Super Fanaticism

Abstract

According to *fanatics*, no matter how good a guaranteed payoff is, it is worse than almost certainly getting *nothing* and a tiny probability of getting some much better payoff. Most find this claim to be, at the very least, highly counterintuitive. But rejecting fanaticism is easier said than done because it is entailed by other highly plausible principles, leaving us with a paradox. This paper aims to deepen the fanatical paradox. According to *super fanatics*, no matter how good a guaranteed payoff is, it is worse than almost certainly getting a *very bad payoff* and a tiny probability of getting some much better payoff. I demonstrate that super fanaticism is entailed by principles no less plausible than those used in the original paradox, before drawing out super fanaticism's implications for interpersonal cases, longtermism, and the fanaticism debate more broadly.

1. Introduction

As you are about to take your final breath an angel appears and tells you that, if you would like, she can guarantee you ten extra years of happy life. Just before you accept her offer, the devil appears and offers you a one in a quadrillion chance of getting 10^{100} extra years of happy life and nothing otherwise. Would you accept the devil's offer instead of the angel's? Most of us, I take it, would not. After all, you can be 99.9999999999999% sure that the devil's risky prospect will come to nothing and you will die now. To put this in perspective, you are more likely to win the jackpot of the UK's national lottery twice in a row, and so opting for the risky prospect is tantamount to giving up ten years of happy life for nothing.

However, according to fanatics this is *exactly* what you should do. More precisely, fanatics claim that no matter how good a guaranteed payoff is, it is worse than almost certainly getting *nothing* and a vanishingly small probability of getting some much better payoff. Given fanaticism's counterintuitiveness it may seem clear that it should be rejected. Sadly, and somewhat predictably, matters are not so simple. Our best developed theories of prospects' values—such as standard expected utility theory and Buchak's (2013) risk-weighted expected utility theory—are fanatical.[1] Even worse, fanaticism is also entailed by the conjunction of two eminently plausible principles. First, *Transitivity*, which states that if A is strictly better than B, and B is strictly better than C, then A is strictly better than C. Second, *Non-Timidity*, which states that a slight decrease in the probability of getting a payoff can always be outweighed by some increase in the payoff's size. We can see this entailment as follows. Let $\mathcal{P}_0$ be a prospect that guarantees any payoff, as large as you like. By *Non-Timidity*, there is some prospect, $\mathcal{P}_1$, that gives a slightly lower probability of getting a payoff than $\mathcal{P}_0$ but whose payoff is sufficiently larger than $\mathcal{P}_0$'s for $\mathcal{P}_1$ to be the better of the two prospects. Likewise for $\mathcal{P}_1$ and every subsequent prospect: by *Non-Timidity*, there is a prospect that gives a slightly lower probability than its predecessor of getting a larger payoff, and which is better than its predecessor. Applying *Non-Timidity* enough times eventually yields a prospect, $\mathcal{P}_n$, that gives a tiny probability of getting some enormous payoff. By *Transitivity*, $\mathcal{P}_n$ is better than the certain prospect with which we began, $\mathcal{P}_0$. Since the payoff guaranteed by $\mathcal{P}_0$ could have been as large as you liked, we have fanaticism.

What to do? One the one hand, *Transitivity* and *Non-Timidity* are incredibly plausible; on the other hand, fanaticism seems unpalatable. In response to this paradox, some have sided with the standard theories and argued that fanaticism is not so bad after all (e.g. Hájek, 2014; Parfit, 1984, 73–75; Wilkinson, 2022), whilst others remain on the fence (e.g. Beckstead, 2013; Beckstead & Thomas, 2021; Russell, 2021). The aim of this paper is to deepen the fanatical paradox by showing that there is a conclusion even worse than fanaticism in the offing: *Super Fanaticism*. Super fanatics claim that no matter how good a guaranteed payoff is, it is worse than almost certainly getting a *very bad payoff* and a vanishingly small probability of getting some much better payoff. More specifically, I shall deepen the paradox by showing that *Super Fanaticism* is entailed by *Transitivity* and a principle no less plausible than *Non-Timidity*, which I shall call

[1] That is, assuming agents' utility functions are unbounded and they are not extremely risk averse (see Beckstead & Thomas, 2021, pp. 8–12).

*Non-Timidity**. In doing so, I hope to cast doubt on the thought that embracing fanaticism is a tenable solution to the original paradox.[2] Here is the plan. §2 introduces *Super Fanaticism* and *Non-Timidity**, and §3 provides an intuitive demonstration of how *Non-Timidity** and *Transitivity* entail *Super Fanaticism* (a formal proof is reserved for the appendix). §4 concludes by highlighting the disturbing ethical implications of super fanaticism—that we should opt for a prospect that almost certainly results in the extermination of every child on Earth but gives a tiny probability of saving a sufficiently large number of lives, instead of saving 200 lives with certainty—as well as its upshots for longtermism and the fanaticism debate.

Before we jump in, some terminology and notation. Let $\mathcal{O}$ be a set of *outcomes*, each of which can be described solely by its payoff, a finite and quantifiable gain or loss of something (dis)valuable. Until §4, we shall assume the payoffs are additional years of life of varying happiness or misery *for the decision-maker*. Whether a payoff is *positive* or *negative* is defined relative to the *status quo*, so, for example, a positive payoff is one that leaves the agent better-off than before.[3] A *prospect* results in different outcomes (or, equivalently, payoffs) with different probabilities and, formally speaking, is a function from $\mathcal{O}$ into the interval $[0, 1]$ such that all of the outcomes' probabilities sum to 1. We shall use $p_1 * x_1 + p_2 * x_2 + \ldots + p_n * x_n$ to represent a prospect that results in payoff x_1 with probability p_1, payoff x_2 with p_2, and so on up to x_n and p_n. For a prospect that results in a (non-zero) payoff of x with probability p and nothing otherwise, we shall simply write $p * x$. Some prospects are better than others, so we shall assume that there is a binary relation, $\succcurlyeq$, on the set of all prospects that represents the 'at least as good as' relation. As ever, strict betterness ($\succ$) and equality ($\sim$) are respectively defined as $\succcurlyeq$'s asymmetric and symmetric parts. We will assume that $\succcurlyeq$ is reflexive and transitive, but it need not be complete.[4]

[2] For readers familiar with the population ethics literature, the difference between fanaticism and super fanaticism parallels the difference between the repugnant conclusion and the very repugnant conclusion (see §4).

