is then bent to the monster's will, bound by its moral obligation to maximize utility. Some Utilitarians respond to Nozick by simply resisting his example. Such an individual seems so far-fetched that one need not worry about it from the perspective of utilitarianism. If such a monster *were* to exist, they

might pose an issue; however, since they do not, they are of no concern. Parfit argues that the repugnant conclusion might offer a monster that resists this response (Parfit). While the population monster dominates by virtue of the quantity of lives lived, rather than their quality, as Nozick's monster does, it still dominates. Furthermore, such a monster is not inconceivable. We can very easily imagine a universe with more people in it. It is worth considering, however, what exactly "repugnant" means, and whether Parfit's conception of repugnancy influences whether we can conceive of this unborn mass as a utility monster.

Part of the Repugnant Conclusion's repugnancy stems from Parfit's phrasing. "Lives barely worth living" hardly sounds appealing. One could just have easily said "lives that are net positive." A still more appealing phrase might be "lives that are just happier than not." Either of these seem more appealing than "barely worth living." Perhaps one way to conceive of the life barely worth living is to engage in a systematic sculpting of one's existence, in search of a bare minimum beyond which life is not worth living. In some ways, this represents a pursuit of philosophy that has found no conclusion in millennia of discussion. It seems quite possible that everyone might be able to pursue an ascetic life of deprivation that is in some sense, still worth living. Perhaps this isn't a bad thing. Monks are not notoriously unhappy. In some ways, they are regarded as happier than most. Monks also shed light on a potential route of empirical salvation for the Repugnant Conclusion; if we can all be happier living as monks and support more monk-living people as a result than a world in which lives are barely worth living, this might be a case in which a world with more people is in fact happier both on average, and in total terms, than a world described by Parfit. A more careful consideration demonstrates this notion clearly.

If happiness can be derived from sources other than material resources, we might reasonably imagine a world in which more people live slightly less happy lives, but this is not a repugnant difference. Consider a world of monks:

> (*Monk World*) Monks populate the earth. As a society, the monks have determined the absolute minimum quantity of resources required to sustain a life; they cannot bring more people into the world by sacrificing more worldly comforts. Yet, these monks are somewhat happy. It would be nice to have a hot shower here or there, or perhaps warm gruel instead of cold, but they try not to focus on that. Instead, they cultivate relationships with others, a cheap source of happiness (presumably they are all fairly social, and at least somewhat agreeable), one which is able to just barely overcome the physical misery they endure on a daily basis.

This world doesn't necessarily seem quite as repugnant. Happiness found in non-material goods might represent a way to concede that Parfit is somewhat right about the Repugnant Conclusion without accepting its status as repugnant; it is instead, a world full of happy people. However, the monks are also vulnerable to familiar objections. Their society is a wildly fantastical, and perhaps cripplingly unrealistic one. The monks also represent a fairly narrow avenue out of servitude; in any world without monks the deeper worries of repugnancy and the sheer mass of the future still hold weight.

Discussion of future lives, thus far, has presumed that futures with more lives in them are better ones. However, this is not necessarily true. It might be the case that some lives are better left unborn – those who are guaranteed to suffer from a terrible illness might be plausible candidates. The Repugnant Conclusion does not include such people – presumably such lives would not be worth living at all. More interesting is whether families face a moral obligation to

reproduce. Parfit offers the "happy child" as a thought experiment to examine such this claim. He asks us to consider a couple that is fairly well off and considering having another child. This child would almost surely live a very comfortable life, and the parents would be made happier by having the child. However, because childbirth is intensive, and because of the strenuous obligations of raising a child, there would certainly be significant drawbacks to having the happy child. In short, the couple is essentially indifferent; they would be perfectly happy having the child, or not. The total utility of the world, however, would increase; the parents would experience no decline in utility, and the addition of the child would add the utility of their life to the universal total.

