Jam Tomorrow and the New Repugnant Conclusion: Puzzles for Longtermism

Max Khan Hayward

Abstract

Longtermists claim that we ought to prioritise projects whose primary beneficiaries will be the inhabitants of the far future, rather than those living in our era. This creates a puzzle. If present people should sacrifice their interests on behalf of future people, should not future people likewise sacrifice their interests on behalf of others living still further in the future? But if this continues indefinitely, all these sacrifices are pointless. Longtermists often justify their

proposals in broadly Act-Utilitarian terms. The puzzle just sketched generates a paradox for an Act-Utilitarian account of moral reasons – the first *Jam Tomorrow* paradox. Still, indefinite deferral would require conditions which are unlikely to obtain. Yet these conditions are precisely outcomes which Longtermists aim to realise. This is the second *Jam Tomorrow* paradox. Variants of Act-Utilitarianism which evade the paradoxes fail to justify Longtermist projects. Indeed, *Total* Utilitarianism, which many Longtermists endorse, faces a worse version of the paradox – the *New Repugnant Conclusion.*

1. Introduction

"The rule is, jam tomorrow, ~~and jam yesterday,~~ but never jam today."
"It must sometimes come to 'jam today'," Alice objected.
"No, it can't," said the queen.
- Lewis Carroll, *Through the Looking Glass*, (with my emendations)

Longtermists think we should be impartial between the interests of those who live today and those who will live in the future. This leads them to endorse striking conclusions about which projects we are morally required to pursue. Because the sentient beings who may exist across the vast expanses of the far future *massively outnumber* those alive in the near-to-medium future, Longtermists claim that we ought to devote the bulk of our efforts and resources to projects most of whose beneficiaries will live far in the future, rather than in our era. As Greaves and MacAskill put it:

> We believe that strong longtermism[1] is of the utmost importance: that if society came to adopt the views we defend in this paper, much of what we prioritise in the world today would change. (Greaves & MacAskill 2021)

But this immediately presents a puzzle. If people *today* ought to accept great sacrifices in order to benefit the inhabitants of the far future, must it not also be the case that the inhabitants of that future – with their even greater technological powers and foresight – should sacrifice *their* interests for the sake of inhabitants of the still further future, and so on indefinitely?[2] But if that happens, what is the point of such sacrifices?

Longtermists also tend to endorse a broadly Act-Utilitarian account of moral reasons. It might seem that this is the best moral theory to justify Longtermism's distinctive practical conclusions.[3] I argue that this appearance is false. When combined with a structurally Act-Utilitarian account of moral reasons, the puzzle just sketched creates unacceptable paradoxes. These paradoxes – the *Jam Tomorrow Paradoxes* and *The New Repugnant Conclusion* – arise from the problem of deferring gratification in pursuit of greater payoffs over an indefinite time-horizon.

To escape these paradoxes, we must either abandon a Utilitarian account of moral reasons entirely or adopt a form of Utilitarianism that does not clearly support Longtermist conclusions. In this paper, I do not advocate a particular solution. Nevertheless, I think my arguments show that Longtermists must look beyond standard Utilitarian moral theories to

[1] Greaves & MacAskill 2021 call their view "strong longtermism", both to contrast with the common of "long-term" in contemporary political contexts to refer to policies that only look to the relatively near future, and to underscore their claim that considerations of the long-term are the most morally important consideration we face. In this paper, I use "Longtermism" without qualification to refer to the movement of which their view is representative.

[2] Longtermists debate whether we are living at the "Hinge of History" (Parfit 2011, 616): that we currently face far greater risks of extinction than past or future people, such that considerations of the long-term are especially pressing today. The issue remains open (Ord 2020 thinks we are; MacAskill 2022b argues against; Mogensen 2024 critiques MacAskill), and seems hard to resolve, but even if "existential risk" reduces in the future, that doesn't mean that future generations won't still have opportunities to sacrifice their interests for the sake of greater benefits accruing to those to come.

[3] Mogensen 2021 provides a classic argument for the connection between Act-Utilitarianism and Longtermist priorities.

justify their conclusions. How deep this divorce must be is a question for further work.

2. Longtermism and Act-Utilitarianism

Not all Longtermists are Act-Utilitarians. Nevertheless, Act-Utilitarianism is the intellectual wellspring and guiding moral orientation of Longtermism and its sister movement, Effective Altruism. Where Longtermists and Effective Altruists depart from strict Act-Utilitarianism, they often do so by adopting a form of deontology that can be characterised as Act-Utilitarianism with *side constraints* – restrictions on utility-maximisation that require choices to respect rights, justice, and so on.[4] I will be discussing any theory whose account of our moral reasons to benefit others is structurally Act-Utilitarian, whether or not the theory also accepts deontological elements, such as side-constraints – these will not be relevant for my argument.

What explains the connection between Act-Utilitarianism and Longtermism? Longtermists start with the observation that the number of sentient beings who will live throughout the duration of the extended future *utterly dwarfs* the number of those living today and in the near- and medium-term future, so long as we prevent catastrophic events that could wipe out sentient life. Moreover, they accept:

> *Axiological Temporal Impartiality*: The value of all subjects' welfare is equally intrinsically important, no matter what position in time they occupy.

Longtermists make two further claims. The first is that people living today *can significantly affect* how good the world will be in the far future (Greaves & MacAskill 2021; MacAskill 2022a). Since there will (or could) be so many people living in the far future, *our opportunities for promoting welfare* are strongly skewed towards the far future. Still, that would have no useable practical implications if we were entirely *clueless* about *which* actions have such beneficial effects (Lenman 2000). Thus, Longtermists also claim that we are not *so*

[4] Berkey 2021 gives an excellent overview of such deontological views in the context of Effective Altruism. Likewise, Greaves and MacAskill offer an argument for Longtermism based on a deontological view on which "when the axiological stakes are very high, there are no serious side-constraints, and the personal prerogatives are comparatively minor, one ought to choose the near-best option,"(2021, p27) arguing that this constraints indeed hold when considering Longtermist proposals. In other words, at least in the context in which Longtermist questions arise, our moral reasons are act-utilitarian in structure. Similarly, Bostrom 2003 argues that "mixed" non-Utilitarian views should accept his conclusions, claiming that, if they permit some moral reasons generated by future interests, then "so long as the evaluation function is aggregative (does not count one person's welfare for less just because there are many other persons in existence who also enjoy happy lives) and is not relativized to a particular point in time (no time-discounting), the conclusion will hold," (p.310).

clueless about the long-term effects of present actions that rational evaluation is impossible (Greaves 2016).

Axiological claims state what is good or valuable, not what any agent ought to do. So *Axiological Temporal Impartiality*, even supplemented with the further premises just mentioned, does not entail the Longtermist view we have *most reason* to prioritise projects affecting the far future. Even bracketing deontological side-constraints, facts about the intrinsic value of an action's outcomes do not *entail* conclusions about the strength of reasons to perform that action. For example, Relational Ethical[5] views allow for *partiality*: agents may have *stronger* reasons to benefit those they have relationships with than to benefit strangers, even if the welfare of both is equally intrinsically valuable.

The distinctive Act-Utilitarian contribution to the Longtermist argument is the claim that an agent's moral reasons are determined solely by the *quantity* of intrinsic value created by the various actions open to her (bracketing side-constraints). Moral reasons are *agent-neutral.* Every agent has most reason to perform whichever available action maximises welfare, regardless of the relationship between the agent in question and the beneficiary of their action. This claim, combined with *Axiological Temporal Impartiality*, *does* imply temporal impartiality about reasons. Not only do the untold billions who occupy the far future intrinsically matter as much as those who live today, but their interests generate reasons in just the same way as do the interests of those living today. Since the former immeasurably outnumber the latter, our reasons to promote welfare skew sharply towards actions that will primarily benefit those in the far future.