[3] Notice that this does not require an *absolute* zero—we are merely stipulating that the zero point is the *status quo*. Thus, we are only assuming that payoffs are measured on an interval scale and therefore remain within the standard von Neumann-Morgenstern framework.

[4] Of course, some—most notably Temkin (2012)—have argued that *Transitivity* is the culprit that leads us to paradox. Since I regard *Transitivity* as sacrosanct, I shall simply assume *Transitivity* to keep things manageable.

2. Super Fanaticism & Non-Timidity*

According to super fanatics, for any finite positive payoff and any finite negative payoff, no matter how large either of them might be, and for any probability, no matter how small, there is some much larger positive payoff such that getting the larger positive payoff with a vanishingly small probability and the negative payoff otherwise is better than getting the first payoff for sure. More precisely:

> *Super Fanaticism.* For any finite positive payoff x, any finite negative payoff z, and any probability q, there is some finite positive payoff y such that $1 * x \prec q * y + (1 - q) * z$.

So, unlike the fanatic who 'merely' claims that for any guaranteed positive payoff it is better to almost certainly get *nothing* but have a tiny chance of getting some much better payoff, the super fanatic claims that it is better to almost certainly get a *very bad* payoff and a tiny chance of getting some considerably better payoff. Returning to the earlier example, the fanatic was willing to forgo a guarantee of ten extra years of happy life in favour of almost certainly getting *nothing* but having a one in a quadrillion chance of getting 10^{100} extra years of happy life. By contrast, the super fanatic would forgo the guaranteed ten extra years in favour of almost certainly getting 1000 *years of torture* but having a one in a quadrillion chance of getting some sufficiently large number of extra years of happy life, say, 10^{1000}.

Our second principle is *Non-Timidity**, which states that a slight decrease in the size of one payoff can be outweighed by *some* (potentially very large) finite increase in the size of another *more likely* payoff. To illustrate, suppose again that you are on your deathbed, but this time the angel offers you 10 extra years of happy life with probability 0.8 and 9 extra years with probability 0.2. Just before you accept her offer, another angel appears and offers you many more years of happy life with probability 0.8, but a slight decrease in the amount of additional happy you life that you get with probability 0.2, down to 8 years, 364 days, and 23 hours. All that *Non-Timidity** states is that there is *some* number of years of happy life—maybe fifty, a hundred, or a greater number still—such that it is better to accept the second angel's offer.[5]

Following Beckstead & Thomas (2021, p. 5), we can make the notion of 'a slight decrease' more exact by introducing a standard of closeness that specifies when two numbers count as close together and so when a decrease from the higher number to the lower number counts as slight.[6] According to one standard of closeness, a decrease in the amount of additional happy life has to be less than one day to count as slight; according to another, a decrease has to be less than an hour. All that *Non-Timidity** says is that there is *some* standard of closeness such that a slight decrease in the size of one payoff can be outweighed by an increase in the size of a more likely payoff. More formally:

> *Non-Timidity**. For some standard of closeness, for any finite payoffs $x_1, x_2, \ldots, x_n$ with probabilities $p_1, p_2, \ldots, p_n$, for any $i, j \in \{1, 2, \ldots n\}$ such that $p_j > p_i$, and for any close together payoffs x_i and y_i such that $x_i > y_i$, there is some finite payoff $y_j > x_j$ such that:

$$p_1 * x_1 + \ldots + p_i * y_i + p_j * y_j + \ldots + p_n * x_n \succ p_1 * x_1 + \ldots + p_i * x_i + p_j * x_j + \ldots + p_n * x_n.$$

To deny *Non-Timidity** is to say that sometimes, no matter how small the decrease from x_i to y_i in the less likely payoff, it can never be outweighed by *any* increase from x_j to y_j in a more likely payoff—no matter how large this increase might be.

*Non-Timidity** can also be illustrated graphically, as it is in Figure 1 below. On the vertical axis we have the size of the prospect's payoffs. Points above the horizontal axis represent positive payoffs, whilst points below the horizontal axis represent negative payoffs. The higher up the vertical axis a point is, the greater the payoff. On the horizontal axis we have the probability

[5] The requirement that the increased payoff be more likely than the decreased payoff is not required for anything that follows, but is included simply to increase *Non-Timidity**'s plausibility.

[6] We are using the idea of a standard of closeness to clarify when the decrease in a payoff's size counts as slight. By contrast, Beckstead & Thomas use a standard of closeness to clarify when a decrease in the probability of receiving a payoff counts as slight. As they do, we shall assume that closeness is symmetric and that, for any given payoff, x, there is an open interval around x within which any number counts as close to x.

that the prospect assigns to getting a given payoff. The wider a payoff's bar is, the greater the probability of getting that payoff. Dashed vertical lines indicate that the payoff is intuitively much greater than it is represented as being on the diagram. To begin with, ignore the shaded regions and consider a certain prospect with payoff x_i. Suppose that we alter the prospect such that there is now a p_i chance of getting a slightly smaller payoff of y_i. This decrease is represented by subtracting the shorter and narrower shaded area on the right from the prospect's payoff. By *Non-Timidity**, this decrease can be outweighed by increasing the size of a more likely payoff by some sufficiently large amount, from x_j to y_j. This increase is represented by adding the taller and wider shaded area on the left to the prospect's payoff.