It seems that one could reasonably argue that the couple ought to have the child. All things equal, the couple would be perfectly happy, and there would be an additional happy life in the world. Parfit concerns himself largely with the question of whether there is a moral reason to give birth to an additional happy life; I.e., all things equal, a universe with more happy lives is a morally better one than one with fewer. In some ways, this almost seems trivially true; more, happy lives seems obviously better than fewer. It might be more interesting to consider a case in which the happy child causes even greater drawbacks to the parents. Perhaps instead of being merely indifferent, the parents *do not* wish to have the child. They value their careers quite highly, and a child would present a significant disruption to their career paths, if not outright scuttling them. However, the child would still live a happy life, and the potential parents are confident in this. Lets refer to this child instead as the *demanding child.* The utility calculation works out such that having the child represents an utterly neutral decision; the world is no better for their existence, and no worse without it.

Whether the potential parents ought to have the demanding child could be demonstrative of intuitions regarding the value of life for its own sake. If we could demonstrate that there is no value to life excepting the utility inherent to it, we might be able to avoid the obligation to reproduce on empirical grounds. If it were somehow possible to demonstrate that caring for the future would represent such significant disutility that not doing so could be a reasonable option, then the present might not be beholden to the future. However, if one were to conclude that an additional life has moral value outside of its total utility, then the chains of servitude might become impossibly difficult to break. Even if there were significant disutility to be had from giving birth to an additional child, the moral value of an additional life might offset such disutility and demand the child be born regardless.

Utilitarians don't have compelling grounds to give birth to the demanding child. At best, they could go either way. Some might find this dismissal of the value of life counter-intuitive, although properly addressing this intuition is beyond the scope of necessity here. Utilitarians are concerned with the principle of utility. Insofar as lives are to be had, they must result in some sort of positive utility; to usher into the world a life of net-negative utility would be a bad action. More interesting then, is the case in which the utility produced by having the demanding child is just barely positive.

If we stipulate that the utility generated by the demanding child is slightly net positive, then we arrive again at the Repugnant Conclusion. Giving birth to an additional life will increase utility, and therefore one ought to usher into existence as many "barely worth living" lives as one can muster. We ought to revisit, then, the Repugnant Conclusion, and more thoroughly examine the causes that drive it. We discussed a potential world of monks that might make the repugnant

component of the Conclusion more palatable. Yet Longtermists might even simply outright reject the notion that the Conclusion is repugnant, even in non-monk worlds.

What exactly constitutes a life barely worth living is important to the Repugnant Conclusion; it is the source of the Conclusion's supposed repugnancy. Parfit presumes that a life barely worth living is a terrible one. It is worse than a life with greater utility – this is definitional – but, put otherwise, it could be called a "happy" life. It is at the very least, a life that is happier than not. This might constitute a range of scenarios. Perhaps it is a Malthusian conception of man, in which one simply grows enough food to reproduce and dies, similar to our monks. Or it could be a life subject to exceptional highs and lows, twists and turns, victories and defeats, only to arrive slightly out ahead when all is said and done. This might be closer to ordinary life than not, which I am hesitant to refer to as repugnant. If the Repugnant Conclusion is not quite so repugnant, then perhaps the Longtermists have little to worry about. They might in fact argue that the Repugnant Conclusion might represent the ideal achievement of their aims. However, even if we set aside the notion of the Conclusion's repugnancy, we can establish that utilitarians might not merely prefer the Repugnant Conclusion to other seemingly better worlds, but that they are actually obligated to pursue its creation.

