Can We Explain the Extinction Intuition?
McMahan and the Moral Significance of the Bad Things that Happen to Perfectly Good People

Just Philosophy (Conference)
Oxford University
June 13-14, 2025

Melinda A. Roberts
The College of New Jersey
robertsm@tcnj.edu

[Draft 2025.06.20]

In part 1, I examine the explanation Jeff McMahan proposes for the intuition that human extinction would be "bad," indeed, the "worst." I suggest that that explanation is incomplete. Though McMahan provides three (I think unassailable) points that get us some way toward the extinction intuition, a fourth *critical* point has yet to have been established. It may be that McMahan hasn't realized just how critical that fourth point is (he may think it simply materializes out of the first three). Or it may be that he considers the fourth point fully established for his reader given its close association with what he has called *the asymmetry* and given, too, the various arguments that he has put forward with the aim of discrediting the asymmetry. Accordingly, in parts 2 and 3, I focus on the two most compelling of those arguments, the *parity* argument and the *populate the distant planet* argument. And I show that those arguments don't quite get us to the result that Jeff wants.

Part 1: McMahan's explanation of the *extinction intuition*

Jeff McMahan notes that some theorists – and surely just a few – consider human extinction *good* and the alternate outcome, human survival, to be a *bad* thing. This position we can call *antinatalism*. Put in comparative terms, antinatalism says that the possible future or world or, in Jeff's terminology, *outcome* in which humans go extinct is *morally better* than the *outcome* in which humans survive for a *very long time* (perhaps forever). Accepting – for the moment – the close connection between how outcomes compare in respect of their moral betterness and how the choices that produce those outcomes are to be evaluated, we can say more: that the extinction choice is *obligatory* and the survival choice *wrong*. And all of that holds, according to the antinatalist, *even if* (a) the additional people who would exist under survival would have been perfectly well off – we can say *happy* – and (b) their existence would not have made things worse for any other existing or future person (human and nonhuman human persons alike).

Consider a quite local example, the *my third child* example, the case of that non-actual third child I could have produced but never in fact did produce. Suppose that that child, had it existed, would have been well off; for the sake of simplicity, we can stipulate that that child's wellbeing would have been *maximized*. Suppose, too, that that child's existence would not have made things worse for me or for any other existing or future person at all. The antinatalist would still say that my bringing that child into existence would have been a "bad" thing to do. My bringing that child into existence would have made things *morally worse* – and indeed *wrong*.

That position, both to Jeff's ears and to mine, seems highly implausible. Under the conditions described, could my having produced that third child really have made things *worse*? Could it really have been *wrong* for me to have produced that child? No.

Moving on, Jeff notes that *most* theorists, "[him]self included," in contrast consider human extinction "bad," and not just bad but "the worst of all *practically possible* events, or outcomes." For purposes here, I'll call Jeff's intuition – that *pronatalist* intuition – *the extinction intuition*.

I'll just underline upfront a quite serious concern the extinction intuition raises for me, and I think, based on how Jeff concludes the paper I am focusing on here, for Jeff too.[1] If human extinction is bad, indeed, the very worst, then it seems that human extinction would also be bad, and make things worse, not just in cases in which other things *are* equal, but also in many cases where things *aren't* equal – in cases where human survival (perhaps indefinitely), *does* make things worse, and perhaps *much* worse, for some, or many, otherwise existing or future people. Depending on the details of our case, that can be a highly troubling result.[2] Yet that extension of the extinction intuition seems hard to avoid once we accept the intuition in its original form.[3] --For purposes of this present paper, however, I am setting aside that quite serious concern.[4]

Jeff's sociological remarks are limited to antinatalism and pronatalism. He doesn't speculate as to just how many of us favor a *third* position: that, other things equal, human extinction makes things *neither better nor worse*. That human *extinction* versus human *survival* isn't, on its own, the important thing – that the important thing is that we do the best we can for each person *does or ever will exist*; that the important thing is that we not leave *any* person up the creek at *any* world where that person *does or will exist*. We are, to quote, but very slightly correct, Narveson, "in favour of making people happy" but, *with restrictions*, neutral about "making happy people."[5] Certain restrictions being necessary for that third position to be considered at all credible, I will it *restricted neutrality*.

Figure 1 sums the three positions just described:

Figure 1: Ranking options for human extinction versus human survival

Persons p1 – pm do or will exist in worlds ("outcomes") w1 and w2; q1 – qn do or will exist only in w1; w1 and w2 exhaust the range of accessible worlds (Jeff writes "practically possible . . . outcomes"[6]) relative to either w1 or w2; an assigned wellbeing level for each person at each world is indicated in terms of natural numbers (which may be considered to have their ordinal value only); nonexistence at a world is indicated by "*."