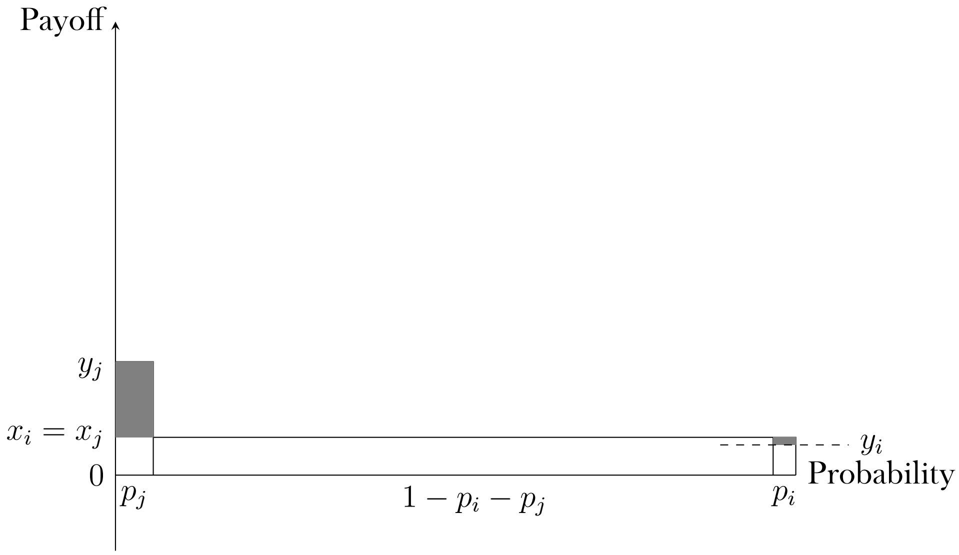

Figure 1: *Non-Timidity**

With *Non-Timidity** in place, we can now clarify its relation to *Non-Timidity*. In what follows, I show that *Non-Timidity** is logically stronger than *Non-Timidity*. However, I think that this increase in logical strength does not translate into a loss of plausibility because, first, the views that satisfy *Non-Timidity* but not *Non-Timidity** are untenable and, second, close variants of the strongest arguments for *Non-Timidity* also entail *Non-Timidity**. However, since I think the first point is decisive, the second point is reserved for appendix B.

One of Beckstead & Thomas' (2021, 15–16) arguments in favour of *Non-Timidity* is that rejecting it also commits one to rejecting *Non-Timidity** (however, they do not give it this name or prove that, together with *Transitivity*, it entails *Super Fanaticism*). Although they do not show this, this is because *Non-Timidity** and *Transitivity* together entail *Non-Timidity*.[7] We can see this as follows. First, notice that via repeated application of *Non-Timidity** a prospect can be made better by decreasing a less likely payoff to zero provided that we increase a more likely payoff by a sufficient amount. This point is illustrated in Figure 2 below. As before, we begin with a certain prospect with guaranteed payoff x_i and *Non-Timidity** allows us to make the prospect better by decreasing its payoff by the smaller grey area on the right provided that we increase its payoff by the larger grey area on the left. But we need not stop there. Next we can further decrease the less likely payoff by a small amount (the smaller blue area on the right) provided that we increase the more likely payoff by some sufficiently large amount (the larger blue area on the left). Although the larger blue area is depicted as being the same size as the larger grey area, the amount by which the more likely payoff must be increased to compensate for a small decrease in the less likely payoff can vary; it need not stay constant. By repeating this procedure for the yellow, red, and green areas, we can reduce the less likely payoff to zero, thereby leaving us with a prospect with a slightly smaller probability of getting a much larger payoff than the original prospect that guaranteed x_i. Since each prospect is better than the last, by *Transitivity* it follows that the final prospect is better than the original prospect. That is, a slight decrease in the probability of getting a payoff can be outweighed by some increase in the payoff's size, which is exactly what *Non-Timidity* states. In a nutshell, *Non-Timidity* allows us to directly trade-off decreases in a payoff's probability against its size, whilst *Non-Timidity** allows us to directly trade-off a less likely payoff's size against a more likely payoff's size. But since a decrease in a payoff's probability can be accomplished indirectly by reducing the payoff's size in some outcomes to zero, *Non-Timidity** allows us to indirectly trade-off a payoff's probability against its size. Thus, provided that one accepts *Transitivity*, *Non-Timidity** is at least as strong as *Non-Timidity*.

In fact, *Non-Timidity** is logically stronger than *Non-Timidity* because a theory of prospects'

[7] Formally, we can state *Non-Timidity* as follows. For some standard of closeness, for any finite payoff x and any close-together probabilities p_i and p_{i+1} such that $p_i > p_{i+1}$, there is some finite payoff $y > x$ such that $p_{i+1} * y \succ p_i * x$.

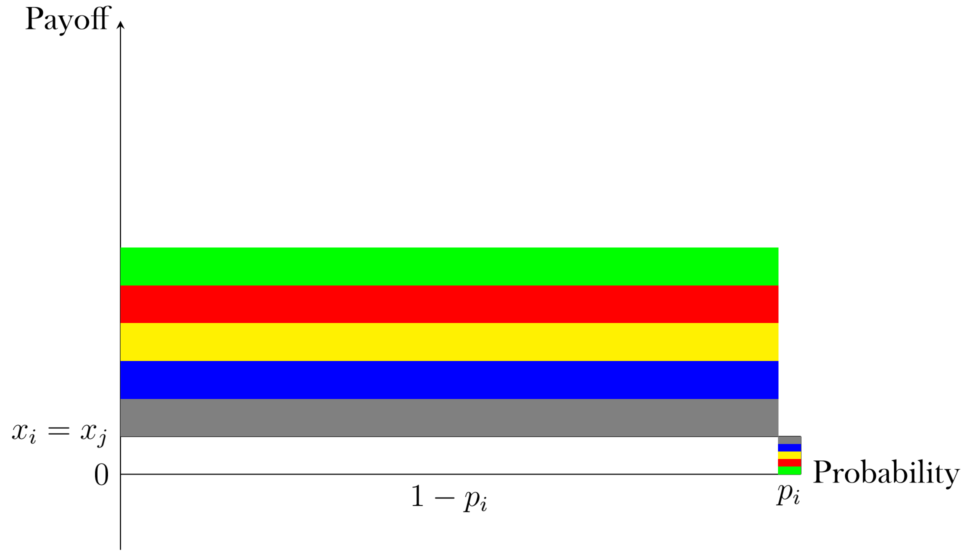

Figure 2: Repeated application of *Non-Timidity**

values could satisfy *Non-Timidity* and *Transitivity* without satisfying *Non-Timidity**. As a toy example, consider a modified version of expected utility theory that incorporates maximal loss aversion as follows: $\mathcal{P}_1 \succcurlyeq \mathcal{P}_2$ iff either (i) the probability of getting a negative payoff given $\mathcal{P}_1$ is strictly less than the probability of getting a negative payoff given $\mathcal{P}_2$; or (ii) the probability of getting a negative payoff given $\mathcal{P}_1$ equals the probability of getting a negative payoff given $\mathcal{P}_2$, and $EU(\mathcal{P}_1) \geq EU(\mathcal{P}_2)$. This theory satisfies *Transitivity* because $>$ and $\geq$ are transitive, and it satisfies *Non-Timidity* for the same reasons as expected utility theory. However, maximal loss averse expected utility theory does not satisfy *Non-Timidity** because, *pace Non-Timidity**, it entails that one cannot make a prospect better by decreasing a less likely payoff such that it becomes negative provided that one increases a more likely payoff's size by a sufficient amount. Thus, *Non-Timidity** is logically stronger than *Non-Timidity*.