A fairly simple observation of the nature of utility demonstrates how we ought to make the future massive. Presumably, resources are subject to diminishing marginal utility; a loaf of bread is worth more to me if I have none, than when I have 100. If we generalize this relationship to all the resources that make life viable, we can observe a similar relationship; Any resources we enjoy strictly past that which is necessary for survival are, in a sense, less productive than the resources employed to get us to a basic level of survival. If resources are rivalrous - that is, only one person can consume them - then any consumption of resources

beyond that which is strictly necessary to survive might impinge upon the ability of another person to survive. Even if there is no other presently living person, one might be obligated to have another child, if everyone could cut back on their own consumption to support it. In short, so long as anyone consumes any resources beyond the level of sustenance, an implicit alternative course of action is available; someone could simply have another child, and this would *by definition* increase happiness, so long as this child's life is slightly net positive in terms of utility. This additional child would do so because any happiness enjoyed by the child is more productive from a resource consumption perspective than any happiness above pure sustenance would be for any other individual in the world.

One might resist this characterization of utility maximization. What about a world in which people are more productive, and as a result, can support more people? There might be an incentive to invest in things that will eventually help support more lives, like medicines or labor-saving machines. Such a world might not look like the Malthusian wasteland I worried the utility maximization calculation would encourage. While this world may differ from the Malthusian one I proposed, it does not differ in a meaningful sense. People would still be forced to sacrifice any utility above the sustenance level, except the resources saved would go towards investments in new technologies or other routes of investment, rather than directly to supporting lives. The central role of life as a vector of utility remains, and the incessant mandate to create life for its own sake also remains as a result.

If we are obligated to reproduce for the future, and live lives sustained by gruel and cornmeal in service of such a goal, the future might indeed be quite repugnant. The math of diminishing marginal utility essentially ensures such a conclusion. For any future that might seem alluring or luxurious, there is surely some way in which that which makes it alluring might

be repurposed to support additional future lives. In short, so long as anyone is enjoying a level of happiness more than barely predisposed to living, we are robbing the world of potentially greater utility. Comfort becomes morally repugnant in itself, with every indulgence purchased at the price of an innocent life.

Our criterion of morality is the maximization of utility; faced with a set of actions, then one which maximizes the good is the one which is right, and the others are wrong. Perhaps this is too simplistic. One might argue that to pass some sort of moral judgement, there has to be a victim of the act under consideration. For example, for some action to be right or wrong, *someone* has to be wronged. If I have pursued some act that has not maximized the potential good in someone's life, when I could have reasonably done otherwise, I have wronged *them*. This would alter our consideration thus far of future generations. Who is it, in the future, exactly, that I have wronged by not giving birth to them? There might be less utility in the world than there would have been otherwise, but who suffers this decrease in utility? Considerations of identity might provide a potential route out of enslavement by the future.

The principle of utility seems to negate any considerations of identity when weighing different futures. The simple criterion of morality is the net utility change resulting from some course of action; who, how, why, or what it affects is of no matter. The action is a means to an end. To argue that one has no obligation to reproduce on the basis of identity – ie, that one cannot harm an unborn child by not giving birth to them – is to in effect abandon the principle of utility in favor of a higher moral consideration. The principle of utility concerns itself with the quantity of utility in the world (or the quantity per person, if one is appealing to average utilitarianism). This must be its primary concern; arguing that the unborn are of no consequence not only ignores

any consideration of the utility they might enjoy but is also fairly outlandish when considered in other contexts.

Some argue that we have no obligation to give birth to the unborn as we have not wronged anyone by not doing so. I find this fairly counterintuitive generally speaking, and it is unclear that utilitarians can appeal to this argument. It hardly seems moral to exacerbate climate change for marginal comforts; destroying the planet would harm the as-yet unborn descendants of present humanity. This seems like a perfectly intuitive judgement, yet it is essentially similar to the arguments surrounding giving birth to additional children. The subjects of utility – the ones who will experience it – are not yet born. Yet we still ought to consider the utility they might experience. If it is net positive, then they ought to give birth to the unborn child. According to the objection's reasoning, by harming the planet we might not be able to argue that we are "harming" anyone in particular. If we further stipulate that this edition of climate change is particularly slow-acting, and will only become apparent in many, many generations, this charge becomes even stronger. Presumably it is still immoral to destroy the planet. Yet it will be neither the present, nor the next generation, nor even the one after that that will experience this disutility. Even appeals to the strength of personal ties fall flat; I feel less connection to my potential unborn offspring ten generations hence than I do to those living in present-day South America (with whom I share little personal connection whatsoever). On what grounds then, if not by appealing exclusively to utility, are we to prevent the destruction of the planet? Utilitarians cannot simply disregard unborn lives as meaningless; if said lives offer the potential to contribute positively to the overall total utility of the world, then they deserve the same treatment as any other potential source of utility. For the utilitarian, utility is utility, born or otherwise.