First camp (*antinatalism*): w1 is "bad"; w1 is worse than w2 Second camp (*pronatalism*): w2 is "bad"; w2 is worse than w1 Third camp (*restricted neutrality*): w2 isn't *worse than* w1; and w2 isn't *better than* w1 (*provided that* no accessible world, including but not limited to w2, is still better for any of q1-qn than w1).[7]		
	w1	w2
p1 – pm	+10	+10
q1 – qn	+10	*

The restriction on restricted neutrality set out in figure 1 – that no other outcome makes things still better for the additional people q1 – qn – can be relaxed in various ways.[8] But having some such restriction in place is critical, as John Broome's argument against Narveson's *unrestricted* neutrality intuition clearly shows.[9] The well-thought-out restriction will recognize that bringing additional possible people into existence creates various moral risks[10] – that it *can* make things worse (contra the mere addition principle) and *can* be wrong even when those additional people would be perfectly happy and their existence makes things worse for no one at all.

Applied to the more local case, we get this result. Restrictions satisfied, the world where I *don't* produce that third child is no better or worse than – is exactly as good as – the other-things-equal world where I *do* produce that third child. Putting the point in deontic terms and more generally: moral agents are morally – though no longer legally or constitutionally in the U.S. – free to produce that additional child or not. As a *type* of maximizing consequentialist, I believe that moral law remains *non*-neutral on many things. But I am pretty sure that morality remains *neutral* on that one thing.

***

I'll just note that, to his credit, Jeff's goal isn't to *install* in each of us the extinction intuition – his own personal intuition that extinction is "the "worst of all *practically*

possible . . . outcomes." I think neither Jeff nor I aim to *install intuitions* in other people's minds or hearts, however deeply held or, so to speak, internally adjudicated we might consider those intuitions to be.

Rather, Jeff's main goal, in this particular paper, is to provide an *explanation* or *justification* for, or *theory* of, the extinction intuition.

That's to his credit as well. We can't just dig our heels in on our intuitions, however deeply held they might be. For our own peace of mind, we need to *understand* whether, and how, a given intuition fits into some broader moral theory we've already carefully inspected *and* enthusiastically confirmed *and yet* continue to investigate. Even more importantly: we need to *explain* our intuitions before we can in good conscience *promote* them to our students or to anyone else. --*Why* must we *explain* ourselves? So that our audience might *for themselves* have a chance both to review the terms of that broader moral theory and to determine just how well the intuition in fact fits that theory.

Accordingly, in this paper, I ask two questions. What is Jeff's explanation for his own deeply held intuition, the *extinction intuition*? What is his *theory*? And, second, does it *work*? *Does* it explain, or justify, the intuition that human extinction is "bad," indeed, among all the "practically possible" – *accessible* – outcomes, the "worst"?

To figure out what that explanation is going to look like, Jeff considers what we shall want to say about Jonathan Glover's "infertility pill" case.[11]

Glover's case imagines a pill that maximizes wellbeing for each now existing person but that also leaves each such person completely infertile, thereby inducing (near-term) human extinction, a "last generation" scenario: a world in which no additional people will ever come into existence at all. Taking the pill, in other words, strips benefits away from (i.e., is bad for, or makes things worse for) *only* those additional possible people "who *could* exist with good lives [witness w3] but will not exist" at all under the choice to take the infertility pill (witness w4) (emphasis added).

Thus Figure 2:

Figure 2: Glover infertility pill case

In w3, many in the existing population *don't take* infertility pill (maintaining fertility but reducing wellbeing for each existing person); in w4, the entire existing population *takes* the infertility pill (inducing infertility in, and maximizing wellbeing for, each of them).

	w3	w4
Many people, all of whom now exist	+8	+10
Many, *many* additional, future, people	+10	*

What the Glover case helps us see, Jeff thinks, is that, given that we think (as Jeff does) that taking the infertility pill makes things *worse* even as it makes things *better for* every existing person, it must be that we think that that's so *in virtue of* what it does to the *people who will never exist at all* under that choice.

That must mean, according to Jeff, that the good explanation – the correct theory – of the extinction intuition will "refer to" the people who could have had "good lives" at w3 but instead, at w4, have no lives at all. The explanation – the theory – will specifically say about those people that they *matter morally*: not (merely) "impersonally" – not (merely) insofar as the wellbeing they are assigned at a world where they exist contributes to *aggregate* wellbeing at that world – but (also) "for their own sakes."

As to that first point, I am in complete agreement with Jeff. Now, unlike Jeff, I don't happen to be trying to explain the intuition that the entire existing population's taking the infertility pill makes things worse – or the extinction intuition itself. Still, as to the *distinct* question – the question what makes for a good explanation *of anything*, or a correct moral theory *of anything* – I agree with the Jeff. The merely possible do indeed matter – *despite* the fact that they never exist at the actual world; indeed, *despite* the fact that they never happen to exist at the world, *actual or not*, that happens to be the target of our ranking or evaluation efforts – "for their own sakes."

Jeff takes a second point from the Glover case as well: that a good explanation – a correct theory – will recognize that the claim that the merely possible matter "for their own sakes" isn't narrowly confined to any special sort of world, to (e.g.,) those worlds where they happen to *exist* or to the *actual* world. A failure of *existential* status at a given world doesn't mean a failure of *moral* status at that world.