That said, it is difficult to see a reason to accept *Non-Timidity* but not *Non-Timidity**. Indeed, it is telling that one of Beckstead & Thomas' arguments for *Non-Timidity* was that its rejection commits one to rejecting *Non-Timidity**. Although *Non-Timidity** is logically stronger than *Non-Timidity*, and so there are views that satisfy the latter but not the former, these views are untenable. The two principles come apart because *Non-Timidity** implies that a small decrease in a payoff such that it becomes (more) negative can be outweighed by a sufficiently large increase in another payoff, whereas *Non-Timidity* is silent on the matter of whether the risk of a loss can be outweighed by some probability of a sufficiently large gain. Thus, unlike *Non-Timidity**, *Non-Timidity* is compatible with theories such as maximal loss averse expected utility theory that impose an embargo on trading increases in the probability of losses for increases in the probability of gains—no matter how unlikely and small the losses are and how probable and large the gains are. Whilst such theories are logically coherent, the embargo that they impose is implausible. If a tiny increase in the probability of a small loss cannot be outweighed by a large increase in the probability of a huge gain then, for example, doctors should not prescribe medications that are all but certain to cure debilitating chronic diseases if they have a one in a billion chance of not curing the disease and giving the patient a minor headache for a minute. Since this is patently false, such views should be rejected. Accordingly, the logical gap between *Non-Timidity* and *Non-Timidity** does not make room for any plausible views. As such, those who accept *Non-Timidity* should also accept *Non-Timidity** since there is no principled rationale for accepting the former but rejecting the latter.

In summary, given that *Super Fanaticism* is even more unappealing than fanaticism and the difference between *Non-Timidity* and *Non-Timidity** is minimal, showing that *Non-Timidity** and *Transitivity* entail *Super Fanaticism* would create a paradox more troubling than the original fanatical paradox.

3. The Super Fanatical Paradox

With these two principles in place, we are now in a position to provide an intuitive demonstration of the claim that *Non-Timidity** and *Transivitity* entail *Super Fanaticism*. Recall that *Super Fanaticism* states that, for any finite positive payoff x, any finite negative payoff z, and any probability q, there is some finite positive payoff y such that getting y with probability q and z otherwise is better than getting x with certainty. So, let x be as positive as you like, z be as negative as you like, and q be as small as you like. What we must prove is that there is some payoff y such that getting y with probability q and z otherwise is better than getting x with certainty.

We saw in the previous section (see Figure 2) that repeated applications of *Non-Timidity** enable us to make a prospect better by reducing a less likely payoff to zero provided that we sufficiently increase the size of a more likely payoff. The key to proving the above claim is that we need not stop at zero. We can decrease the less likely payoff's size by further small increments until the payoff has some negative value, z. Provided that we increase the more likely payoff's size by a sufficient amount each time, each prospect is better than the previous one. This process is depicted in Figure 3 below.

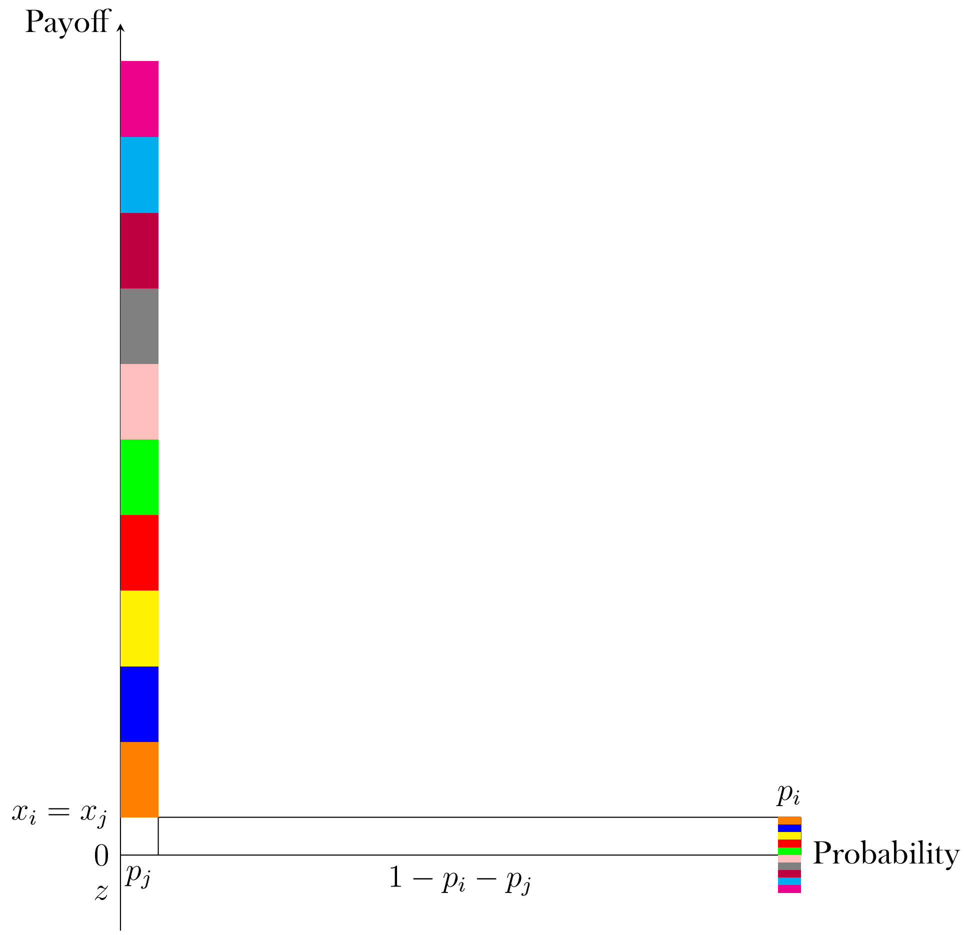

Figure 3: Negative payoffs via repeated application of *Non-Timidity**

We then repeat this process for the remainder of payoff x_j, each time ensuring that the probability of the payoff being decreased is less than p_j. Let y_j be the size of the positive payoff once the only other payoff is z. This prospect, which we shall call $\mathcal{P}^*$, is depicted in Figure 4 below.