It must also be noted that even without considerations of the future, the unborn, or rather, those with the potential to live, might represent a utility monster in their own right. Consider the world in which we decide to commit ourselves to borderline penury and devote our resources to supporting a future population. The burgeoning future need not enslave us; the handful of additional lives that might be had by means of self-imposed destitution might do just fine. Perhaps we simply ought to reduce all consumption, at the nearest instant, to the absolute minimum level and begin to reproduce. In short, perhaps there need be no utility monster at all. Simple and absolute devotion to maximizing total utility will achieve the same result as serving any one being. What is more interesting, however, is whether the uncomfortable nature of the utility monster changes as a result of a change in the monster's form. Perhaps the monster seems less monstrous, if, instead of a concrete entity, it simply becomes devotion to a number (total utility). Or rather, instead of a number, a series of individuals, many of whom would perish if any one of them attempted to live a happier life.

The Repugnant Conclusion seems at first, quite similar to Longtermist visions of massive future populations. However, the two potential monsters are still subtly different. While utilitarians have at least some grounds to immediately resist the repugnant conclusion, Longtermists have no such plausible response. The utilitarian might say that *if* it were the case that a world with many people in it would maximize utility, then the Repugnant Conclusion would hold. However, they might resist the Conclusion by rejecting the premise that such a world would maximize utility. The Longtermists, however, are immediately committed to such a world. Longtermism is motivated by the idea that the future *will* have many, many people, and that serving these people *will* be the utility maximizing course of action. The monsters differ somewhat in their appearance. For utilitarians, the conclusion is repugnant because of its low

standard of living. The Longtermist monster might consist of many such similarly repugnant worlds spread across time, or instead, enough time periods in which the world seems hardly repugnant at all, but there are simply enough generations that future lives outweigh those of the present. Even though this Longtermist scenario is not repugnant in the same sense, the mass of the future might still play a similar role; because the unborn are so numerous, they might ruin any consideration that the living would have in a utility calculation.

We have so far specified two types of utility monster: the far and near future unborn. There is a tension between these two monsters. It is possible that one subsumes the other. Consider a world in which society recognizes its utilitarian obligation to reproduce. It devotes all its resources to bringing about this world and is swamped in population. Yet at any given point within this world, the future still looms large. Even a world that is operating at the maximum utility possible at any point is still lorded over by the future. It is simply impossible for the present millions, even if we have brought the previously unborn millions into being, to ever achieve parity with the future. For any world that is operating at its carrying capacity, there might exist some reasonable number of future generations of similar or even greater size. And, as we have established thus far, these generations carry equivalent moral weight. It seems then, that the monster hidden in the far future might devour the one hidden in the near future; if we were considering who we ought to prioritize in terms of resource devotion, the unborn of the far future might override those of the near future. It seems at least plausible that this might represent some sort of improvement on considering two monsters.

Considering whether the far-future monster lords over the monster lurking in the near future is cold comfort at best. It is fundamentally an empirical consideration, which is fickle ground to begin with. It also isn't necessarily clear that the two monsters are necessarily as

different as they initially seem. While they do concern different time horizons, both essentially are monsters comprised of the corporate entities of the unborn. The time horizon isn't quite as important; it merely demarcates an arbitrary temporal cutoff. This seems to show that the real force of the Longtermist example lies not in the size of the *far-future* per-se, but rather it lies in the sheer mass of the unborn that constitute the far-future. However, another interesting relationship between the moral burden presented by different populations and time can be observed in the "investing" cycle of happiness.