As Railton says (1984), Act-Utilitarianism is a theory of *objective reasons for action*, *not* of how agents should deliberate.[6] It may not always be best for agents to *aim* to promote utility impartially, or try to *calculate* how to do so. But however agents deliberate, what they have most objective moral reason to *do* is to promote utility impartially.

We can now see why the Act-Utilitarian account of the structure of objective moral reasons seem so congenial to Longtermist conclusions. If an agent's reasons precisely track the amount of intrinsic value created by each action open to her, with no other weighting factors

[5] See Peter (forthcoming) for an excellent discussion of contemporary Relational ethical views.

[6] While this view has become orthodox among Act-Utilitarians seeking to respond to self-effacingness worries, Longtermists do sometimes portray their broadly Act-Utilitarian arguments as a subjective decision-theory (eg Greaves & MacAskill 2021). However, they do not argue that agents must always explicitly deliberate in Act-Utilitarian terms. Presumably, they think that explicit Act-Utilitarian deliberation is in fact optimific in the context of Longterm planning, so their view is compatible with this objective formulation of Act-Utilitarianism.

such as relationships or temporal or physical distance, Longtermists can argue:

> 1) The value of all subjects' welfare is equally intrinsically important, whatever position in time they occupy.
> 2) The number of welfare subjects who will exist in the far future utterly dwarfs the number of welfare subjects alive in the near-to-medium-term future.
> 3) There are (identifiable) actions open to us now, that will significantly affect the welfare of those who will live in the far future.
> 4) Each agent has most objective moral reason to perform the action, out of those open to her, which creates the most intrinsic value.
> *Therefore:* The actions we have most reason to perform are those whose beneficiaries will mostly exist in the far future, rather than the near-to-medium term future.

Premise 4 is the distinctive contribution of the Act-Utilitarian view of objective moral reasons, and is required for the argument to be valid. That is why a broadly Act-Utilitarian view seems so congenial to Longtermism.

I'm now going to tell a story that should make us doubt whether an Act-Utilitarian account of objective moral reasons is internally coherent when we consider the problem of promoting welfare over the long-term future.

3. Jam Tomorrow

After the near-misses of the second Trump Administration – the nuclear standoff with Russia, the SpaceX catastrophe, and the tense months of the Anthropic Alignment scare – the tide in policy circles swung decisively in favour of Longtermism. No longer merely exhorting change from the confines of academia, or bending the ears of a few sympathetic philanthropists, Longtermists found a place at the heart of the American government with the establishment of the Department for Longterm Planning. The Department's independence and funding were irreversibly secured with a Constitutional Amendment, and staff were free from oversight or risk of termination by elected politicians. Within the department there were teams dealing with existential risk, AI alignment, pandemic preparedness, and other projects.

Sam headed up the Energy Planning Unit. While AI researchers were constantly innovating new, socially-valuable uses for LLM-based technologies, these required large amounts of energy. The EPU had recently placed an order of 10th Generation Solar panels for a vast new solar farm that would provide additional energy capacity for the AI division's projects. These would be ready for delivery in one year and would last 10 years before needing to be replaced.

However, Sam's colleague, Sasha, pointed out to him that researchers at Cambridge had recently made a breakthrough in solar panel optimisation. Gen 11 panels would be significantly more efficient, and last 11 years. Although the EPU had made a down-payment on the Gen 10 panels, the constant reduction in solar panel prices meant that the same number of Gen 11 panels could be purchased with the remaining allocated funds. However, they would take one extra year to deliver. Never one for sunk costs thinking, Sam concluded that he ought to change his order. The services powered by the farm would be beneficial, but they were not urgent, and the interests of those living two to thirteen years hence were not less important than the interests of those living one to eleven years from now. The choice was simply a matter of trading a smaller gain in the near future for a greater gain in the slightly-more-distant future.

A few months later, Sasha came to Sam with more exciting news. A spinoff from MIT had made a further breakthrough. Gen *12* panels would last 12 years, and would be cheaper and even more efficient than Gen 11 panels. But they would not be ready until 3 years hence. Still, consistency compelled Sam to order the switch – why accept a smaller benefit in two years when something better could be had in three years for the same price?

It was another few months later that Sasha came to Sam with yet more good news. A further breakthrough from a Chinese manufacturer would provide Gen 12 panels with a novel built-in battery-storage functionality, allowing the plant to supply clean energy when the sun didn't shine, for a lower cost than the EPU's original supplier. The problem struck Sasha and Sam at the same time. The march of technology wasn't going to stop any time soon. Indeed, the safer bet seemed to be that it would only accelerate. Gen 12 panels would soon be superseded by Gen 13, energy storage would continue to improve, one day the renewable energy technologies they knew would be obsolete. Freed from short-term thinking, Sam and Sasha – and those who came after them – would always be faced with new opportunities to forfeit lesser near-term gain for greater long-term payoffs. But when would it ever end? Would rational planners like Sam and Sasha ever have good reason *not* to defer gratification? Peering into the depths of the limitless future, would it not always make sense for temporally-impartial altruistic planners to continue sacrificing the present for greater benefits in the future? But if *no one* was ever going to cash in on their investments, then. . . what was the point? If deferred gratification would always be morally rational, and if planners always did what morality required, then no one would ever get any gratification. But, in that case, why ever bother deferring?

4. The Paradox

Sam and Sasha are thinking like temporally-impartial Act-Utilitarians – that is, like orthodox Longtermists engaged in long-term planning. They give no greater weight to the interests of people in the far future than to those alive today, and devote their resources accordingly.

But nothing in the story hinges on whether they *think* in Act-Utilitarian terms – this is not an argument against Act-Utilitarianism as a subjective decision-guide. All that matters is that they *actually choose* the options that seem to be required by their objective moral reasons. If such deferring choices are in fact what Act-Utilitarian reasons demand, then the end result will fail to maximise utility if three further conditions obtain:

> a) Their work will be continued in future by other Longtermists.
> b) There will, indefinitely, continue to be opportunities to trade short-term gratification for greater benefits further down the line.
> c) There will, indefinitely, continue to be welfare subjects who can be benefitted.

Together, these assumptions seem to lead to the conclusion that no one will ever enjoy the benefits for which they and their successors are sacrificing their respective presents, for it will always be rational to defer gratification *for one more year* to gain a greater benefit. This is not just morally regrettable, but seemingly contradictory – for if no-one will ever cash in on the investments of the past, then the utility gained by deferring is *nil.*

This gives us the first version of the *Jam Tomorrow* paradox:

> *Jam Tomorrow 1:*
> 1.1. It is morally rational to make a deferring trade-off, in which we sacrifice the option of a lesser present benefit for the option of a greater future benefit.
> 1.2. If it is morally rational to make a deferring trade-off in any individual case, then it is rational to make any member of an infinite series of identical trade-offs.
> 1.3. If any deferring trade-off is a member of such an infinite series, then its expected gain is negative, so it is not morally rational.

If our tomorrows are endless, if the production of jam continues always to improve, and if choices are always made by rational planners, then the consumption of jam will *always* be deferred until tomorrow. But this is pointless, because you cannot eat jam that will only ever appear tomorrow.