By *Transitivity*, it thereby follows that $\mathcal{P}^* \succ \mathcal{P}_0$. That is, getting negative payoff z with prob-

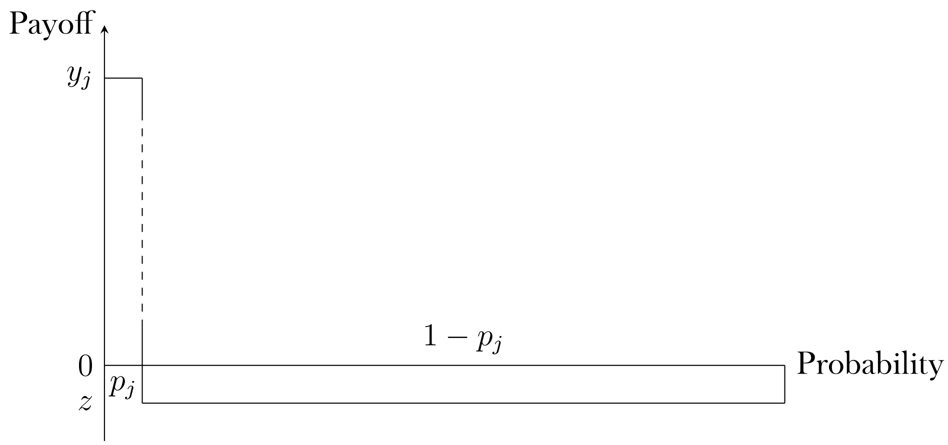

Figure 4: Super fanaticism

ability $(1 - q)$ and y_j otherwise is better than getting x for sure. Since nothing in the proof required x, z, and q to take any specific values—that is, x and z could have been as good and bad (respectively) as you liked, and q could have been as small as you liked—*Non-Timidity** and *Transitivity* entail *Super Fanaticism*.

4. Upshots

I shall conclude by making three observations concerning how the super fanatical paradox extends to interpersonal cases, its implications for longtermism, and the ramifications of the super fanatical paradox for the fanaticism debate more broadly.

Beginning with the former, the foregoing discussion assumed that we are in an intrapersonal context where a single decision-maker chooses between payoffs that *they* receive. However, the super fanatical paradox's implications in interpersonal contexts, where the payoffs are received by individuals other than the decision-maker, are far more disturbing. Assuming that future people's lives have some positive value, fanaticism entails that it would be better to donate money to an organisation where the donation would fractionally increase the probability of humans going on to populate the solar system instead of giving the money to the Against Malaria Foundation (AMF) and saving, say, 200 lives with certainty.[8] Whilst this is tantamount to letting 200 people

[8] Of course, in reality one could never be certain that donating to AMF would save a specific number of lives.

die for nothing, it pales in comparison to the implications of super fanaticism. Consider the following example:

> *Malaria Nets vs. Evil Cabal.* You have $1 million that you can donate to either of two institutions: AMF or Evil Cabal. With your donation, AMF can save 200 lives with certainty. Alternatively, Evil Cabal will use your donation to fund their project of creating a superintelligent AI that will exterminate every child on Earth. However, there is a one in a quadrillion chance that the cabal's programmers will make a mistake and the AI will turn out to be benevolent and enable humans to populate the Milky Way. In this case, 10^{36} is a conservative estimate for the number of future lives given our current evidence (Greaves & MacAskill, 2021, p. 8).

It is one thing to, in effect, let 200 people die for nothing. But in this case the alternative to donating to AMF is all but certain to result in the extermination of roughly 2 billion children, and so pursuing it is tantamount to letting 200 die so that billions can be killed. But this is *precisely* what the super fanatic is committed to doing. This is beyond fanatical. It is outright monstrous.

The increased disturbingness of fanaticism in interpersonal cases is readily—though need not be—explained by what has become known as "the separateness of persons" (see Rawls, 1999, p. 24). Those sympathetic to this line of thought claim that there is a morally significant difference between allowing a single individual to incur a loss so that *they* receive a greater benefit, versus allowing one individual to incur a loss so that *another* individual receives a greater benefit. This is because, in the latter case, the individual who incurs the loss is not compensated by the gain received by the other individual. Accordingly, to treat these two cases alike is to treat the two individuals as though they were a single individual and thereby fail to respect their separateness. Likewise, in intrapersonal super fanatical cases, at least the individual who is almost certain to incur a loss has some chance of receiving an astronomical benefit. By contrast, in interpersonal super fanatical cases, the individuals who almost certainly incur a loss have no such chance.

One might hope to avoid super fanaticism's unwelcome implications in interpersonal cases by appealing to constraints against harming, since these would forbid imposing almost certain losses on some so that others have a tiny probability of receiving benefits. Sadly, this will only work provided the constraints are absolute and forbid inflicting harm regardless of how much good doing so brings about. But contemporary deontologists typically allow that such constraints have thresholds—that is, it is permissible to inflcit a given harm provided that doing so produces a sufficient amount of good—to avoid the extreme verdict that one cannot pinch someone's arm in order to save a billion lives (e.g. Kamm, 2007, 30–31; Thomson, 1990, Ch. 6). Since *Non-Timidity** also says that some sufficiently large gain can offset a small decrease in a payoff—even if that means that the payoff becomes (more) negative—non-absolute constraints cannot shield us from super fanaticism's implications in interpersonal cases.

This brings us nicely to the second point: super fanaticism's implications for longtermism. For present purposes, it suffices to define longtermism as the claim that improving the long term future is the most important moral issue we presently face.[9] The intimate relation between longtermism and fanaticism is revealed by the central argument for longtermism which, in brief, goes as follows (cf. Beckstead, 2013; Greaves & MacAskill, 2021). The potential number of lives in the far future is vast and so there are potential far future outcomes that are enormously valuable, likes ones in which there are quintillions of blissful lives. Moreover, there are prospects available to us that offer tiny probabilities of producing these enormous payoffs, such as research into beneficial AI. Crucially, because the size of a payoff in which these blissful lives exist is so large, it follows via expected utility theory that these long term prospects are better than near term prospects that give high probabilities of bringing about comparatively modest payoffs, such as purchasing insecticide-treated bednets to save lives in sub-Saharan Africa.

Notice, however, that expected utility theory only entails that such long term prospects are better than the near term alternatives *because* it is fanatical—a non-fanatical theory that ignored payoffs with sufficiently small probabilities would favour near term prospects. Leading longtermists regard their reliance on a fanatical theory of the value of prospects as one of the main threats to their view. Thus, Greaves & MacAskill (2021, p. 25) write, "We regard [non-fanatical

[9] For a more precise definition, see Greaves & MacAskill (2021, 3–4).

decision theories] as one of the most plausible ways in which the argument for strong longtermism might fail". The emergence of super fanaticism further aggravates these worries. Not only does super fanaticism pile additional pressure on expected utility theory and thereby cast further doubt upon the longtermists' central argument, but it also highlights further implications of longtermism.