A consideration of personal responsibility demonstrates the potential conundrum facing the demands of the future upon the present. Suppose Paul is a devoted utilitarian. He has been since his youth. He recognizes that, because of the compounding nature of interest, saving early will reap significant rewards later. He pursues this notion to its extreme; he spends all of his youth, and nearly all of his working life saving in an extreme manner. He does not travel, dine out, and subsists upon rice and beans, with the goal of investing his savings. We would refer to Paul as someone taking part in the "investing" cycle of happiness. He does this until his death, never spending his retirement. Perhaps we face a similar scenario to Paul in considering the future. We are making extensive sacrifices to ensure that the future is slightly better off. Yet by the time the future rolls around, and becomes the present, it too is subject to the demands of the future. The new-present then duly commits itself to a life of penury in service of the massive future, and the cycle begins anew.

We are sacrificing in service of the unborn millions in the future. But these unborn millions have effectively, no start and no end. At any point in the continued existence of humanity there are, presumably, and barring some horrific extinction-level event, more people yet to live. If humanity were to continue indefinitely, who would be the recipient of our hard

work? If every successive generation is merely committed to slaving for the future, and there is always a future, then there seems to be no end in sight. We would be condemned to simply save for a future that can never enjoy the fruits of our labor, a never-ending consideration of the unborn millions.

While this Longtermist treadmill does seem to present a problem for Longtermism, it does not seem to help us reject Longtermism generally for the purpose of defending EA. What remains is the fact that even if no population ever enjoys the payoff from their tireless investments in their posterity, investing in one's future generations is still the utility-maximizing course of action. Choosing the utility-maximizing course of action is never a question of whether anyone will necessarily enjoy the utility, it is merely a question of what the utility maximizing course of action would be.

Maximizing total utility has presented us with a troubling situation. If everyone acts to maximize total utility, then we find ourselves in a reproductive race to the bottom. Total utility increases, but *no one* is happier as a result. We can only find solace in our mutual support of the existence of others. Without our sacrifice, they would not be able to live their barely-better-than-worthless lives. Longtermists ought to find no discomfort with such a conclusion. This could be entirely consistent with their position; considering the long-term future is valuable because more utility will be experienced in the future than in the present, even if it is a seemingly miserable lot. In such a world there will be many, many more descendants than there otherwise might have been, the better maximize utility. This might reasonably reveal then, a more fundamental tension at the heart of utilitarianism, or at least total utilitarianism. We ought to maximize utility. However, when we do this, everyone's lives become worse; we are only better for having lived more moral lives, although that hardly heats homes or fills stomachs. In short, that which at first

appears to value improving the happiness of the world, in fact incentivizes sacrificing it at the individual level.

The average-utilitarians of the world might be sitting smug at this point. Rather than maximize total utility, the average utilitarian wishes to maximize the average utility of a given population. It seems that, because of the moral mathematics of diminishing utility, average and total utilitarianism must be incompatible in some sense; any world in which the average utility is above the subsistence level is one in which total utilitarianism would demand more births, and any world in which the average level of utility is at the subsistence level is one in which average utilitarianism would suggest that fewer births is a better course of action. Therefore, if we are to save any form of utilitarianism, we must choose one or the other. However, even if we do abandon total utilitarianism, it is unclear that this saves utilitarians from enslaving the present, and in fact invites other unsavory conclusions.