Here, too, I am in complete agreement with Jeff. We both happily reject "moral actualism."[12]

And, third and last, Jeff maintains that the good explanation – the correct theory – will have it that, in some sense or another, a happy person's never existing at all strips that person of a *benefit*. Jeff doesn't want to say that the world where the child exists is

better for that child than the world where the child never exists. He, that is, rejects *nonexistence comparability*, on grounds that I do not completely understand or think that we can, in the end, credibly endorse.[13] He does, however, want to recognize that it *matters*, to that person, whether that person exists or not – that that person has an *interest* (we might casually say) in existing that we do well not to simply ignore. (Among other things, ignoring the interest in the happy person's existing militates against taking the interest that the miserable person has in never existing into account.)

As to this third point, too, I am in complete agreement with Jeff.

Those three points together constitute Jeff's explanation or justification or theory of the extinction intuition. To sum up:

> (1) merely possible people matter for their own sake, and not (only) *impersonally* or as a function of an increase (or decrease) their existence at a world contributes to (or subtracts from) *aggregate* wellbeing at that world;
>
> (2) just as people who do or will exist at the *actual* world have moral status, so do people who are *merely possible* relative to that world; *moral actualism* is false (I accept, though Jeff perhaps does not, that people who are merely possible relative to the actual world have *exactly the same* moral status as people who do or will exist at the actual world) and
>
> (3) the person's coming into existence at a given world and being "happy" (i.e., have a positive wellbeing level) at that world constitutes a *benefit* to that person at that world (I accept, though Jeff does not, that existence in that case makes things *better for* the person as compared against that person' never existing at all).

As noted, I consider each of these three points unassailable and, for purposes here, accept them as such. I won't work through the arguments that support that position here. I'll just note that I see each point as solidly confirmed by the work that has been done in the area of population ethics over the last handful of decades, including work by Jeff himself, as well as by Broome, Arrhenius, Rabinowicz, Temkin, Parfit (of course) and many others.

***

But there's a problem. Even if we accept that it's entirely appropriate to complement points (1) – (3) with various independent and uncontroversial moral principles – and we should accept that that's so – we *still can't get to Jeff's extinction intuition*. There's a gap, a leap of logic. He needs a fourth point.

Let's say that a *loss* (I am happy to say *harm*) is sustained by a person at a world whenever that person is (in Jeff's terms) deprived of a benefit otherwise available to that person or (in terms I accept) that world makes things *worse for* that person than some other accessible world makes things for that person. Then – remember, we're trying to get to the result that human extinction is not just "bad" but the "worst" –

> (4) any loss any person sustains at any world has (roughly the same) *moral significance* as any other (similarly-dimensioned) loss sustained by that same person at any alternate accessible world; any such loss sustained by any such person at any such world *counts against* the world in which it is sustained, *whether the person does or will exist in that world or not*.

In other words, we might want to say, with Jeff, that *all people*, merely possible or not, matter morally at all accessible worlds, *but* that it's not the case that *all their losses* matter morally. *I* might matter morally and just as much as anyone else does, whether I exist in the actual world or not. It doesn't follow that all the *losses* that I might sustain, including those I sustain at worlds where I never exist, also matter morally.

Consistent with (1) – (3), we might, in other words, accept the *loss distinction thesis*.[14]

> *Loss distinction thesis*: The *loss* a person p sustains at a future x compared against a future y is *morally significant* – i.e., *counts against* x, potentially making x *worse than* alternate accessible worlds (including but not limited to y) and the choices made at x *wrong* – *only if* p *does or will exist* in x.[15]

Jeff needs point (4) in order to complete his explanation of the extinction intuition. But he can't get (4) without ruling out the loss distinction thesis. And we can't simply assume that the loss distinction thesis is itself false without begging the very question we are trying her to resolve.

***

Let's back up. The loss distinction thesis insists that what we can all *existential* losses – the losses sustained by a person at a world where that person never exists – have no moral significance at all. The loss distinction thesis, at the same time (as a mere necessary condition on when a loss has moral significance), makes room for other moral principles (which themselves may be entirely noncontroversial) to step in to declare that what we can call *ordinary* losses – e.g., the loss a person sustains in a world (whether the actual world or not) where that person is hit by a car; the loss a child sustains when brought into an avoidably burdened existence – have full moral significance.[16]

The loss distinction thesis itself, to my ears, is grounded in a highly intuitive idea that I have elsewhere called the *existence condition*.

> *Existence condition*: Where worlds y and z are accessible relative to world x, x is *worse than* y, and a choice made at x is *wrong*, *only if* a person p *does or will exist* at x and x is worse for p than z (where z may, but need not, be identical to y).

A cautionary note: the idea captured by the existence condition, as well as the loss distinction thesis, is one that many theorists have found highly intuitive even as they have taken themselves to be compelled to reject it. On more than one occasion,

however, it seems that it's the formulation that compels rejection and not the underlying intuition.[17] The existence condition, in contrast, both says less (it doesn't go beyond the intuition) and does less (it doesn't, e.g., compel us to reject various other, completely uncontroversial, moral principles). Thus it's narrowly drawn, providing only a *necessary*, and not a *sufficient*, condition on moral worseness and moral wrongdoing, and the necessary condition itself is more generous – easier, that is, to satisfy – than what we see in various competing formulations.