Longtermists typically claim that we should forgo almost guaranteed modest near term benefits in favour of incredibly unlikely far future benefits—instead of purchasing bednets, we should fund research into beneficial AI. Though true, this is an understatement. Since their underlying theory of the value of prospects is not merely fanatical but super fanatical, they are actually committed to claiming that present generations should not merely forgo modest gains but should actually incur tremendous losses to their welfare if doing so has a minuscule probability of producing a sufficiently large far future payoff. To make this concrete, imagine that a small elite could enslave almost everyone on Earth and devote its resources to construct a Dyson sphere, a megastructure that entirely ensloses a star and its surrounding solar system. This would allow humans to harvest the majority of the star's energy, thereby enabling human civilisation to grow almost beyond comprehension. However, life for the enslaved would be unimaginably wretched as almost all resources are devoted to the sphere's construction, and they are separated from their loved ones and allocated to wherever their skills can be put to best use. According to longtermists, provided the number of future lives is large enough, the elite should attempt to construct the sphere even though they are all but certain to fail. For some, this may be too much to stomach.

Turning to the broader ramifications of the super fanatical paradox, it is telling that the difference between fanaticism and super fanaticism parallels the difference between the repugnant conclusion and the very repugnant conclusion. The repugnant conclusion states that, for any number of lives with very high positive welfare, there is some much larger number of lives with very low positive welfare that is better. The very repugnant conclusion states that, for any number of lives with very high positive welfare and any number of lives with very negative welfare, there is some much larger number of lives with very low positive welfare such that a population comprised of the all lives with very negative welfare or very low positive welfare is better than the population comprised of all the lives with very high positive welfare (Arrhenius, 2003, p. 168).

Both the repugnant conclusion and fanaticism involve trade-offs between quality and quantity. In the case of the repugnant conclusion, the quality and quantity are the lives' welfare and the number of lives, respectively; whilst in the case of fanaticism, it is the probability of receiving a payoff and the payoff's size. They prey on the fact that if decreases in quality—be it welfare or probability—can always be outweighed by increases in quantity—be it the number of lives or payoff size—then, for any quantity of high positive quality, there is a much larger quantity of lower positive quality that is better. The very repugnant conclusion and super fanaticism press this point further. If decreases in quality can always be outweighed by increases in quantity then, for any quantity of high positive quality and any quantity of negative quality, there is a much larger quantity of lower positive quality such that the larger quantity of lower positive quantity outweighs both the quantity of negative quality and the quantity of high positive quality.

Thus, given the close parallels between these claims, super fanaticism has much the same implications for the fanaticism literature as the very repugnant conclusion has for the population axiology literature. First, it highlights that theories which entail these claims—such as total utilitarianism in the case of the very repugnant conclusion, and expected utility theory in the case of super fanaticism—face far stronger objections than was initially thought. Second, just as proponents of axiologies that entail the repugnant conclusion have sought to debunk its apparent repugnance (e.g. Huemer, 2008), so too proponents of theories that entail fanaticism have sought to debunk its fanatical appearance:

> Given how widespread these intuitive mistakes [in probabilistic reasoning] are, we should give little weight to our intuitions about what we should do in cases of low probability, including those which lead us to recoil from fanatical verdicts—with a little more scrutiny, those intuitions may appear foolish too. (Wilkinson, 2022, p. 452)

But just as attempts to debunk the intuitions against the repugnant conclusion struggle to debunk the far stronger intuitions against the very repugnant conclusion, one might doubt whether attempts to debunk our intuitions against fanaticism are able to debunk the far stronger intuitions against super fanaticism. Perhaps the fanatics' debunking arguments are up to the challenge, but super fanaticism certainly raises the bar.

Finally, the fanatical paradox requires us to choose between being fanatical, saying that sometimes a small decrease in the probability of receiving a payoff cannot be outweighed by any increase in its size, or denying the transitivity of better than. None of these options are appealing. But given that super fanaticism is far worse than fanaticism whilst the difference between *Non-Timidity* and *Non-Timidity** is at best marginal, the options forced upon us by the super fanatical paradox are even bleaker than those available in the wake of the fanatical paradox. We must either: (i) deny *Transitivity*, and say that A can be strictly better than B and B strictly better than C without A being strictly better than C; (ii) reject *Non-Timidity**, and admit that sometimes a small decrease in a payoff cannot be outweighed by any increase in the size of a more likely payoff; or else (iii) embrace *Super Fanaticism* and claim that, no matter how good a guaranteed payoff is, it is worse than almost certainly getting a very bad payoff and a tiny probability of getting some much better payoff. Now that really is an unappetising choice.

References

Arrhenius, G. (2003). The Very Repugnant Conclusion. In K. Segerberg & R. Sliwinski (Eds.), *Logic, Law, Morality: Thirteen Essays in Practical Philosophy in Honour of Lennart Åqvist* (pp. 167–180). Uppsala: Uppsala University.

Beckstead, N. (2013). *On the Overwhelming Importance of Shaping the Far Future* (Doctoral dissertation, Rutgers University).

Beckstead, N., & Thomas, T. (2021). A Paradox for Tiny Probabilities and Enormous Values. *Global Priorities Institute Working Paper Series*, No.7, 1–39.

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

Greaves, H., & MacAskill, W. (2021). The Case for Strong Longtermism. *Global Priorities Institute Working Paper Series*, No.5, 1–42.

Hájek, A. (2014). Unexpected Expectations. *Mind*, 123, 533–567.

Huemer, M. (2008). In Defence of Repugnance. *Mind*, 117, 899–933.

Kamm, F. M. (2007). *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. Oxford: Oxford University Press.

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

Rawls, J. (1999). *A Theory of Justice* (Revised ed). Cambridge, MA: Belknap Press of Harvard University Press.

Russell, J. S. (2021). On Two Arguments for Fanaticism. *Global Priorities Institute Working Paper Series*, No. 17, 1–41.

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

Thomson, J. J. (1990). *The Realm of Rights*. Cambridge, MA: Harvard University Press.

Wilkinson, H. (2022). In Defense of Fanaticism. *Ethics*, 132, 445–477.

