CHAPTER 8

Should we colonize (interstellar) space?

Joseph Gottlieb
Department of Philosophy at Texas Tech University, TX, United States

Chapter Outline

8.1 Introduction

This paper explores the following question: should we colonize space? By "colonize space" I mean the creation of self-sustaining space colonies, including, eventually, extra-solar space colonies. I understand our question as a normative, moral question. It is not a question about whether we can, technologically speaking—now, or in the future—colonize space, although it plausibly assumes that we can, since if we cannot, the moral question is otiose.[1]

Our question is a specific application of the more general question of what we should do, morally speaking. A reasonable first answer to this question is this: we should choose whichever course of action leads to the best outcomes. I assume that, what outcome is best is at least in large part a function of the *welfare* or *well-being* in that outcome.[2] This is not to deny that there might be other values whose importance hold

[1] I will, however, discuss specific technical details as they arise, relevant to the main thread. For now, it is worth noting that there is likely no in principle barrier to space colonization, including interstellar colonization. Indeed, intrasolar colonization might be feasible (if not advisable) with current or near future technology. For discussion, see Beckstead (2014).

[2] I assume that welfarist considerations are wide, insofar as I deny that one state of affairs can be better or worse than another only if there are some individuals who are the same in both. So, in this way, I allow that there is *impersonal value*, not just *personal value*. Suppose I am comparing two states of affairs: one ("A") with 10 people suffering from severe nausea, and another ("B") with 10 different people, but where only five people suffer from severe nausea. The state of affairs A is not worse *for* anyone, since presumably for that to be the case, the people in A would also have to be in B. Thus, if B is worse than A—and it seems that it is—it is worse in virtue of having more negative impersonal value. For this distinction, see Parfit (1984). In Sect 2.1, I briefly entertain what happens if we drop the assumption that there is such a thing as impersonal value.

Interstellar Travel
DOI: https://doi.org/10.1016/B978-0-323-91360-7.00009-4

independent of their contribution to well-being, but it is certainly a central value and one that we can compare across different states of affairs while holding other values fixed. So, if the world-path in which we colonize space leads to better welfare outcomes than the world-path in which we do not, then, after accounting for any side-constraints, we ought to colonize space.

To many, whether colonizing space will lead to better welfare outcomes may seem like a settled matter. Space colonization is frequently viewed in highly-optimistic terms, with visions of spinning cities at Lagrange points, lush, perfectly terraformed planets, near unlimited material resources, and near utopian societies dominating. The seeding of this conception of space colonization has a complicated history, with ties to architectural visions—think here of the O'Neil Cylinder (1976) and its attendant sociological commentary—science fiction, and the Cold War. A space-based future, it is often presumed, is a Star Trek future. So, the thought goes, of course we should pursue large-scale space colonization. To do anything less would be a moral failing.

However, the tide of optimism might be slowly turning. There has been an influx of arguments for the claim that space colonization will lead to catastrophic outcomes, posing a serious *suffering risk* ("s-risk")—a choice uniquely poised to create astronomical amounts of suffering (e.g., Torres, 2018; Tomasik, n.d.; Duedney, 2020; Kovic, 2021). Here, the conclusion is, whatever its other merits, space colonization is, on balance, a terrible idea, and certainly not anything we should pursue in the foreseeable future.

I do not aim to fully justify the pessimistic attitude; likewise, I do not aim to fully justify the optimistic attitude. Instead, my central aim is to articulate a framework from which we can assess these competing attitudes, and thus decide whether the world-path in which we pursue colonization leads to better welfare outcomes. No framework is perfect, but a virtue of the one to follow is that it considers space colonization within its proper scope (cf. Gottlieb 2022): an undertaking that, no matter its extent, is an inherently long-term endeavor, and will have immense impacts on how well our future unfolds.[3] The framework itself is neutral on whether we ought to colonize space, but provides the necessary scaffolding for helping us settle the matter insofar as it directs us to the relevant normative and empirical questions.

Naturally, the framework has limits. One limit is that in the first instance it only speaks to moral reasons to (or to not) colonize space. (That's consistent with there being other kinds of reasons. Perhaps space colonization will be exciting or inspiring.