***

Again, if the loss distinction thesis (and, with it, the existence condition) is at least credible, if it's even a contender, then Jeff's own explanation of the extinction intuition isn't even close to complete.

Is the loss distinction thesis at least credible? Or is it purely arbitrary, completely ad hoc? I don't think that we can say that it's arbitrary or ad hoc. The distinction drawn seems on the face of things at least plausibly relevant to moral analysis. We have, on the one hand, the sorts of losses people avoidably sustain by virtue of the fact that they *exist and suffer* at any given world – *ordinary* losses – and, on the other, the losses they avoidably sustain at worlds by virtue of the fact that they *never exist at all* at those worlds – that is, *existential* losses.

Putting the point in deontic terms and quite generally, we can say this: we don't have to have children, but if we do have them, we should make things better for them rather than worse. Your parents and mine were obligated to choose in ways that make things better for us rather than worse. But they were *never obligated to bring us into existence to begin with*. [18]

Again: until we rule out the loss distinction thesis, we can't get (4). And until we get (4), we can't explain the extinction intuition.

Part 2: The "parity" argument

Jeff disagrees. He thinks (I think he thinks) that he, in effect, has already secured point (4) by way of discrediting – destroying – the *happy child half of the asymmetry*, the very asymmetry that Jeff himself famously articulated awhile back and immediately, in that very same paper, questioned.[19]

The *miserable* child half of the asymmetry is widely accepted, including by Jeff. According to the miserable child half, it, other things equal, makes things worse, and is wrong, to bring the miserable child into existence. It's the *happy* child half that Jeff questions, and, in the paper I'm focused on here, argues strenuously against: it doesn't, other things equal, make things worse, and isn't wrong, *not* to bring the happy child into existence.

Now, I agree that, if the happy child half of the asymmetry must go, so must the loss distinction thesis go, and, with it, the existence condition.

In this paper, I'll look at two of Jeff's (I think most impressive) strategies for discrediting the happy child half of the asymmetry.

***

The first, discussed in this part 2, is the *parity* argument. There, he uses the miserable child half of the asymmetry to make the case against the happy child half.

The argument is this. By a certain "parity of reasoning" with what we want to say about the *miserable* child, we should now say what Jeff wants us to say about the *happy child*. After all, as points (1) – (3) establish, both children matter for their own sakes whether they exist at any special world or not, and for both children their coming into existence or not is a *big deal*. Much is at stake for both; whether we put the point in terms of losses sustained or benefits forgone, it matters to both whether existence happens for them or not.

Jeff's three (unassailable) points don't, in other words, apply only to the predicament of the miserable child. They apply to the predicament of the happy child, and just as well.

We should accept that the *parity of reasoning* between the two cases is *extensive*.[20] But that doesn't mean it's *complete*.

The loss distinction thesis – which we can't yet have ruled out, without begging the very question at hand – shows us just where parity fails.

For that thesis would have us analyze the two halves of the asymmetry in two very different ways.

Regarding the miserable child half, we'll say this: *consistent with* the loss distinction thesis, the miserable child's loss at the world where that child exists and is miserable has full moral significance. It *counts against* that particular world: other things equal, it makes the world where the child exists and suffers worse than the world where the child never exists.[21]

But regarding the happy child half of the asymmetry, we'll say this: Because the happy child *never exists* at the world where that child isn't benefitted, that is, where that child sustains a *loss*, that benefit deprivation, that *loss*, *doesn't* count against that world: other things equal, it leaves the world where the child never exists no worse than the world where the child does or will exist.[22]

Thus – without begging the very question we're trying to resolve – we are forced to recognize that the parity argument doesn't, after all, discredit the happy child half of the asymmetry. That half of the asymmetry still intact – the loss distinction thesis still

intact – we are still missing point (4). And without point (4), we have yet to provide a complete explanation – a correct theory – of Jeff's extinction intuition.

Part 3: Populate-the-distant-planet case; *Interstellar*

A second strategy Jeff puts to work to discredit the happy child half of the asymmetry is based on his "populate-the-distant-planet" case.

In that case, we have the choice whether or not to populate a distant planet with a very great number of future human beings. If we do populate the planet, it's "statistically certain" that many, many of those future people will have lives "well worth living" *and* that a proportionately smaller number of people will have lives less than worth living, that is, "bad lives."

Figure 3: Jeff's populate-the-distant-planet case

	w5 (world where we choose to populate the distant planet)	w6 (world where we choose *not* to populate the distant planet)
p1 . . . pn (many, many people)	+10	*
q1 . . . qm (many people, but many fewer people)	-10	*

According to Jeff, if the happy child half of the asymmetry is correct, then we are "require[d] *not* to populate the planet" (emphasis added).

Why is that? We all agree that the misery of the miserable people in that case has full moral significance; it *counts against* the world where we choose to populate the distant planet.