My argument is *not* that Act-Utilitarianism endorses this absurd conclusion. Rather, it faces a paradox. If Sam and Sasha and their counterparts down the ages continue to defer gratification, then each of them has traded something for nothing. Each of them had it in their power to bring about a world with more utility, and failed to do so. Act-Utilitarians cannot accept that, in so doing, each planner did what she objectively had most reason to do. But they cannot explain how any of them had objective moral reason to do otherwise.

Why is this? As mentioned above, Act-Utilitarians endorse a principle of *Agent-Neutrality.* This implies:

> If A has most reason to ϕ in circumstances C, then B has most reason to ϕ in circumstances C*, unless there is some relevant difference between C and C*.

What differences are relevant? If B does not have the option to ϕ in C*, then B does not have most reason to ϕ. Or C* might differ from C such that ϕing is not the option open to B which creates the greatest marginal utility compared to B's other options, even though ϕing is the option open to A which creates the greatest marginal utility compared to A's other options.

If conditions a) – c) obtain, then there may be no such difference between the situation of Sam and Sasha and the situations faced by their successors. So, if Sam and Sasha have most reason to defer gratification, then their successors, who are identically-situated in the relevant sense, also have most reason to defer. If one of their identically-situated successors does *not* have most reason to defer, then it must be that Sam and Sasha and all their other counterparts also do not have most reason to defer. Of course, so long as *at least one* future planner refuses to defer, and cashes in on the investment, then each previous planner did what they had objectively most reason to do since every prior case of deferral increases the eventual payoff. But it cannot be that this future planner actually has most reason to terminate the series, unless there is some relevant difference between her and those who came before her. After all, it would always be better to wait just one more year before cashing in.

Can Act-Utilitarians avoid the paradox in a world where a)-c) obtain, by rejecting one of the claims 1.1. – 1.3.? For now, I'll assume that Act-Utilitarians will not reject 1.3., claiming that each agent did what she had most objective moral reason to do, even though the net result was a massive utility loss.[7] But, given Agent-Neutrality, if Act-Utilitarians want to claim that *any* of the agents in an infinite series fails to do what she has most reason to do, they must claim that *all* of them fail to do what they have most reason to do, since they there is no relevant difference between them. Moreover, since agents would be doing what they have most reason to do by deferring in any *finite* series of trade-offs, it seems they must then claim that *some* agent has most reason to terminate such a series, in order to prevent it becoming infinite. Thus, Act-Utilitarians need to show how their account of reasons can generate the following three claims:

> i) In accepting a deferring trade-off, an agent may be doing what she has most reason to do, if the series of deferrals is finite.
> ii) In accepting a deferring trade-off, an agent is not doing what she has most to do, if the series of deferrals is infinite.
> iii) In a series of deferring trade-offs, there is always an agent who has most reason to terminate the series by rejecting the opportunity to defer (even if there is no relevant difference between her and her predecessors who did have most reason to defer).

We can see how hard it is to meet these conditions by considering how Act-Utilitarians might attempt to reject the other two premises of the *Jam Tomorrow 1* paradox.

They might deny 1.1.. Act-Utilitarianism entails that it is always morally correct to exchange a smaller near-term utility payoff for a greater utility payoff in the further future. But 1.1. speaks not of utility payoffs, but of *options* to obtain a utility payoff. For such an option to yield utility, it needs to be *taken*, not merely available. If the option to attain a greater utility *will not in fact be taken*, there is no reason to forego a lesser payoff to acquire the option. So an Act-Utilitarian might instead claim:

> 1.1.* It is morally rational to make a deferring trade-off, in which we sacrifice the option of a lesser present benefit for the option of attaining a greater future benefit, if and only if the latter option will itself be taken.

But this principle is implausible. True, agents should not sacrifice the option for a short-term

[7] In this section, I assume Actual Act-Utilitarianism for ease of exposition, which claims that whether or not an agent has acted rightly depends on whether or not she in fact brought about the most utility. In the next section, I discuss Expected Act-Utilitarianism and show how the paradox applies there too. Note that Actual Act-Utilitarians can agree that expected utility is what should guide us in assessing plans and evaluating the actions of others, even though the ultimate moral goal is to maximise actual utility.

benefit to gain the option of a greater future benefit *in cases where the greater future benefit will irrationally be left untaken* – for example, because a future counterpart fails to take the option due to akrasia, sloth, or perversity. But Sam and Sasha's story is nothing like this. Their worry is that their morally-rational counterparts will forego their options to acquire utility *only in order to trade them for still better options.* It is hard to see how utilitarians can find anything to complain of here. There is no plausible principle that a present utility cannot be traded for options for greater utility *unless that the utility will be realised at some determinate point in the future.*

To see this, imagine a series of arbitrary Days and arbitrary units of Benefit (B). If:

> I) It is morally rational to forego 1B on Day 1 in exchange for 2B on Day 2, *where I will take 2B on Day 2.*

then it must also, by parity, be the case that:

> II) It is morally rational to forego 2B on Day 2 for 4B on Day 3, *where I will take the 4B on Day 3.*

But *axiological temporal neutrality* implies that:

> III) 4B *that I take* on Day 3 is better than 2B *that I take* on Day 2, and 2B *that I take* on Day 2 is better than 1B *that I take* on Day 1.

By transitivity, it follows that:

> IV) 4B *that I take* on Day 3 is better than 1B *that I take* on Day 1.

1.1.* says that I should sacrifice 1B on Day 1 for 2B on Day 2 *only if* I will take 2B on Day 2, but if reasons perfectly track the goodness of outcomes, as Act-Utilitarians claim, so that agents always have objective moral reason to choose a better outcome over a worse, IV implies that:

> V) It is morally rational to sacrifice 1B that I take on Day 1 for 4B that I take on Day 3.

and since Act-Utilitarians accept standard principles of means-ends rationality, and in this case:

> VI) The only way to take 4B on Day 3 is to have the option to take 4B on Day

> 3, and the only way get from having the option to take 1B on Day 1 to having the option to take 4B on Day 3 is to sacrifice the option to take 1B on Day 1 for the option for 2B on Day 2, and to sacrifice the option to take 2B on Day 2 for the option to take 4B on Day 3.

this straightforwardly entails that:

> VII) It is morally rational to sacrifice the option to take 1B on Day 1 for the option to take 2B on Day 2, *in order to trade this option for the option to take 4B on Day 3.*

In other words:

> VIII) It is morally rational to sacrifice the option to take 1B on Day 1 for the option to take 2B on Day 2, *even if I do not foresee that I will take the 2B on Day 2.*

Thus, Act-Utilitarianism entails the falsity of 1.1.*. And a parallel argument can be constructed for any analogue of 1.1.* claiming that an agent should sacrifice a benefit today for the option of a future benefit only if someone will realise the benefit at some determinate point in the future. So Act-Utilitarians cannot generate a moral distinction between finite series of deferrals and infinite ones by pressing on the distinction between options and realised utilities.

Act-Utilitarians might attempt to formulate a principle that *directly* prohibits infinite series of deferrals. They might say that it is always morally rational to trade-off the option of a smaller near-term utility payoff for the option of a greater longer-term utility payoff *where that trade-off forms part of a finite series,* a deferring trade-off is *not* morally rational *where that trade-off forms part of an infinite series.* In other words, they might attempt to deny 1.2. After all, it is will-known that infinities pose special problems for Act-Utilitarianism,[8] and it's only if a series of deferring trade-offs continues infinitely that deferring has negative utility.