[3] How long, exactly? Let's set aside interstellar colonization and focus on something less prosaic: colonizing Mars. Again, assume that this colony must be self-sustaining and also fairly sizable—say 1 million people, a relatively modest population compared to most major Earth-based cities. How long would it take to create a colony of this size? Take food production is a minimal proxy: if Cannon & Britt (2019), one hundred years is the lower bound for building a Martian colony capable of feeding one million settlers.

These are reasons to colonize space, maybe even good reasons, but they are not obviously moral reasons.) And with that focus in mind, a central point will not just be the potential s-risks of space colonization, but also space colonization's role as a mitigator of *existential risks* ("x-risks") more generally, including *extinction risks*. How space colonization bears on our having (1) a good future and (2) our having a future at all, is paramount in ascertaining what moral reasons there are for pursuing space colonization to begin with. So, we'll begin by setting out the framework's normative and empirical assumptions (Sect.2), and with them in place, we'll then (Sect. 3) turn to the x-risks, and to what extent they are exacerbated or mitigated in a world where we colonize space.

8.2 The framework

Our framework involves the following core claims:

(1) The value (or disvalue) at stake in choosing among actions that impact the long-term trajectory of civilization is astronomical.

(2) When choosing between options A (e.g., pursuing space colonization) and B (e.g., not pursuing space colonization), a primary (though not sole) determinant of which state of affairs (or "worlds") brought about by A versus B is better is the welfare of the individuals in those distinct state of affairs.

(3) A necessary condition on an individual's being a welfare subject—an individual that can have some welfare level or other in the first place—is that the individual has the capacity for phenomenal consciousness.

(4) In light of (1), whether we should focus on extinction risk mitigation (increasing the probability of there being a future at all) or on s-risk mitigation (increasing the probability of that future is good, assuming we don't go extinct), depends on whether the expected value of the future is positive or negative, which in turn depends in part on how much weight one gives to suffering versus happiness, and the ratio of future happiness to suffering.

I won't defend (2). While I assume a wide, impersonal conception of welfare (recall fn. 2), claim (2) does not require that only welfare matters—I don't assume *welfarism*—and nothing to follow requires any specific account of welfare. We could be hedonists about well-being, we can be satisfactionist's about well-being, or something else still. While this framework might be especially attractive to utilitarians, it does not presuppose any normative ethical framework, utilitarian, or otherwise. Every normative theory places at least some emphasis on choices that optimize welfare, and every normative theory recognizes that we have reasons, even if defeasible, to promote welfare.[4] So, my focus will be on (1), (3), and (4). I take these up in turn.

[4] Rawls (1971: 30), who rejected utilitarianism *qua* normative ethical theory, recognized as much.

8.2.1 The astronomical value of the future

The potential future scale of our civilization is astronomical, with an astronomical amount of value at stake. Extrapolating from the fact that *homo sapiens* have existed for over 200,000 years, Snyder-Beattie et al. (2019) estimate that the expected future lifespan of humanity, bounded only by natural extinction threats, is 87,000 years. However, we are not a typical mammalian species. We have at least some measures of control over our destiny. With a further 1 billion years until the Earth is no longer habitable (Wolf & Toon, 2015); and trillions of years until the last star formations (Adams & Laughlin, 1999: 34), the future is in principle ours for the taking, and one that could be populated with an immense number of future welfare subjects.[5]

These observations have stark ethical implications. They suggest that, given the potential scale of future civilization—and the fact that, in expectation, the vast majority of welfare subjects and the vast majority of what we value, whatever it may be, is in the future—we should be particularly concerned with how our future unfolds. This view is called *longtermism*, and its proponents include Bostrom (2003; 2013), Beckstead (2013); Cowen (2018); Ord (2020), and Greaves & MacAskill (2021).[6] Space colonization, then, should be of special interests to the longtermist, since it is reasonable to assume that the vast majority of expected value (or disvalue) in the future resides only in a world-path that includes colonization, since it is these scenarios that contain, in expectation, the most expected welfare subjects (cf. Bostrom, 2003).