But according to the happy child half of the asymmetry – and the loss distinction thesis – the happiness of the happy people is devoid of moral significance: it *doesn't count in favor of* that same world.

But with nothing on hand to balance out the morally significant misery of the many people, we seem forced to conclude that the choice to populate the planet makes things worse and is wrong.[23]

But that result seems suspect, even to my hardened ears. Now, Jeff wants to say that the choice to populate the distant planet is *required*, i.e., *obligatory*. That's another issue. The issue I'm raising now is this: surely, as Jeff at least suggests, the choice to populate the distant planet is at least *permissible*.

On that basis, Jeff rejects the happy child half of the asymmetry.

I love this argument! For one thing, it's brilliantly provocative and *so close* to a perfect argument! For another, it situates us squarely in the land of *Interstellar*![24]

In that film – "film"? – a Professor Brand uses deception to convince a pilot, Cooper, to take on the job of travelling through the universe and time via wormholes, black holes and tesseracts to identify a faraway planet in a faraway galaxy suited for sustaining human life and avoiding human extinction. Setting aside the deception – that's just not Jeff – I'll just say I that I see a bit of Jeff in Brand. Brand, played by Michael Caine, wants the incredibly complex populate-the-distant-planet-with-new-people project to succeed, *no matter what that means for the people who do or will otherwise exist*. (*Complex* doesn't *begin* to describe Brand's project, which contemplates not just identifying that faraway life sustaining planet but also packing rocket ships with vats full of frozen human gametes and embryos and artificial wombs that will then jet off in the direction of that life sustain planet and give the human species a fighting chance at survival.) But I think it's Cooper, played by Matt McConaughey, who has the correct moral calculation in mind: he's not going to participate in the populate-the-distant-planet-with-new-people project if that requires *abandoning* the people who do or will exist on Earth and seem doomed to face a very bad end – life ravaged by climate change or nuclear catastrophe or both (winds, fires, dust, poverty, no crops beyond corn, and soon no corn). If forced to choose, Cooper is going to stay behind and do *the best he can* to make the lives of those existing and future earthbound people go better. But he's deceived by Brand into thinking that the goals can *both* be achieved. All it will take is for the earthbound scientists to solve the "problem of gravity" while Cooper's off looking for the life sustaining planet. (The science of *Interstellar* is *perfect*, my logic students tell me, and the moral thesis of the *Interstellar*, as I interpret it, isn't bad, either.)

***

But let's go back to Jeff's case and Jeff's argument. And let's suppose that we agree with him that it's *permissible* – if not, as he says, *required* – to choose to populate the distant planet. Does it follow that the happy child half of the asymmetry is false?

I don't think that it does. I think we can easily accept *both* the permissibility of the choice under scrutiny *and* retain the happy child half of the asymmetry. What makes the choice permissible isn't that the happiness of the happy future people is functioning to *balance out* the misery of the miserable future people. What makes the choice permissible, rather, is what the choice looks like from the perspective of each additional possible person.

So let's take a closer look at *that*. Jeff writes that it's "statistically certain" that, among a large possible future population (consisting of both the p-people and the q-people), a far smaller population (consisting of just the q-people) will have lives less than worth living – that is, miserable lives.

Let's, then, consider what, at a granular, *one-person-at-a-time*, level, that means. What does it mean for *each and every member* of that larger population?

It's widely accepted – and I accept – that probabilities bear on the moral evaluation of our choices. (Surely that's so, even if we ultimately reject standard expected value theory.[25]) We can then ask: for each person in that larger population, based on the information available to agents prior to choice, what are the chances that that person will have an existence worth having?

They are, clearly, very great. Of course, it's still a gamble. For each such person, there's *some* chance that that person will end up with a miserable life. But for each such person the game is still worth the candle: there's a still greater chance that the person will end up with a very good life.

We accept, in more routine, same person cases that those kinds of chances can make a choice that would otherwise be wrong perfectly permissible.

Consider the "it's your own child" case. Suppose that your own child – a child who, whatever choice you make, does or will exist – will be left with a lifetime wellbeing level of zero if you do *nothing*. Suppose that, instead of doing *nothing*, you could do *something* that creates a very high probability of greatly improving things for your child. But it's a gamble: the choice to do something doesn't just create a high probability of success; it also creates a low probability of leaving your own child with a life *far less* than worth living.

Even in the case where the gamble turns out badly – as it will in some possible worlds – the choice to gamble is still plausibly understood as permissible.

We can, I think, say the same thing about Jeff's populate-the-distant-planet case. Judged from a granular, one-future-person-at-a-time perspective, the choice to populate the distant planet is permissible.

But then the choice to gamble is permissible, not because the happiness of some people is functioning to *balance out* the misery of still other people. Rather, it's permissible because, for each such person, that choice has a different function entirely: we can simply say for purposes here (and even if we in the end eschew standard

expected value theory) that, for each possible future person, it *maximizes probability-related anticipated value* (and does so, let's just underline, even at worlds where the gamble turns out badly).