A The Super Fanatical Paradox

Theorem 1. *Non-Timidity* and Transitivity entail Super Fanaticism.*

We recall the definitions of *Non-Timidity** and *Super Fanaticism*:

> *Non-Timidity**. For some standard of closeness, for any finite payoffs $x_1, x_2, \ldots, x_n$ with probabilities $p_1, p_2, \ldots, p_n$, for any $i, j \in \{1, 2, \ldots n\}$ such that $p_j > p_i$, and for any close together payoffs x_i and y_i such that $x_i > y_i$, there is some finite payoff $y_j > x_j$ such that:

$$p_1 * x_1 + \ldots + p_i * y_i + p_j * y_j + \ldots + p_n * x_n \succ p_1 * x_1 + \ldots + p_i * x_i + p_j * x_j + \ldots + p_n * x_n$$

Super Fanaticism. For any finite positive payoff x, any finite negative payoff z, and any probability q, there is some finite positive payoff y such that $1 * x \prec q * y + (1 - q) * z$.

Proof. Fix any real numbers x, z, and q. Take a prospect, $\mathcal{P}_0$, that guarantees payoff x:

$$p_1 * x + \ldots + p_i * x + p_j * x + \ldots + p_n * x$$

where $p_j = q$ and $p_i < p_j$ for all $i \neq j$. Fix the standard of closeness as in *Non-Timidity**, and let c be such that any two numbers whose absolute difference is at most c count as close. By the Archimedean property of the real numbers, there is some natural number m such that $x - m \cdot c < z$. Let n be the smallest number such that this inequality is true. Define:

$$x_i = \begin{cases} x & \text{if } i = 0 \\ x - i \cdot c & \text{if } 1 \leq i \leq n - 1 \\ z & \text{if } i = n \end{cases}$$

In other words, $x_0 = x$, and we keep subtracting c until we are within distance c of z, at which point $x_n = z$. Note that $|x_i - x_{i-1}| \leq c$ for all $i \in \{1, 2, \ldots, n\}$. Use *Non-Timidity** with the parameters in $\mathcal{P}_0$ to define y_1. Similarly, use *Non-Timidity** with the parameters in $\mathcal{P}_{i-1}$ to recursively define y_i. Then, applying *Non-Timidity** n times yields the prospect:

$$p_1 * x + \ldots + p_i * z + p_j * y_n + \ldots + p_n * x.$$

Call this prospect $\mathcal{P}_n$. By *Transitivity*, $\mathcal{P}_n \succ \mathcal{P}_0$. Repeating this process for all of the remaining payoffs yields prospect $\mathcal{P}_{(n-1)n}$:

$$p_1 * z + \ldots + p_i * z + p_j * y_{n(n-1)} + \ldots + p_n * z.$$

By *Transitivity*, $\mathcal{P}_{(n-1)n} \succ \mathcal{P}_0$. This is *Super Fanaticism* with $y = y_{n(n-1)}$, as desired. □

B Extending the Strange Dependence Argument

Beckstead & Thomas' (2021, 16–18) 'strange dependence' argument for *Non-Timidity* shows that *Non-Timidity* is entailed by two eminently plausible principles.[10] First:

> *Weak Dominance*. For any probabilities p, q such that $p > q$, and any finite payoff y, there is some finite payoff $x > y$ such that $p * x \succ q * y$.

That is, for any payoff y and any probability q of receiving that payoff, it is better to get some larger payoff x with higher probability p. Second:

> *Separability*. For any prospects $\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3$, each with finitely many payoffs, if $\mathcal{P}_1 \succ \mathcal{P}_2$ then $\mathcal{P}_1 + \mathcal{P}_3 \succ \mathcal{P}_2 + \mathcal{P}_3$.

Where $\mathcal{P}_1 + \mathcal{P}_3$ is the prospect that results from summing $\mathcal{P}_1$'s and $\mathcal{P}_3$'s payoffs. Thus, *Separability* states that if $\mathcal{P}_1 \succ \mathcal{P}_2$, then altering their payoffs by the same amounts leaves this ranking unchanged. To illustrate, let $\mathcal{P}_1$ and $\mathcal{P}_2$ be prospects that pay £10 and £5 if a coin lands heads, respectively, and nothing otherwise. Clearly, $\mathcal{P}_1 \succ \mathcal{P}_2$. Moreover, suppose that $\mathcal{P}_3$ pays £2 if the coin lands tails and nothing otherwise. So, $\mathcal{P}_1 + \mathcal{P}_3$ pays £10 if the coin lands heads and £2 if the coin lands tails, whilst $\mathcal{P}_2 + \mathcal{P}_3$ pays £5 and £2, respectively. *Separability* entails that $\mathcal{P}_1 + \mathcal{P}_3 \succ \mathcal{P}_2 + \mathcal{P}_3$. And this seems correct: if we start with one prospect that is better than a

[10] This presentation of Beckstead & Thomas' argument broadly follows Russell's (2021, 6–8). Russell casts doubt on *Separability* by showing that a version that is not restricted to prospects with finitely many payoffs is inconsistent with an even more plausible principle, the principle of stochastic dominance (i.e., for any prospects $\mathcal{P}_1, \mathcal{P}_2$ and any payoff x, if $p[\mathcal{P}_1 \geq x] \geq p[\mathcal{P}_2 \geq x]$ then $\mathcal{P}_1 \succcurlyeq \mathcal{P}_2$; and if there is also some x such that $p[\mathcal{P}_1 \geq x] > p[\mathcal{P}_2 \geq x]$ then $\mathcal{P}_1 \succ \mathcal{P}_2$). Regardless of whether we think that the inconsistency of the unrestricted version of *Separability* with stochastic dominance in infinite cases undermines *Separability*—that is, regardless of whether we think the strange dependence argument succeeds—the point is that arguments for *Non-Timidity* also support *Non-Timidity**.

second prospect, and we alter them in exactly the same way, then the altered first prospect should still be better than the altered second prospect.

The argument from *Weak Dominance* and *Separability* to *Non-Timidity* goes as follows. We must show that there is some sufficiently large payoff y such that increasing the size of the payoff from x to y outweighs a slight decrease in the probability of receiving a payoff from p_i to p_{i+1}. Letting $y = x + b$ and $p_i = p_{i+1} + q$, this is to say that we must show that there is some b such that increasing the payoff's size by b outweighs a slight decrease of q in its probability. Let $b > x$, and consider the prospects in Table 1 below.