[8] The main worry in the literature is that in a universe with an infinite population, agents may face choices between actions that each produce infinite welfare. The aggregate marginal utility resulting from each will be the same (i.e. infinite), even if one results in a world with greater cumulative amounts of welfare than the others at each point in time. Orthodox Act-Utilitarianism seemingly fails to discriminate between these options. See Nelson & Garcia (1994) for this challenge; classic responses include Vallentyne (1993), Lauwers & Vallentyne (2004) & Mulgan (2002), and more recently Hong & Russell (forthcoming) and Francis (ms). The Jam Tomorrow paradox of infinite deferral is distinct from this and survives even if there is a solution to the familiar problem of choosing between non-equivalent infinite utility gains.

If performing an infinite series of deferring trade-offs were a single action performed by a single agent, Act-Utilitarians could explain why this agent did not have most moral reason to act thus. But, of course, that is not the case. Act-Utilitarians don't just need to claim that performing an infinite series of deferring trade-offs is morally bad; they need to claim that when a series of agents perform a succession of deferring trade-offs that constitute an infinite series, *none* of them did what she had most reason to do, *even though* they would have most reason to perform each of a succession of otherwise-identical deferring trade-offs that constitute a *finite* series. Act-Utilitarians who accept an Eternalist metaphysics might accept this, and say that even if deferring trade-offs (*finite) are intrinsically identical to deferring trade-offs (*infinite), there is still a fact of the matter as to whether a series of deferring trade-offs will terminate or not. If the series fails to terminate, then each agent acted objectively wrongly in deferring.

However, even this only gives us claim ii) – a prohibition on performing trade-offs that form part of an infinite series. But it's not enough just to evaluate agents' actions from the perspective of eternity, as though it were simply a brute fact that no-one chose to terminate the series. If morality requires that agents *do* perform each member of a finite series of deferring trade-offs, per i), but *do not* perform each member of an infinite series of deferring trade-offs, then it must require iii), that some agent terminate a series of trade-offs. A prohibition on performing an infinite series of trade-offs does not explain how anyone has most reason to *terminate* an ongoing series, since there is no point at which a series changes from being finite to infinite.

I do not see how an Act-Utilitarian theory of reasons can explain why agents must, at some point, have a reason to cease deferring jam for tomorrow. But then they must either reject finite deferrals of jam, which exchange less for more, or endorse endless postponement of jam, which fruitlessly exchanges something for nothing. And they cannot accept either of these conclusions.

5. An Irrelevant Problem?

You might think that this is a relatively unimportant puzzle, since the undesirable result – the endless postponement of gratification – would only follow from Sam and Sasha's initial decisions to defer so long as further conditions obtain. To recap, these are:

a) Their work will be continued in future by other Longtermists.

b) There will, indefinitely, continue to be opportunities to trade short-term gratification for greater benefits further down the line.

c) There will, indefinitely, continue to be welfare subjects who can be benefitted.

But a) - c) are rather *outré* possibilities. If Sam and Sasha foresee that some of their successors will *not* be rational Longtermists, that the march of technological progress will slow or stop, or that the future of sentient life will end, they can cease worrying about the *Jam Tomorrow* paradox. Given the likelihood that at least one of these will someday obtain, Sam and Sasha need not greatly fear paradox in sacrificing the present – some people, somewhere in the future, will most likely benefit from their investments, for the series of trade-offs will probably come to an end eventually.

But this should be cold comfort for Longtermists. *But for the Jam Tomorrow problem*, it would *not* be a bad thing if the planners of tomorrow were morally and rationally perfect, or if technology continued to improve, or if the future of sentient beings continued indefinitely. On the contrary, these are precisely the goals that Longtermists strive so hard to bring about! Preventing the extinction of sentient life is, if anything, the central goal of Longtermism.[9] Hence the Longtermist focus on identifying and preventing "Existential Risks" (e.g. Bostrom 2002, 2003; Ord 2020). Longtermists think we should devote considerable resources to colonising space because, at some point in the future, the Earth will become uninhabitable. Some look beyond this horizon, and aim to foster the development of digital technologies that will allow us to "upload" our consciousnesses to a virtual world, freed from the perils of embodied existence.[10] Some even dream of computing technology based on space-time crystals (Li *et al* 2012) that could function in a low-energy cosmos, allowing conscious beings to survive the "heat death" of the universe.[11]

Allied to this is a dedication to promoting technological progress (Bostrom 2003). Longtermists think that promoting the long-term welfare of sentient beings requires us to invest resources in artificial intelligence, quantum computing, transhumanist enhancement, space exploration and so on. And, of course, Longtermists want to have their goals *implemented* – which means that they would like to ensure that present and future decision-makers (be they government planners, billionaires or corporate leaders) make choices

[9] "The early extinction of the human race would be a truly enormous tragedy" (MacAskill 2022a, p176); "Premature human extinction would be astronomically bad." (Greaves & MacAskill 2021, p11).

[10] A classic case for both space colonisation and the creation of "non-biological" conscious agents is Bostrom 2003.

[11] See, for example, §2 of this discussion in one of the major Effective Altruism / Longtermist online communities: https://www.lesswrong.com/posts/7nH7R9eqYWSPvF4Qm/immortality-a-practical-guide

which are in line with Longtermist principles.[12] These are the projects which Sam and Sasha's colleagues in other teams at the Department for Longterm Planning would be working on. But it is only conditional on the *failure* of these paradigmatic Longtermist projects that Sam and Sasha's choice to accept deferring trade-offs would not lead to the endless, and pointless, deferral of gratification. They defer because they are Longtermists, but if they hope to escape the disastrous implications of the *Jam Tomorrow 1* paradox, then they must, in some sense, be relieved to foresee that the central projects of Longtermism will probably fail. And that seems no less paradoxical.

So we can state a second version of the paradox:

> *Jam Tomorrow 2:*
> 2.1. It is good if agents make deferring trade-offs that form part of a finite series, but bad if agents make deferring trade-offs that form part of an infinite series.
> 2.2. An agent's deferring trade-off forms part of a finite series only if one of the following fails to obtain:
> i) her successors are perfectly morally rational.
> ii) her successors get the opportunity to make further deferring trade-offs.
> 2.3. It would not be bad if i) and ii) were to obtain.

Thus the Act-Utilitarian account of objective moral reasons generates a paradox of axiology. It is good that Sam and Sasha perform deferring trade-offs,[13] so long as they do not form an infinite series, since by so doing they exchange a smaller benefit for a greater one. But in order for any such series to be finite, someone must terminate it. Since an Act-Utilitarian account of reasons cannot generate the result that anyone will have most objective moral reason to terminate a series of deferring trade-offs, then for a series to be terminated it must either be the case that some future planner does not get the opportunity to keep deferring, or that she selects an option that she does not have most objective moral reason to select. But if it is good that Sam and Sasha get opportunities to perform deferring trade-offs and act on those opportunities, then it cannot be bad that their successors get such opportunities and act on them. So Sam and Sasha's decision is good only if something else, that is not

[12] Chapters 3 & 4 of MacAskill 2022a focus on affecting future values to bring them more into line with Longtermist goals: "changing values has particularly great significance from a longterm perspective" (p57). Likewise, MacAskill 2022b argues that a core Longtermist project is "using one's time to grow the number of people who are also impartial altruists." (p335 fn13).

[13] Many Longtermists explicitly endorse the view that "rather than spending now, society could save its money for a later time" (Greaves & MacAskill 2022 p.15; see also MacAskill 2023), for example because future people will likely be in a better position to achieve long-term benefits with the resources we lead to them.

good, occurs further down the line.