Now, to say that maximization of probability-related anticipated value is *one* path to permissibility, *one* sufficient condition for permissibility among others, is not at all to rule out still other paths. The happy child half of the asymmetry – and the loss distinction thesis – gives us still others. Thus, another path to permissibility (another sufficient condition for permissibility) might be – and is, in the populate-the-distant-planet case, though *not* in the it's-your-own-child case – to leave that person out of existence altogether.[26]

To conclude this part 3, let's talk about possible worlds. That the *choice* to gamble is permissible doesn't mean that each *world* at which agents choose to gamble is at least as good as the world where agents don't choose to gamble.

Consider the it's-your-own-child case. I don't think anyone thinks that the world where you permissibly choose to gamble *but things turn out badly* for your child (where your child is left with a life less than worth living) is at least as good as a second world where you don't gamble and the child is left as-is (at the zero-wellbeing level) or a third world where you do gamble and things turn out well. The first world is obviously worse than the latter two.

Ditto the populate-the-distant-planet case. Consistent with the position that the choice to populate the distant planet case is permissible, we can still say that the world in which that choice is made and the miserable people are forced into existence for the benefit of their happier and more multitudinous brethren is worse than the world where that choice isn't made and that future population never exists at all.[27] And I think that that's exactly the right way to rank the relevant worlds in that case.

Conclusion

I've suggested here common ground between Jeff's view and my own. Despite the fact that a certain number of people will have "bad" lives in Jeff's populate-the-distant-planet case or in the extinction case if we, respectively, choose to populate the distant planet or choose to work to avoid human extinction, it may well be that, other things equal, both choices are *permissible.*

But the claim – his own extinction intuition – that Jeff wants to explain isn't a claim about the mere *permissibility* of the choice to populate the distant planet or to avoid human extinction. The claim he wants to explain – his own extinction intuition – is much stronger than that: it's that those choices are both "required"; that they're both *obligatory* – and that the alternate choices – and, specifically, the extinction choice– is "bad" and indeed, the "worst." The complete explanation of that much stronger intuition, Jeff's own extinction intuition, we still don't have.

References

Glover, Jonathan [_____].

McMahan, Jeff [_____]. (introduction of "the asymmetry").

McMahan, Jeff 2024. "Human Extinction and the Morality of Procreation," University of Pavia.

Narveson, Jan. 1976. "Moral Problems of Population." In Michael D. Bayles, ed., *Ethics and Population*, 59-80. Schenkman.

Roberts, M. 2024. *The Existence Puzzles*. Oxford University Press.

Roberts, M. 2025. "Two Dogmas of Population Ethics." *Utilitas* (forthcoming).

Weinberg, Rivka 2017. *The Risk of a Lifetime*. Oxford University Press.

[1] McMahan 2024. Jeff ends that paper not with a hard-and-fast, categorical *conclusion* but rather a *dilemma*. Specifically, the logic of his argument

> [D]oes, however, leave us with a dilemma . . . . If the reason to cause a well-off person to exist is relatively weak, then the reason to cause a better-off rather than a less well-off person to exist will also be relatively weak *and* the reason to avoid extinction will also be weaker than many of us believe it is . . . . If, on the other hand, the reason to cause well-off person to exist is significantly *stronger*, it *will* explain the importance of avoiding extinction but will also ground a *duty* to have children far more often than most of us will be willing to accept.

[2] Can it really be permissible for me to bring that third child into existence, even when my doing so makes things worse, and perhaps much worse, for my two other existing or future children?

[3] If we then bring probabilities to bear, whether by way of standard expected value theory or otherwise, we then seem inevitably faced with the *fanatical conclusion*: the idea that it's permissible, or even obligatory, to reduce wellbeing, perhaps radically, for perhaps *all* existing and future people at a given world, in order to secure at that same world the *tiniest chance* that the human species will survive forever. See Roberts 2025 forthcoming and Roberts 2024 (Chap. 6).

[4] For discussion, however, see Roberts 2025; Roberts 2024.

[5] Narveson 1976.

[6] Jeff's uses the terms "practically possible" in describing a certain relation between possible outcomes, or worlds. I'm assuming the relation he has in mind to be roughly what I have in mind when I say that one world is *accessible* relative to another. We are both declining to discuss what is correct to say in terms of moral evaluation across the entire range of all *logically possible worlds* (including, e.g., worlds agents lack the ability, power or resources to bring about, e.g., worlds where we live forever, or worlds where there's no such thing as gravity, or worlds where we undo the past).

[7] The sufficient condition on when the more populated world is at least as good as the less populated world is very stringent. But it's just one possible sufficient condition on w1's being at least as good as w2. Here's an example of a more relaxed sufficient condition and one that I've elsewhere proposed: w2 isn't *better than* w1 (i.e., w1 is at least as good as w2), *provided that, that is, at least in the case in which)* no third outcome exists that is available (I've used the term *accessible*) *and* isn't ruled morally out of bounds *and* makes things still better than w1 for any of q1-qn. There will be others!