It was noted earlier that the single-entity utility monster still plagues average-utilitarianism – it simply requires a monster that receives, for any population of *n* persons, *n* times as much utility from any given resource as the average member of the population. This might be implausible, bit it is hardly any less plausible than an entity that suffers no bound on utility-enjoyment to begin with (as the original Nozickian total-utility monster must). Do the unborn millions (an alternative form of the monster) also affect average-utilitarianism? They do not. Consider the choice faced by an average-utilitarian when considering whether to reproduce. They only ought to do so if the world with their child in it could reasonably be expected to enjoy a greater average utility. This consideration places a greater concern on the denominator involved in the utility calculation; empirically, maximizing average-utility often involves increasing utility and holding population constant (or even decreasing it). Population growth, unless new births

provide geniuses that contribute greatly to overall utility, works against maximizing average-utility. Faced with unborn millions, the average-utilitarian shrugs; they are only concerned with those unborn that might contribute to increasing the average utility of the world.

Average-utilitarianism might also free us from the overriding interest of the future, although its freedom is largely contingent on empirical considerations. At first, it seems as though average-utilitarianism might still be subject to the influence of the future. If the future consisted of very many small generations, so long as there were a chance of *enough* generations existing, their interests might outweigh those of the present. However, two facets of assessing average-utility make this an unlikely outcome. The first is that, taken to its extreme, average-utilitarianism would likely advocate for very small generations (presuming a fairly high bound on the amount of utility that any one individual can enjoy). This would allow many resources to be enjoyed only by a few people at a time. However, the unborn millions exert their influence via their mass; if they must be few at any one time, then they must be temporally disparate. We might reasonably then discount their existence on the basis of probability; it is difficult to assign probabilities into the *far, far* future. These considerations, however, are entirely contingent on empirical analysis of the resources available to the world, and the certainty with which we can forecast the future. It might also be the case that the most likely future contains large, long-lasting generations whose interests override our own. This possibility would again doom us to servitude.

At best, average-utilitarianism might free us from serving the future. However, this solution is less compelling within the context of utility maximization. The principle of utility compels us to adopt utility as our criteria of moral good. I have attempted to maximize utility. But this led me to strange places; not only did the interests of the future far supersede our own,

but so too did the interests of the potential unborn millions that could be born in the near future. So, I have turned to average-utilitarianism in the hope that it might decrease the moral import of the future by preventing it from leveraging its size. Yet this seems like a strange turn, in some regard. It seems reasonable to maximize utility if utility is the criterion of good. It seems less intuitive that we ought to maximize average utility. Are we compelled to maximize the good in the world, or rather do we wish for all persons to each have as good a life as possible? The latter allows us to disregard the potential of additional lives containing net good, if this net good is less than the average. This is because the focus is on improving the average rather than the total. But this seems at odds with adopting utility as a criterion of good.

Parfit captures this tension with what he calls Mere Addition. He offers a comparison between two worlds: world A, and world A+. World A contains some number of people enjoying happy lives. World A+ contains the same number of happy people enjoying the same number of happiness, with an additional group of people that are separated from those so far mentioned, that enjoy a slightly lower level of happiness, but are still quite happy. The residents of each happiness group in A+ are simulatively unaware of each other's existence. He defines this scenario more generally as Mere Addition, which occurs "when, in one of two outcomes, there exist extra people who (1) have lives worth living, (2) who affect no one else, and (3) whose existence does not involve social injustice" (Parfit 420). Mere addition relates significantly to the tension between adopting utility as a criterion of moral good and pursuing a maximization of average-utility.

Mere addition implies certain evaluations that when, taken in the context of the principle of utility, stretch any reasonable application of the principle itself. Parfit offers a short example

demonstrating the effects of Mere Addition on average utilitarianism. He begins by considering Eve and Adam:

> On the Average Principle, the best history might be the one in which only Eve and Adam ever live. It would be worse if, instead of Eve and Adam, a billion billion other people lived, all with a quality of life that would be almost as high. Though this claim is hard to believe, it is not absurd. The second history is in one way worse. It is bad that no one's life is quite as good as Eve and Adam's would have been (Parfit 420).