Prospect	p_{i+1}	q	$1 - p_{i+1} - q$
$\mathcal{P}_1$	b	0	0
$\mathcal{P}_2$	0	x	0
$\mathcal{P}_3$	x	0	0
$\mathcal{P}_1 + \mathcal{P}_3$	$x + b$	0	0
$\mathcal{P}_2 + \mathcal{P}_3$	x	x	0

Table 1

Since $b > x$ and $p_{i+1} > q$, *Weak Dominance* entails that $\mathcal{P}_1 \succ \mathcal{P}_2$. By *Separability*, it follows that $\mathcal{P}_1 + \mathcal{P}_3 \succ \mathcal{P}_2 + \mathcal{P}_3$. That is, for any probability $p_i = p_{i+1} + q$ of getting any finite payoff x, it is better to get some larger payoff $y = x+b$ with a slightly lower probability p_{i+1}, which is what *Non-Timidity* says. Thus, those who wish to avoid fanaticism by rejecting *Non-Timidity* must either reject *Weak Dominance* or *Separability*; and given that *Weak Dominance* appears unimpeachable, they must therefore deny that $\mathcal{P}_1 + \mathcal{P}_3 \succ \mathcal{P}_2 + \mathcal{P}_3$ and reject *Separability*. But this comes with its own costs. Imagine that prospects $\mathcal{P}_1$ and $\mathcal{P}_2$ are different diets whose payoffs are additional years of happy life for you, whilst $\mathcal{P}_3$ is a prospect whose payoffs are additional years of happy life for an alien on a distant galaxy. In this context, denying *Separability* amounts to saying that which diet you should opt for depends on what is happening in this distant galaxy. Hence the name, 'the strange dependence argument'.

A more complex version of this argument supports *Non-Timidity**. Two things must be mentioned before we can show this. First, *Weak Dominance* must be slightly strengthened to say that for any payoff, it is better to get some larger payoff with *at least as high* a probability:

> *Dominance*. For any probabilities p, q such that $p \geq q$, and any finite payoff y, there is some finite payoff $x > y$ such that $p * x \succ q * y$.

So, *Dominance* entails that for any payoff and any probability of receiving that payoff, getting *some* sufficiently larger payoff with *the same* probability is better. By contrast, *Weak Dominance* is silent on the matter. Given the plausibility of this claim, the move to *Dominance* does not weaken the ensuing argument in any significant way.

The second is a lemma showing that *Separability* and *Transitivity* entail *Weak Separability*, according to which:

> *Weak Separability*. For any prospects $\mathcal{P}_1$, $\mathcal{P}_2$, $\mathcal{P}_3$, $\mathcal{P}_4$, each with finitely many payoffs, if $\mathcal{P}_1 \succ \mathcal{P}_2$ and $\mathcal{P}_3 \succ \mathcal{P}_4$ then $\mathcal{P}_1 + \mathcal{P}_3 \succ \mathcal{P}_2 + \mathcal{P}_4$.

In other words, if we start with one prospect that is better than a second prospect, and we alter the better prospect in a more favourable way than the worse prospect, then the altered first prospect should still be better than the altered second prospect. We can see the entailment as follows. Suppose $\mathcal{P}_1 \succ \mathcal{P}_2$. By *Separability*, $\mathcal{P}_1 + \mathcal{P}_3 \succ \mathcal{P}_2 + \mathcal{P}_3$. Suppose $\mathcal{P}_3 \succ \mathcal{P}_4$. By *Separability*, $\mathcal{P}_2 + \mathcal{P}_3 \succ \mathcal{P}_2 + \mathcal{P}_4$. By *Transitivity*, $\mathcal{P}_1 + \mathcal{P}_3 \succ \mathcal{P}_2 + \mathcal{P}_4$.

To show that *Dominance* and *Separability* entail *Non-Timidity**, we must show that for any slight decrease in the size of a less likely payoff from x_i to y_i, there is some sufficiently large payoff y_j such that increasing the size of a more likely payoff from x_j to y_j outweighs the slight decrease in the less likely payoff.[11] Letting $y_i = x_i - a$ and $y_j = x_j - b$, this is to say that for any slight decrease a in a less likely payoff, there is some sufficiently large increase b in a more likely payoff that outweighs the slight decrease in the less likely payoff. Let b be such that $x_j + b > x_i + a$ and let δ be the difference between them (i.e. $x_j + b = x_i + a + \delta$). Now consider the prospects in Table 2.

[11] I will illustrate this for prospects with two non-zero payoffs, but the argument straightforwardly generalises to any prospects with a finite number of non-zero payoffs.

Prospect	p_j	p_i	$1 - p_j - p_i$
$\mathcal{P}_1$	$a + \delta/2$	0	0
$\mathcal{P}_2$	0	a	0
$\mathcal{P}_3$	$x_j + \delta/2$	0	0
$\mathcal{P}_4$	x_j	0	0
$\mathcal{P}_1 + \mathcal{P}_3$	$x_j + a + \delta = x_j + b$	0	0
$\mathcal{P}_2 + \mathcal{P}_4$	x_j	a	0
$\mathcal{P}_5$	0	$x_i - a$	0
$\mathcal{P}_1 + \mathcal{P}_3 + \mathcal{P}_5$	$x_j + b$	$x_i - a$	0
$\mathcal{P}_2 + \mathcal{P}_4 + \mathcal{P}_5$	x_j	x_i	0

Table 2

Since $p_j > p_i$ and $a + \delta/2 > a$, *Dominance* entails that $\mathcal{P}_1 \succ \mathcal{P}_2$. Moreover, since $x_j + \delta/2 > x_j$, and the probability of receiving a payoff given $\mathcal{P}_3$ or $\mathcal{P}_4$ is the same, *Dominance* entails that $\mathcal{P}_3 \succ \mathcal{P}_4$. Thus, by *Weak Separability*, $\mathcal{P}_1 + \mathcal{P}_3 \succ \mathcal{P}_2 + \mathcal{P}_4$. So, by *Separability*, $\mathcal{P}_1 + \mathcal{P}_3 + \mathcal{P}_5 \succ \mathcal{P}_2 + \mathcal{P}_4 + \mathcal{P}_5$. That is, for any finite payoffs x_i and x_j with probabilities $p_i < p_j$, there is some finite payoff $y_j = x_j + b$ such that getting y_j with probability p_j and $y_i = x_i - a$ with probability p_i is better than getting x_j with probability p_j and x_i with probability p_i, which is what *Non-Timidity** states. Thus, those who wish to reject *Non-Timidity** must either reject *Dominance* or *Separability*; and given that *Dominance* seems no less unimpeachable than *Weak Dominance*, they are therefore committed to denying *Separability* and admitting the strange dependence.