[8] See Roberts 2024 (Chapters. 3 and 4).

[9] Where the only difference between two possible alternate futures x and y is that one additional moderately well off, moderately *happy*, person p does or will exist in x, and

where there exists a third option z just like x except that p is still *better off* in z than in x, we need to say that x is worse not just than *z* but also than *y*.

[10] The risk of existence language I owe to Weinberg 2017. I don't think the recognition that existence involves risks means, however, that we must accept her, or indeed any, version of *moral actualism*.

[11] See Glover [____]. The stipulated wellbeing levels are mine. I'll just note that confounding factors seem unavoidable in Glover's example. It's hard to imagine things being clearly better for each existing person if each such person is denied the plusses of having their own baby or enjoying anyone else's. It's even harder to imagine that wellbeing is the kind of thing we get out of a pill, even if it's a pill that produces good health and a long life. However, to appreciate Jeff's point, we must try to set outside those confounding factors: we must, with Jeff, assume that, *for reasons that have nothing to do with those confounding factors*, extinction is bad, that is, that w4 is worse than w3.

To then see what the theory Jeff is after is going to look like, we then simply ask "*Why* is it bad?" *Why* is w4 worse than w3?

A preliminary question: is this a case we can actually accept? Are we confident we can process in any intuitive way the (I think outlandish) stipulations that it requires? (That existing people aren't terribly bothered by the fact that they constitute the "last generation"; that wellbeing can be got from a pill?) Well, let's suppose that we can.

[12] *Addition plus* demonstrates why that second position is virtually incontrovertible.

Addition plus	x	y	z
p	+10	+12	+5
q	*	+1	+5

In x, happy person p exists and q doesn't; in y, p is even happier in part because the minimally happy q also exists (organ donor; slave); and in z, p is a worse off than in x or y and q is substantially better off than in y, such that, in z, p's and q's wellbeing levels are the same. q is merely possible relative to x; if x is the actual world, then q is merely possible relative to the actual world. But q still matters morally (and every bit as much as p does). Why? It's q's moral status, in combination with the avoidably low wellbeing level q has in y relative to z, that explains why it's OK for p to have the reduced wellbeing level that p has in x: why x is, after all, at least as good as y.

[13] See Roberts 2024 (Appendix [A]). That small, technical difference between us, however, isn't material, so far as I can see, for purposes of anything I want to say in this present paper.

[14] I, for reasons known only to some prior self, previously called the *loss distinction thesis* called "variabilism."

[15] By implication from the loss distinction thesis, and given that the *losses* and *gains* are just ways of talking about the same set of facts, the morally significant *gain* is limited to the gain that *reverses* the morally significant *loss*, i.e., the loss sustained at a world where the person who sustains the loss does or will exist. On that position, a gain accrued by a given person at a given world can have moral significance even if that person *never exists at that world at all*. We thus – looking ahead to the discussion of *the asymmetry* in part 2 below – position ourselves to explain why it is that the world where the completely miserable person *never exists* is better than the world where that person exists and suffers. (Versions of a *gain* distinction principle that mechanically substitute the relevant terms into the loss distinction thesis – *gain* in for *loss*, e.g. – miss that point. The principle that deems a person's gain at a world to have moral significance only if that person exists at that world generates the clearly problematic and intuitively false result that a person can't ever be better off having never existed at all.

[16] The loss distinction thesis – like happy child half of the asymmetry and the existence condition – is categorical in nature. Thus it rules out the two tier view. Putting that view in the terms I've suggested here, it implies that a sufficient number of *existential* losses will balance out – or even more than balance out – a lesser number of ordinary losses. (Let's assume that all of the losses, of either sort, make things *worse for* the person who sustains them to exactly the same degree, e.g., that each loss leaves the person with a zero wellbeing level (i.e., none at all) when that same person (just as easily) could have had a wellbeing level at a fulsome 10.) What is that "sufficient number"? A problem with the two tier view is that its advocates tend not to say. But let's suppose it's twenty. Do we make up for causing (or allowing) one two year old child's wellbeing level to fall from a fulsome 10 to the zero level by way of bringing twenty additional children into existence? I don't think that we do. Do I make up for mistreating my existing child by bringing an additional twenty children into existence? I don't think that we can. See note [20] below (bobcat kitten).

[17] As but one example, see note [17] above. (mechanical substitution of *gain* for *loss*).

[18] Let's apply the loss distinction thesis to the case of the tiny bobcat kitten. (I'm truly *not* trying to install an intuition here, but just to make the case that there really is an open question.) To complete the explanation – to bridge the gap from Jeff's three points to the extinction intuition – we'd need to say something like this:

the "plight" of the never-existing, arguably zero-wellbeing, tiny kitten makes the world where I never bring that kitten into existence to begin with *worse*

in roughly the same way that

the plight of the zero-wellbeing, tiny kitten I now hold in my hands and (i) fail to rescue from whatever horror will otherwise befall it or (ii) badly mistreat makes *that* world morally worse.