There's no paradox *in general* in saying that sometimes unfortunate occurrences make beneficial results possible. The problem is that the costs, in this case, are *unnecessary* for the realisation of the benefits. Nothing *forces* future planners to keep deferring gratification endlessly, just because sentient life continues forever and there continue to be opportunities to exchange smaller near-term benefits for the option of greater future payoffs. Planners could always just choose to cash in on the investments of their forebears. It is only if future planners always choose to accept trade-offs that the boon of an endless future, full of opportunities to trade less for more, implies the disaster of infinite deferral. *But for the unforced choices of Longtermist planners*, nothing about an endless future of technological improvement precludes the ample enjoyment of jam.

This is why the paradox arises even if we do not stipulate that the future of sentient life is *actually* infinite, so long as it is *indefinite*, in the sense that it is always *possible* that sentient life survives for a further generation. Normally it is precisely a view to the future that rationalises deferring gratification. At each point in time, Longtermists think it good that sentient life continues for a further generation. At each point in time, Longtermists think it good to sacrifice a smaller benefit in the present for a greater benefit which the next generation can enjoy. Yet it only takes iterations of these good things for sentient life to persist forever whilst never enjoying jam.

This also shows why Act-Utilitarians cannot avoid *Jam Tomorrow 1* by defining objective moral reasons in terms of *expected utility*. Expected-Act-Utilitarians might reject 1.3., and claim that Sam and Sasha and their successors do what they have most objective moral reason to do *even if the series of deferring choices does in fact go on forever*, since a)-c) were objectively improbable, so the series was *unlikely* to continue eternally. As Doody (2022) points out, Expected-Act-Utilitarians accept that objectively rational choices may lead to disaster, if the disaster was sufficiently improbable at the time of choice. But Longtermists are precisely working to bring about a)-c), and thus to make them more probable. If they succeed, the *Jam Tomorrow 1* paradox arises even on Expected Act-Utilitarianism; if they think it *good* that a)-c) obtain, but *bad* that deferral continues forever, they face *Jam Tomorrow 2*.

Thus, the *Jam Tomorrow 2* paradox shows that there is a contradiction between the Act-Utilitarian account of reasons and the Longtermist's temporally-impartial welfarist axiology, even if the future of sentient life does not actually end up being infinite. To avoid this contradiction, Longtermists must accept one of the following:

> A) It is sometimes better that agents to fail to do what they have most objective moral reason to do, rather than doing what they have most reason to do.
> B) The standard Act-Utilitarian account of objective moral reasons is incorrect.
> C) It is not bad if sentient life comes to an end in the future.

I'm going to argue that none of these options should be palatable to Longtermists.

6. Rational Irrationality vs Moral Immorality

According to the Act-Utilitarian account of reasons, if Sam and Sasha have most reason to defer, then so do each of their successors. Yet, for a series of deferring trade-offs to yield a benefit, it must end. This requires one of their successors to do something other than what she has most reason to do. Can Act-Utilitarians accept A), and claim that sometimes it is better if agents do something other than what they have most reason to do?

Non-Consequentialist theories, which accept the doctrine of *The Priority of the Good over the Right*, accept that it sometimes leads to a better outcome if agents fail to do what they have most reason to do. The bystander on the bridge has decisive moral reason *not* to push the large man into the path of the trolley. But it would lead to a better outcome if he did. On such views, our objective moral reasons do not always support the action that leads to the best results.

But it is precisely *because* Non-Consequentialist theories say that agents are sometimes morally required to do something other than what is best that Utilitarians reject these views, accusing them of "rule worship" – putting moral principles above the interests of people. Act-Utilitarians accept the converse doctrine – *The Priority of the Good over the Right.* If an agent is choosing between action ϕ and action ψ, and ϕ has better consequences, then she has more reason to choose ϕ than ψ; if ψ has the best consequences, she has most reason to choose ψ. Thus, Act-Utilitarians seem committed to rejecting A) – what is morally best for an agent to do is never other than what they have most objective reason to do.

This might seem too strong. Theories of rational choice sometimes allow that there can be cases of "rational irrationality" – where it would be rational to make oneself, at some future point, choose in an irrational manner, precisely because so-doing would bring about the best consequences. Can Longtermist Act-Utilitarians borrow from this tradition to claim that it is sometimes morally best for someone to contravene their objective moral reasons?

In many cases of "rational irrationality" what is at stake is the agent's *subjective deliberation*. Parfit's (1984) case of the Robber shows that making myself incapable of rational deliberation can lead to better results. The Robber threatens to kill my family if I do not give him the code to the safe, but realising that I am incapable of understanding or responding to his threat, he should rationally abandon his attempt at extortion. But if the Robber nevertheless *will* slaughter my family if I don't accede to his demands, then do I indeed have most *objective* reason to give him the code. Likewise, the upshot of the "Paradox of Hedonism" and Railton's related "Paradox of Consequentialism" (1984) is not that agents ever *objectively* have most reason to perform suboptimal actions. Rather, achieving the best outcomes sometimes requires us to *subjectively deliberate* in a way that doesn't explicitly aim at maximisation. Moreover, due to the quirks of human psychology, there are important goods – such as relationships – that are only open to people who are disposed to think in ways that sometimes lead them to make objectively irrational choices. But Utilitarians should still hold that it would be *better* if human psychology were not like this, if the greatest benefits could be enjoyed by agents whose style of deliberation didn't predictably lead them to select objectively irrational, suboptimal options.

But in any case, nothing in the *Jam Tomorrow* problems hinges on how agents happen to deliberate. The paradoxes arise when agents *in fact* do what they have most objective reason to do, regardless of how they subjectively deliberate. Nor do the problems arise from quirks of human psychology, or the threats of malign agents. They arise from the very nature of following Act-Utilitarian objective moral reasons over indefinite time horizons. So these cases are of no help.

More relevant are the Paradoxes of Infinite Choice, such as Pollock's (1983) problem of EverBetter Wine or Artzenius, Elga and Hawthorne's (2004) case of Trump in Hell. In these cases, the norms of orthodox Decision Theory imply that agents are rationally required to select each of an infinite series of choices, whose aggregate effect is disastrous. Perhaps responses to these problems can solve the *Jam Tomorrow* paradoxes.

However, we cannot simply transpose solutions developed for a theory of strategic deliberation to a theory of objective moral reasons. In cases where infinitely repeating a certain type of choice leads to disaster, Artzenius, Elga and Hawthorne accept that *each choice is rational*. But when they call these choices "rational", they are speaking of the structural norms of Decision Theory. On their view, orthodox Bayesian principles of rationality are so theoretically well-supported that they are unwilling to reject them, even if they entail that rational choices can sometimes predictably lead to avoidable disasters. These principles are the best available guides for decision-making in ideal contexts, but they are not *perfect* guides to strategic choice – it simply turns out that no set of practical principles is perfect in this way.

But the Jam Tomorrow problem is stated in terms of *objective moral reasons.* It is disappointing, but coherent, to conclude that there is no general set of strategic principles that always lead agents to their strategic goals – that the best theory of strategic rationality gives imperfect advice. But it *is* incoherent to endorse *The Priority of The Good Over The Right* and also hold that agents sometimes have most objective moral reason to bring about a preventable moral disaster. Principles of strategic decision-making can be imperfect guides to strategic success. But theories of objective moral reasons are not supposed to be merely useful guides to achieving our goals – they aim to define what we objectively ought morally to do. The norms of what we objectively ought morally to do cannot be imperfect guides to what we objectively ought morally to do. The *Priority of The Good* claims that what is morally best is always what we ought morally to do (if we can). If one of an agent's options leads to better outcomes than alternatives, she must have more objective moral reason to do it. If an account of objective moral reasons fails to deliver this result, then it is not just an imperfect set of advice; it is a false moral theory.