This claim is hard to believe in part because it stretches any credulous appeal to maximizing utility. It might more reasonably be referred to as the maximization of the utility of *some*, or rather a prioritization of the *average*. Regardless, Parfit then continues to detail truly outlandish conclusions that follow from the Average Principle. He revisits Eve and Adam:

> The Average Principle has other implications which *are* absurd. Suppose that Eve and Adam had lived these wonderful lives. On the Average Principle it would be worse if, *not instead but in addition*, the billion billion other people lived. This would be worse because it would lower the average (Parfit 420).

This is clearly an outlandish conclusion. If the original motivation to adopting utility as a criterion for good was its intuitive appeal, then mere addition must spell the end of average-utilitarianism. How could it possibly be that the mere addition of good to the universe is to be regarded as a bad outcome? This is counter-intuitive in a sense completely different from the conclusions that total-utilitarianism led us towards. The Repugnant Conclusion at least results in a stupendous amount of utility, if achieved in an intuitively repugnant manner. Average-utilitarianism, by contrast, has led us to prefer what is, on any reasonable view, a worse world to

a better one. To argue otherwise would be to assign a primacy to the mathematical construct of an average over utility itself.

So, we ought to reject average-utilitarianism. Where does this leave our search? We hoped that an appeal to averages might diffuse the import of the future. It very well might have. However, adopting average-utilitarianism entails essentially abandoning the moral intuitions that undergird the principle of utility. Given that saving this very principle from the utility monster was our original goal, we are at best, back where we started.

It seems then, that the moral mathematics of utility trap us within the Repugnant Conclusion, and by extension, within the maw of the utility-monstrous unborn millions. We originally worried that the far future might present a burden on our decision-making process. And to a significant degree, it still does; the Longtermists have proved correct, insofar as one is unconcerned with the burden this might present. Considering the far-future still seems to be the utility maximizing course of action. However, we have also discovered that there is another monster lurking in the near-future as well. Because we might support additional life by reducing the resources we expend on comforts, we might be compelled by the possible existence of millions of additional lives to drop everything and work to make these lives a reality.

Average utilitarianism no longer presents a possible way to trim the moral weight of the future. I began by arguing that utilitarians are in fact required to bring about the Repugnant Conclusion, as a result of the obligation to give birth so long as the child's life will be net positive in terms of utility, and as a result of the nature of diminishing marginal utility. I argued that the obligation to the Repugnant Conclusion presented a utility monster in its own right and motivated the potential shift to average utilitarianism. However, upon considering Parfit's Mere

Addition paradox, I found average utilitarianism to be completely unpalatable, and rejected it as a result, eliminating the possibility of its use an objection to Longtermism.

VI

Concluding Remarks

After numerous attempts to lessen the burden of the future on the moral calculations of the present, I have failed to do so. I provided a conception of EA that used total-utilitarianism as its moral system; that which maximizes utility is right. I then considered the role of the future by means of Longtermism, which mandated that one consider the far-future effects of one's actions, as this would maximize total utility due to the millions of unborn humans yet to live. I argued that the corporate mass of these unborn humans represented a form of the Nozickian utility monster that was quite real, and not merely a wildly unrealistic figure of Nozick's imagination. I then expended considerable effort investigating ways of avoiding valuing the future so highly, in the name of protecting the moral import of the present.

I considered whether the undue burden of the future was a result of using expected value to assess the *ex ante* payoff of future decisions. To support such a conclusion, I considered a series of examples that demonstrated pitfalls in using expected value to assess the payoff of decisions. I proposed considering the variance of the payoff of a decision as a way to avoid the importance of the future; because the payoff of many far-future decisions is highly variable, avoiding high-variance payoffs might have blunted the influence of the future in present-day decision making. While this may serve to mitigate the role of the future in some situations, like asteroid deflection, it failed to completely rule out the possibility of a situation in which the

future had some low-variance, high expected value avenue that came at a cost to the present in terms of forgone opportunity.