The choice between pursuing large-scale space colonization including (eventually) interstellar colonization and not pursuing space colonization is what Parfit (1984: 356) called a *different people-different number choice*. Such choices are different from *same people choices* and *different people-same number choices*. In same people choices, irrespective of which course of action I pursue, all and only the same welfare subjects live; presumably, whether I snap my fingers next Wednesday has no bearing on the number and identities of the welfare subjects that will live. By contrast, given basic facts about biology, if John and Jane decide to conceive a child today instead of tomorrow, they will have made a different people-same number choice.[7] Yet space colonization—and indeed just

[5] These references are from Greaves & MacAskill (2021). As they emphasize, what matters here is the expected number of future welfare subjects. If we simply assume a 1% chance of civilization surviving beyond the habitability of Earth (in 1 billion years), and assume the same per-century population as today (10 billion per), then in expectation, there are at least 10^{15} people to come (MacAskill & Greaves, 2021). Or, if you think that there's a 10% chance that civilization survives 10 million generations till the end of the Earth, then, in expectation, there will be one million future generations, that is, the future will be at least 100,000 times bigger than the present.

[6] Note that as defined here, longtermism is not the same as *strong longtermism*, which in its axiological variant, says that in a large class of decision situations, the action that is (*ex ante*) best is that whose effects are (*ex ante*) best for the long-term future. This view says that the impact on the far future is the most important feature of our actions today. For present purposes, I don't endorse this stronger view. See Greaves & MacAskill (2021) for discussion, and also a defense of the deontic variant of strong longtermism.

[7] We can assume that Jane and John cannot have more than one child at a time, and that their child, whoever it is, chooses to remain childless.

about any action that makes a significant difference to how the long-run unfolds—is a different people-different number choice: it changes not just who is born but how many people are born. In this respect, space colonization is like the creation of the first transcontinental railroad, just at an even larger scale.

Now, it is well-known that making value comparisons under such different people-different number scenarios is notoriously tricky (Parfit, 1984; Greaves, 2017). To do this is to engage in *population ethics*, and to offer a *population axiology*. It would be a strength of our framework if one could endorse its claim that astronomical value is at stake in deciding whether to colonize space (and in deciding amongst any actions that have impacts on our long-term trajectory) irrespective of one's population axiology, or at least if one could endorse it on the minimally plausible axiologies. Can you?

It seems that you can (Mogenson, 2021; see also Beckstead, 2013). Consider first *totalism*. This is the easy case, for its an additive axiology. This means it treats the value of a state of affairs as a sum of values contributed by each value bearing entity in that state of affairs, taken independently (Tarsney & Thomas, 2020). On totalism, if human beings go extinct within the near future, as opposed to persisting for another 1 billion years with a population of at least 1 billion people at any point in time, very little of the loss in value will be due to the effects on people living at the time of extinction. Most of the value lost is due to the fact that trillions of people who would otherwise have enjoyed (we can presume) on-balance worthwhile lives are never born. Because there are so many potential people, the expected value of any action that lowers the risk of an extinction is astronomical. As such, if space colonization can meaningfully mitigate extinction risks, the expected value of space colonization is astronomical.[8]

Much of the same conclusion holds even if you prefer a nonadditive axiology like *averagism*. On averagism, the value of state of affairs is the sum of the welfare of every individual existing in that state of affairs divided by the number of welfare subjects in that state of affairs. Our generation arguably has the best quality of life in history. We bring up the historical average. If we assume that quality of life will continue to improve, future generations will bring up the historical average even further. As Mogenson (2020: 4) puts it, "when the future population is very big relative to the present, what happens

[8] Totalism is not uncontroversial, as it implies the *repugnant conclusion*. (We can imagine two world-paths where, so long as it has a sufficiently larger population, will always rank higher than the other, even if each of its inhabitants have lives that are just barely worth living, thus making only the minimal positive contribution to the total welfare at that world.) I do not think this is a reason to reject totalism for two reasons. First, our moral intuitions about these scenarios are not obviously trustworthy (e.g., Tannsjo, 2002). Second, *any* population axiology has apparently untoward consequences, whether it is the repugnant conclusion or perhaps something worse. In other words, we have *impossibility theorems* to the effect that, no matter which axiology we choose, we will violate at least some compelling intuitions (Ng 1989; Arrhenius, 2000). For a discussion of the role of the repugnant conclusion in population ethics or value theory more broadly—specifically, why the avoidance of the repugnant conclusion should no longer be a central desiderata for theorizing—see Zuber et al (2021).

over the long-run plays a much bigger role in determining the average welfare of the total population of everyone who ever lives than what happens to present people."[9]