For this reason, Artzenius, Elga and Hawthorne's (2004) solution for avoiding disaster in infinite choice scenarios cannot be transposed to the case of Sam and Sasha. They claim that, although it is rational for agents to make each choice in the series, it is also rational for agents at the start of the series to *bind*[14] those later in the series to act irrationally. It is rational every day of the week to exchange today's allocation of jam for more jam tomorrow; but it is also rational for the White Queen, on Monday, to do something which will *force* her to eat the jam on Friday.

But this cannot help the Act-Utilitarian. Given Agent-Neutrality, if Sasha and Sam have sufficient reason to bring it about that the deferral of gratification ceases in, say, 200 years,[15]

[14] There is a rich literature on binding and rational choice; the *locus classicus* is Elster (1979).

[15] Of course, no termination strategy will be optimal – for each point Sam and Sasha could pick to enforce cashing in, it would be better if they picked a later date. Or if they followed a "mixed strategy" and used a randomising machine to select the future-binding date, it would always be better if they employed an alternative mixed strategy that set termination at 1 year after the point selected by the machine, and so on (this argument follows following Pollock's 1983 discussion). Even if they overcame this challenge – for example, by adopting Landesman's 1995 moderate satisficing view (discussed in the next section) and simply picking a *good enough* date or strategy, Sabina would still have moral reason to accept great costs to alter

by binding some future planner (call her Sabina) *to cease accepting deferring trade-offs,* then Sabina must have sufficient reason to cease deferring *without needing to be bound.* If Sam and Sasha are morally justified in forcing Sabina to do it, Sabina is morally justified in doing it anyway.[16] In that case, binding is an unnecessary cost. But since Agent-Neutrality cannot allow that Sabina *does* have most reason to reject a deferring trade-off if Sam and Sasha have most reason to accept a relevantly-identical trade-off, then it cannot be the case that Sam and Sasha have most reason to bind Sabina.

So Act-Utilitarians must reject A). They can accept that there can be such a thing as rational irrationality, but they cannot accept that there is such a thing as moral immorality.

7. Tweaking Act-Utilitarianism

Perhaps a better option for the Longtermist is B) – to reject the orthodox Act-Utilitarian theory of objective moral reasons, in the hope that there is a nearby alternative which justifies Longtermist practical conclusions without generating the *Jam Tomorrow* paradoxes.

One possibility appeals to the notion of a *discount rate.* Longtermists sometimes attempt to adjust for the deep uncertainty infecting predictions about the far future by affording far-future utilities a lower weighting in deliberation than more proximate ones. But this is *not* a claim about our objective moral reasons. If only we could be more certain about the far future, we would have no reason to discount – the point of using a discount rate in our subjective deliberations is to help us to do what we have most objective moral reason to do. Indeed, the claim that far-future utilities are *objectively* as reason-giving as near-term ones – *Temporal Impartiality about Reasons* – is precisely the feature of Act-Utilitarianism that Longtermists appeal to in justifying their practical proposals![17]

But even if Utilitarians abandoned temporal impartiality, claiming that utilities give weaker *objective* moral reasons to agents the further they are from the moment of choice, this would not solve the *Jam Tomorrow* paradoxes. In the story of Sam and Sasha, near-term utilities

the binding mechanism so that termination would be deferred for (say) another 1000 years.

[16] Hayward (2024) makes a similar case, in the context of collective action problems, for why Act-Utilitarians who accept *Agent Neutrality* cannot endorse binding.

[17] Greaves & MacAskill, for example, say that "One feature of expected utilitarianism that is near-essential to our argument is a zero rate of pure time preference. With even a modest positive rate of pure time preference (as e.g. on "discounted utilitarian" axiologies), the argument would not go through. Our assumption of a zero rate, however, matches a consensus that is almost universal among moral philosophers…" (2021 pp18-19).

are *not* being sacrificed for *far*-future benefits. At each point, the trade-off only defers gratification for an extra year. Iterations of relatively near-term trade-offs suffice to reach the disaster of infinite deferral. We need only keep postponing jam until *tomorrow* to deny ourselves jam forever. A discount rate steep enough to prevent even such brief deferrals would not only rule out the far-future-oriented projects which Longtermists endorse, but even more moderate concern for the future.

Another option appeals to *satisficing*. Again, we need to be clear as to what this view is about. As a subjective decision-guide, it seems fine to say that we needn't always worry about *maximising* utility, so long as we promote utility *enough*. But our problem is not about subjective decision-making. Sometimes satisficing is stated as a view about moral *obligation*. We are not *obliged* to maximise the good, only to create enough good in the world. But I have not framed the paradoxes primarily in the language of obligation; I spoke of moral reasons. So long as Satisficers agree that there is always *more* reason to select options which create *more* utility than the alternatives, then my statements are compatible with their views (and those of Scalar Utilitarians, who eschew talk of obligation altogether).[18]

However, as Landesman (1995) suggests, paradoxes of infinite choice might motivate a more moderate role for satisficing in an Act-Utilitarian theory. Normally, agents have most reason to perform the act that maximises utility (or, if faced with multiple optimific options, to take one of them). But a maximising function can be undefined in an infinite domain, and so in some choice-contexts there is simply no such thing as an act that maximises utility. In such cases, the Act-Utilitarian might say that we have *sufficient* reason to perform any act that is good enough.

Yet even this modification does not solve our problem. If selecting a finite series of deferring trade-offs were a single action performed by a single agent, this view would justify avoiding infinite deferral. An infinite series is worthless, there is no *one* finite series that is maximally good, so the agent may morally choose any finite series that is good enough, even though, whichever series she chooses, there is a better one available. But the *Jam Tomorrow* story involves no such choice. Indeed, we can set up the problem without any reference to maximisation. At no point are Sam and Sasha or any of their successors attempting to perform the impossible action of selecting the maximally good finite series. Rather, each agent is simply deciding whether or not to accept a single deferring trade-off. This generates

[18] Indeed, once framed in terms of reasons, I am unsure what debates about the threshold of "obligation" amount to, since Act-Utilitarians already grant that agents needn't be blamed, or feel guilty, whenever they fail to maximise utility, since such acts of excessive recrimination would themselves rarely maximise utility.

a paradox for the moderate satisficing view just as for the maximising view. If the moderate view solved the problem, it would imply that some agent will have a reason to terminate the series. That implies that all other agents are choosing in the context of finite deferral. In all finite cases, it is better to make a deferring trade-off than to reject it. Thus, each agent has more reason to defer. But repeated iterations of such a choice is all it takes for deferral to last forever.

Thus, the moderate satisficing view, that agents are permitted to plump for any good enough option if there is *no* optimal choice available, will not solve the problem. No agent is in that situation. For any of Sam and Sasha's successors to be justified in terminating the series, we would need a more radical satisficing view which says that agents are justified in choosing inferior options *in general*, not only when asked to select a best option out of an infinite menu.

But this more radical satisficing view faces two problems. First, it still doesn't imply that anyone did anything *wrong* in a scenario where every agent chooses to defer, and no one ever gets any jam. Second, and more worryingly for Longtermists, it undermines the justification for their distinctive policies. If agents are justified in rejecting deferring trade-offs *in general*, then there is no pressure to prioritise policies, like space colonisation, which aim maximise benefits in the far future, even at the expense of the present. The radical satisficing view at best leaves us with the rather anodyne conclusion that what we owe to the future is to do something *good enough* for it. Familiar philanthropic goals, such as vaccine development and poverty alleviation, are surely *good enough* for us to be justified in choosing them.