Reconsidering expected value failed to decrease the influence of the future within utility calculations. So, I turned towards a temporal discount rate of some kind, with the hope that this discount rate would erase any influence of the far-future on the present within utility calculations. Unfortunately, Derek Parfit has already supplied a series of highly compelling rebuttals of any temporal discount. I considered and accepted his objections to temporal discounting and concluded that this too could not decrease the influence of the future.

Having exhausted two significant routes for avoiding the influence of the future, I examined the moral obligation to have children, whether utilitarians were obligated to bring about the Repugnant Conclusion, and whether the Repugnant Conclusion itself represented an additional utility monster with which EA would have to contend. I explored the possibility of resorting to average-utilitarianism as respite from the monsters lurking in the future and the present. I then considered Parfit's Mere Addition Paradox. I concluded that the Mere Addition Paradox provides decisive reason to abandon average-utilitarianism completely, and concluded that again, the present is subject to the considerations of the future.

I finally consider now, what exactly is to be done about all of this. Three significant means of objecting to Longtermism and the burden of the future have failed. Seeing as this is the case, it is worth revisiting the original structure of the objection that stems from Longtermism. Longtermism might require the present to pay a significant price *X* for the sake of some payoff for the far future *Y*. If *X* were sufficiently large enough, and *Y* were marginal and spread over many, many people, I argued that *X* might represent a morally offensive price to pay for such an increase in utility *Y.* Utilitarianism, and by extension, EA, seems to commit us to pay *X*, so long

as doing so is the utility maximizing course of action, independent of the actual form of *Y* (be it significant good concentrated in a handful of people, or negligible good for trillions). There then remain two courses of action. Either paying *X* is *not* actually a morally offensive action, or I must reject utilitarianism. Recall again the uncertain disease. Achieving a slight increase in the certainty of preventing a head cold for trillions might come at the cost of millions of unnecessary deaths. While forbearing on a final ruling, I take this to be my final conclusion: if there exists some situation in which the present must pay some morally offensive price to secure marginal gains at the individual scale for future trillions, and we cannot question the moral intuition surrounding this situation, then we are compelled to reject utilitarianism.

Works Cited

Beckstead, Nick, and Teruji Thomas. "A Paradox for Tiny Probabilities and Enormous Values." *Noûs*, vol. n/a, no. n/a. *Wiley Online Library*, https://doi.org/10.1111/nous.12462. Accessed 13 May 2023.

Effective Altruism - MacAskill - Major Reference Works - Wiley Online Library. https://onlinelibrary.wiley.com/doi/abs/10.1002/9781444367072.wbiee883. Accessed 13 May 2023.

Greaves, Hilary, and William MacAskill. *The Case for Strong Longtermism - Hilary Greaves and William MacAskill (Global Priorities Institute, University of Oxford)*. 14 June 2021, https://globalprioritiesinstitute.org/hilary-greaves-william-macaskill-the-case-for-strong-longtermism-2/.

MacAskill, William, et al. *Giving Isn't Demanding*. Oxford University Press, 2018. *DOI.org (Crossref)*, https://doi.org/10.1093/oso/9780190648879.003.0007.

Nozick, Robert. *Anarchy, State, and Utopia*. Basic Books, 1974. *Open WorldCat*, http://www.gbv.de/dms/bowker/toc/9780465002702.pdf.

Parfit, Derek. *Reasons and Persons*. Clarendon Press, 1987.

Peterson, Martin. "The St. Petersburg Paradox." *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2022, Metaphysics Research Lab, Stanford University, 2022. *Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/archives/sum2022/entries/paradox-stpetersburg/.

Singer, Peter. "Famine, Affluence, and Morality." *Philosophy and Public Affairs*, vol. 1, no. 3, 1972, pp. 229–43.

---. *The Life You Can Save: How to Do Your Part to End World Poverty*. Random House trade pbk. ed, Random House, 2010.