According to the loss distinction thesis, that first loss, that *existential* loss, *doesn't* count against the world where that loss is sustained. Consistent with the loss distinction, that second loss, that *ordinary* loss, very much counts against the world where it's sustained: it makes that world, relatively to still others, *worse*.

But is it really clear to us that the loss distinction thesis is false? Do we really think that – to extend the point – it would have been permissible to leave the one kitten to its misery (to leave it to the horror of the zero-wellbeing existence) – *provided that* we *balance things out* by taking steps to bring a *distinct* kitten, a *happy* kitten, into existence? *I don't think that we do.*

What if it's not just one additional happy kitten, but two, or twenty? Would *that* make it permissible to leave the one kitten, the kitten I hold in my hands, to suffer? *I don't think that it would.* What is called the *two tier view*, according to which the losses sustained by the twenty kittens might, in theory, add up to less than the loss sustained by the one, thereby giving support to the position that it's wrong, and makes things

worse, to leave the one kitten, the kitten I hold in my hands, to suffer. But before we can say one way or another, the theorist behind the two tier view must tell us what the relevant ratios are. If 20 never existing kittens' losses don't tilt the balance in favor of letting the one suffer, do 30? Do 40? The better view is, I think, that the numbers don't matter: that what matters is whether the person who sustains the loss does or will exist in the world in which that person sustains that loss.

[19] McMahan [___].

[20] The happy child, just like the miserable child, matters for that child's own sake, and leaving that child out of existence altogether doesn't strip that child of its moral status (*moral* status, again, being independent of *existential* status), and bringing either child into existence can have the effect of burdening that child or benefiting that child (making things worse for that child or making things better for that child).

[21] To put things in more concrete terms – and avoid the vague "counts against" – we might consider adding a maximizing, Pareto-like principle – principle that *granulates* over the population rather than *aggregates* – to our moral theory.

Where each person who does or will exist in y also does or will exist in x, and x is worse than y for at least one person (e.g., the miserable child) who does or will exist in x, and y is worse than x for not person who does or will exist in x, x is worse than y (y is better than x).

We then easily get the result that the future x that includes the miserable child is worse than the otherwise similar future y that excludes the miserable child.

[22] In more concrete terms:

Where each person who does or will exist in y also does or will exist in x, and *even if* x is better than y for at least one person (e.g., the happy child) who does or will exist in x, y is worse than x (x is better than y) *only if* y is worse than x for at least one person who does or will exist in y.

This principle easily implies that y isn't worse than x: the necessary condition is failed in the happy child case; y is worse for the happy child than x, but y isn't worse than x for anyone who does or will exist in y.

[23] Ditto human extinction; thus Jeff writes that, "unless bad lives can be morally offset by a sufficient number of good lives, the balance of moral reasons *may* actually *favor* extinction."

[24] The people Cooper wants not to abandon – to continue doing his best for – on Earth include his own very bright daughter (his son being relegated throughout the film to the ranks of the supposedly very dull). Brand, as noted in the text, deceives Cooper into thinking that he can both take on the job of avoiding human extinction and also continue to do his best for those who do or will exist on earth. But in the end, to Brand's own surprise, that's *just what happens*; the earth-bound scientists, led by

Cooper's own very bright daughter (we don't learn what ever happened to his poor son), *do* solve the "problem of gravity."

[25] See note [3] above (fanatical conclusion).

[26] Of course, in the it's-your-own-child case, you don't at the time of choice have the option of leaving your child out of existence altogether. We can't change the past. I'll just note that, to my ears at least, the choice *not* to gamble seems very possibly flat out wrong. At least, the choice *not* to gamble, made at those worlds where that choice, as you all-but-knew it would ahead of time, turns out badly, seems, to my ears, wrong. You *all but know*, given the probabilities, that the gamble will reap substantial benefits for your own child. And you *certainly* know that the choice to gamble is your best way forward – that it's your *best* way, no *perfect* way being available, of doing the best you can for your child.

Admittedly, that way of looking at the case may seem to raise a question about what I said earlier. If we are convinced that the choice to gamble in the it's-you-own child case is obligatory, must we say the same about the populate-the-distant-planet case? That there, too, the choice is obligatory?

No. *Even if* it's obligatory to gamble in the it's-your-own-child case, it doesn't follow that it's obligatory to gamble in the populate the distant planet case. To choose not to populate the distant planet is simply to choose to leave a whole lot of people out of existence altogether. And that – per the happy child half of the asymmetry, per the loss distinction thesis, per the existence condition – is perfectly permissible (and – again – not an option in the it's-your-own-child case).

[27] Or so a principle I've elsewhere called the *Pareto reduction principle* would (more concretely) insist:

Where each person who does or will exist in y also does or will exist in x, and x is worse than y for at least one person who does or will exist in x, and y is worse than x for no person who does or will exist in y, x is worse than y (y is better than x).

More generally: ordinarily, there's a close *conceptual* connection between the evaluation of choice and the ranking of worlds in respect of their overall betterness. When probabilities pop into the picture, however, that connection often fails. The it's-your-own-child case is clearest evidence of that. But there are an infinite variety of other such cases.