Perhaps there are other modifications to the Act-Utilitarian theory of reasons which avoid the *Jam Tomorrow* paradoxes whilst still justifying Longtermist policies. But I am doubtful. The vista of infinite deferral that haunts Sam and Sasha is composed of steps which are individually so small that it is hard to see how they could be blocked by any Utilitarian view that did not also undermine Longtermism.

8. A Welcome End?

Had we but world enough and time,
This coyness, lady, were no crime[. . .]
But at my back I always hear
Time's wingèd chariot hurrying near[. . .]

Now let us sport us while we may,
And now, like amorous birds of prey,
Rather at once our time devour
Than languish in his slow-chapped power.
- Andrew Marvel, *To His Coy Mistress* (my excerpts)

This brings us to C). The disastrous conclusions of the Jam Tomorrow paradox needn't arise if the imminent demise of sentient life vitiated the very possibility of continued deferral – if we really are about to die, it's time to eat all remaining jam. And it seems unlikely that the future of sentient life will continue indefinitely. As Marvel's poem suggests, the prospect of an ending creates a certain urgency, prompting us (at least the poet hopes) to cease deferring our gratification to the future. But can we accept that the eventual demise of sentient life is *good*?

Longtermists are, as noted, committed to preventing the extinction of sentient life – it is the justification of many of their most distinctive and dramatic policy prescriptions. And, *but for the Jam Tomorrow Paradoxes,* it is hard to see why to fight for the preservation of sentient life should have any end-point. Even if Longtermists don't figure the *eternal* survival of sentient life as an explicit goal, nevertheless, at each point in time, we would expect Longtermists to *strive* for the continuation of (non-abject) sentient life into the future. So accepting C) seems deeply in tension with the practical projects of Longtermists.

However, *Utilitarians* don't *have* to think it morally good that sentient life continues forever. On some views – such as *Person-Affecting* versions of utilitarianism – the demise of sentient life would be an inherently neutral thing, since this need not, by itself, harm anyone (depending on how it came about). On that view, we ought to defer gratification for the benefit of future people *if it is true that there are going to be future people*, but there would be nothing innately regrettable about the fact that sentient life was doomed to end.[19] The Person-Affecting Act-Utilitarian is not, thus, committed to viewing the indefinite continuation of sentient life as a good thing, or striving to bring it about.

Still, this is only a partial dissolution of the *Jam Tomorrow 2* paradox. The argument showed that the Act-Utilitarian must see it as *good*, not merely *not bad,* that one of a)-c) not obtain. To say that the end of sentient life would be inherently neutral is very different from

[19] "On such a person-affecting view, human extinction would be bad only because it makes past or ongoing lives worse, not because it constitutes a loss of potential worthwhile lives." (Bostrom 2003, p312); "According to the spirit of a person-affecting approach, premature extinction is in itself at worst neutral." (Greaves & MacAskill 2021, p18).

welcoming it. And it would be just bizarre to welcome it *only in order to escape a paradox of rational choice.* After all, even if a)-c) obtained, we are not *forced* to do what Act-Utilitarian reasons demand. Future planners could, at any point, cease deferring and eat the jam, if they chose to. Why welcome the apocalypse just in order to make rationally-permissible a good outcome which we could bring about anyway? Why not simply to *reject* the proffered theory of moral reasons?

In any case, most Longtermists reject Person-Affecting views. The distinctive projects pursued by Longtermists aim to prevent the extinction of sentient life, not just to ensure that such lives will be good *if they happen to exist.* Accepting a Person-Affecting view would undercut the justification for these projects. Thus, most Longtermists accept *Total* Utilitarian views. Indeed, as I will now argue, Total Utilitarianism leads to a paradox even worse than *Jam Tomorrow.*

9. The New Repugnant Conclusion

Unlike Sam and Sasha, consigned to the rather dull Energy Planning Unit, Anusha held a coveted position in the Existential Risk division. As as a brilliant analyst and planner she soon rose to occupy a leading position in the team. And, in one sense, she was well-placed. For she had truly absorbed the import of *Total* Utilitarianism. She believed that moral duty called upon her – and everyone else – to do all they could to maximise the total utility realised throughout the future history of the universe. And this meant, as Parfit (1984, p453) had so perspicaciously pointed out, that the difference between a calamity which extinguished 99.9% of sentient life, and one which eradicated it all, was of far greater moral significance than the difference between no calamity at all and the death of 99.9% of thinking beings. For in that band of survivors lay the most important thing of all – the possibility of a future. The happiness of the billions upon billions who could descend from those survivors would make the vast loss of life – tragic as it was – pale into relative insignificance. Morally speaking, it was the survival of sentient life that mattered most of all.[20] This was why Anusha believed that existential risks – no matter how small they should be – were, by far, the most important concern facing humanity.

But, in another sense, her position was unfortunate. Anusha's tragedy was that she simply hated the work. She was brilliant at cool calculation, the careful weighting of risks – as

[20] "Even an extremely small reduction in extinction risk would have very high expected value." (Greaves and MacAskill 2021, p11).

an actuary of sentient life, she had few peers. But it gave her no warm glow. Anusha was a people-person at heart. Before she encountered Longtermism, she had planned to be a community organiser; but she had become convinced that this was an inefficient use of her days on earth, and so, morally reprehensible. Now she sat in front of a computer year in, year out, as her personal relationships faded away, hobbies neglected, passions unsated. She hated her life. But this, she reasoned, was of little importance. If she had but the smallest chance to make any slight difference to the possibility that life was preserved, that far outweighed the small matter of one woman's unhappiness.[21]

It was only natural, given Anusha's standing in the Department, that Sasha and Sam came to her with their problem. That was when the horror of her situation fully dawned on Anusha. For, if she was anomalous in her talents, this fact was not what made the difference in the grim choice between the interests of the myriad inhabitants of the endless future and the happiness of one individual. Any person could reason – should reason – that she ought to do all she could, to the point of leading a life worse than no life at all, to contribute anything to the cause of sentient survival. And this awful justification would never expire. As she explained to Sam and Sasha, the future of sentience could never be assured without ongoing work – all thinking beings who would ever live would hold that precious flame in their hands. And so, they too would be morally obliged to sacrifice everything they could to create even the smallest increase in the chance that it should stay alight. But if every person did all morality demanded, and surrendered every chance of joy to preserve the unbounded future, then what was it all for? Preserving life was only valuable if that life was *good.* But if each generation drove itself into misery to preserve the next, then no one would ever benefit from the sacrifice. The possibility of future joy would be bought at the price of a reality of unending misery. In a dark moment, Anusha wished that the future extinction of sentience were truly inevitable, for only then might morality permit someone – anyone – to live for today, rather than for tomorrow.

[21] "Even if we use the most conservative of these estimates, which entirely ignores the possibility of space colonisation and software minds, we find that the expected loss of an existential catastrophe is greater than the value of 10^{16} human lives. This implies that the expected value of reducing existential risk by a mere one millionth of one percentage point is at least a hundred times the value of a million human lives. The more technologically comprehensive estimate of 10^{54} human brain-emulation subjective life-years (or 10^{52} lives of ordinary length) makes the same point even more starkly. Even if we give this allegedly lower bound on the cumulative output potential of a technologically mature civilisation a mere 1 per cent chance of being correct, we find that the expected value of reducing existential risk by a mere one billionth of one billionth of one percentage point is worth a hundred billion times as much as a billion human lives." (Bostrom 2013, p18).

10. The Paradox

Anusha's predicament, the *New Repugnant Conclusion*, arises from her commitment to Total Utilitarianism: the view that what is morally valuable is the *total* amount of happiness that exists – morality's concern is not just to make people happy, but to make happy people.[22] On that view, the preservation of sentient life is of almost limitless value. Thus, every person, for all eternity, must be morally required to sacrifice *everything* to ensure almost *any* increase in the probability of sentient survival. But a history in which everyone follows this prescription is not just regrettable – it is, in Total Utilitarian terms, of *negative* value. In that sense, the New Repugnant Conclusion is worse than the Old – rather than a future where billions lead lives *barely* worth living, we face an eternity in which endless generations lead lives that *are not worth living at all.*

Of course, there is a contingent fact at work in this story too. In Anusha's case, the work of preserving humanity's future robs life of all its joy. But that needn't be true for everyone. Presumably, many Longtermists *enjoy* their work. So long as some people enjoy doing what Longtermism demands, then their work is not self-defeating, and no repugnance ensues.

But this is a surprising result for an Act-Utilitarian. For it shows that Total Utilitarians in a world of potentially-endless sentient life can avoid Repugnance only if either:

> I) Some people do not do what they have moral reason to do.
> II) For some people, reasons of self-interest align with the demands of moral reasons.

Act-Utilitarians think a virtue of their theory is that it gives optimific guidance in a world where *some* people do not do the right thing. But I) says that their theory *doesn't* give optimific guidance in a world where *everyone* does the right thing! Yet, as we've seen, Utilitarians must think it is always good that agents do what they have objective moral reasons to do.

As regards II), it has been characteristic of Utilitarianism after Sidgwick to *deny* that self-interest must align with morality – morality is supposed to trump prudence, vitiating Sidgwick's worries about the "Dualism of Practical Reason". Indeed, the claim that the demands of morality are *not* limited by the prudential burdens they place upon moral agents is a foundational motivation of Effective Altruism, and thence of Longtermism. It is not clear whether a moral view which accommodates II) would serve to justify the practical

[22] To reverse the Narveson's (1973, p90) famous formulation.

prescriptions of Longtermists.

11. Conclusion

I am not sure what to make of these arguments. Some negative claims seem clear to me. Endlessly deferring gratification is not morally rational. We cannot always sacrifice our lives today for the lives to come, no matter their number, for it is only in our todays that happiness is realised. Jam must not always be saved for tomorrow. Yet in a world with an indefinite future, orthodox Act-Utilitarianism can't deliver the right result – it must endorse deferring trade-offs in general, exchanging less today for more tomorrow, but it cannot explain why anyone has moral reason to cease a series of deferral. And Longtermists cannot judge it *good* if external factors – such as unpreventable extinction or the disappearance of Longtermist agents – made deferral simply unavailable. This isn't so much a paradox of infinity, as a paradox of *avoiding* infinity when each step on the road to infinity is morally indistinguishable from the last, and each seems to make the world better. Some alternative forms of Act-Utilitarianism – Person-Affecting rather than Totalist views, or the radical satisficing view of reasons in place of maximising views – do something to lessen the blow, although they do not entirely solve the problem. But these views also undercut the justification for Longtermist projects.

Longtermists may be sanguine. I have focussed on Act-Utilitarianism, and my arguments have been theoretical and technical. But Longtermism is a social and political movement whose identity is perhaps now more connected to its practical prescriptions than its historical utilitarian underpinnings. Longtermists are already seeking non-utilitarian justifications for their projects, and my arguments may simply encourage this work. But I do not think it will be easy to find *any* justification for the radical demands of Longtermism that avoids something like the Jam Tomorrow paradoxes. If we should sacrifice our today for the world of tomorrow, why should not the people of tomorrow sacrifice their interests, which we had fought so hard to promote, for the lives still to come?

Longtermism has achieved what few academic movements achieve, and found influence in the world beyond the academy, in the minds of philanthropic billionaires and in the boardrooms of tech companies whose products may transform the world. It has gained this influence in large part through the moral clarity of its arguments. But if I am right, the arguments are not so clear, after all.

Bibliography

Arntzenius, F., Elga, A. & Hawthorne, J. (2004). Bayesianism, Infinite Decisions, and Binding. *Mind 113 (450):251 - 283.*

Berkey, B. (2021). The Philosophical Core of Effective Altruism. *Journal of Social Philosophy 52 (1):93-115.*

Bostrom, N. (2002). Existential risks: analyzing human extinction scenarios and related hazards. *J Evol Technol 9 (1).*

Bostrom, N. (2003). Astronomical Waste: The Opportunity Cost of Delayed Technological Development. *Utilitas 15(3):308–314.*

Bostrom, N. (2013). Existential Risk Prevention as Global Priority. *Global Policy 4 (1):15–31.*

Doody, R. (2022). Don't Go Chasing Waterfalls: Against Hayward's "Utility Cascades". *Utilitas 34 (2):225-232.*

Elster, J. (1979). *Ulysses and the Sirens: studies in rationality and irrationality.* New York: Cambridge University Press.

Francis, T. (ms) Population Axiology without Identity

Greaves, H. (2016). Cluelessness. *Proceedings of the Aristotelian Society 116 (3):311-339.*

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

Hayward, M. K. (2024). III—Doing Our 'Best'? Utilitarianism, Rationality and the Altruist's Dilemma. *Proceedings of the Aristotelian Society 124 (1):49-70.*

Hong, F. & Russell, J. S.(forthcoming). Paradoxes of Infinite Aggregation. *Noûs.*

Landesman, C. (1995). When to Terminate a Charitable Trust? *Analysis 55 (1):12 - 13.*

Lauwers, L. & Vallentyne, P. (2004). Infinite utilitarianism: More is always better. *Economics and Philosophy 20 (2):307-330.*

Lenman, J. (2000). Consequentialism and Cluelessness. *Philosophy and Public Affairs 29*

(4):342-370.

Li, T. et al. (2012). Space-Time Crystals of Trapped Ions. *Physical Review Letters, 109(16).*

MacAskill, W. (2022a) *What We Owe to the Future.* London: OneWorld.

MacAskill, W. (2022b). Are we living at the hinge of history? *In Ethics and Existence: The Legacy of Derek Parfit.* eds. J.McMahan, T.Campbell, J.Goodrichand K. Ramakrishnan, 331–57. Oxford: Oxford University Press.

Mogensen, A. L. (2021). Moral demands and the far future. *Philosophy and Phenomenological Research 103 (3):567-585.*

Mulgan, T. (2002). Transcending the infinite utility debate. *Australasian Journal of Philosophy 80 (2):164 – 177.*

Narveson, J. (1973). Moral Problems of Population. *Monist, 57(1), 62–86.*

Nelson, M. T. & Garcia, J. L. A. (1994). The Problem of Endless Joy: Is Infinite Utility Too Much for Utilitarianism? *Utilitas 6 (2):183-192.*

Ord, T (2020). *The Precipice: Existential Risk and the Future of Humanity.* London: Bloomsbury.

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

Parfit, D. (2011). *On What Matters, Vol. 2.* Oxford: Oxford University Press.

Peter, F. (forthcoming). Relational Moral Demands. *Proceedings of the Aristotelian Society, 2025.*

Railton, P. (1984). Alienation, Consequentialism, and the Demands of Morality. *Philosophy and Public Affairs Vol. 13, No. 2. 134–171.*

Vallentyne, P. (1993). Utilitarianism and infinite utility. *Australasian Journal of Philosophy 71 (2):212 – 217.*