Things get trickier when we look at theories which entail an *axiological asymmetry*, like *person-affecting views*. Put impressionistically, these theories emphasize making people happy, not making happy people.[10] Even if we assume that people will have a high quality of life in the future, a fan of person-affecting views will insist that there is no opportunity cost in failing to create these people. This is because person-affecting views reject any distinction between personal and impersonal value. In other words, they insist that, if outcome *O** is to be better than outcome *O*, then *O** must be better *for someone*. And if outcome *O** is one where the only difference with *O* is that we create far more future people with lives worth living—because, say, we colonize space—then *O** won't be better on that count. It won't be better, the proponent of a person-affecting view will say, because we cannot make positive comparisons about the well-being of a person who lives in *O** but not in *O*. As Broome puts it:

> *If it were better for a person that she lives than that she should never have lived at all, then if she had never lived at all, that would have been worse for her than if she had lived. But if she had never lived at all, there would have been no her for it to be worse for, so it could not have been worse for her.*
>
> ***(1999: 168)***

To be sure, this does not imply that the far future has no value, or that we have no obligations to future generations. But it does imply that insofar as the future has value, and insofar as we have reasons to shape the future, either through space colonization or something else, they will stem from the interests of people alive today, or perhaps those in the past.[11] That puts a dent in our claim about the astronomical value of the future. However, these reasons may be powerful nonetheless, since the expected value of x-risk mitigation could still be quite large on a person-affecting view, even if "quite large" is not astronomical. That may be motivation enough for most people.

Now as described, such person-affecting views are strict insofar as they place zero intrinsic value to creating additional future lives with positive well-being. My own view, following Beckstead (2013), is that this is implausible. More plausible—though still false by Beckstead's (Ibid.) lights—is a moderate person-affecting view. Here, at least some intrinsic value is assigned to the creation of additional welfare subjects with lives worth living, albeit less than improving the lives of those who are currently alive, or the

[9] Like totalism, averagism faces a serious worry. In particular, it faces the *sadistic conclusion*: on averagism, it is better to add some people with negative well-being than to add a larger number of people with positive well-being, assuming the latter is below the prior average.

[10] This slogan comes from Narveson (1973).

[11] For a discussion of the relationship between our obligations to the past and the value of future generations, see Kaczmarek & Beard (2020).

improvement of additional lives. Yet with the numbers on the table—conservatively, 10^{15} future welfare subjects in expectation (see fn. 4)—the astronomical value thesis plausibly stands even on a moderate person-affecting view.

8.2.2 Welfare subjects

We noted earlier that our approach is welfarist only in the sense that a large part of how we evaluate different states of affairs is in terms of the welfare in those states of affairs, understood impersonally. Any framework that hopes to evaluate states of affairs where we pursue colonization (soon, later) compared to those where we do not pursue colonization (ever, many centuries from now) in term of their relative betterness should also have something to say about the following foundational question: what exactly counts as a welfare subject, that is, an entity that has some welfare level or other?

Following Lee [ms], I suggest that for an entity to be a welfare subject, and so have some welfare level or other, if and only if that entity is (capable of being) *phenomenally conscious*. To be phenomenally conscious is roughly to be sentient, for there to be "something it is like" for you to see, smell, hear, feel, and think (Nagel. 1974). This is not the same as being awake or responsive to stimuli; it is for you to have a subjective, experiential perspective on the world. As a theory of welfare subjects, this view is plausible. Although there might be relational reasons for valuing a rock you found as a child, the rock is not, independently of your own sentimental attachments, the kind of thing that can be better or worse off. It makes sense of why there are debates about the ethical status of organisms whose status as conscious entities is unclear—very simple organisms, early fetuses, artificial intelligences, subjects in persistent vegetative states, amongst others (Lee, ms). Consciousness is ethically significant because it is in virtue of being conscious that something can be better or worse off. Suppose there is a planet we want to colonize has microbial life. While the presence of microbial life might present some reasons to not colonize that planet, one such reason won't be that, in a world-path where colonization on this planet occurs, we have actualized a possibility with lower well-being than the relevant alternatives, at least assuming microbial life is not phenomenally conscious.[12,13]

[12] Are there views on which microbial life is conscious? The closest candidate is probably *panpsychism*, simply because on that view everything is conscious in some sense. The more nuanced versions of this view typically distinguish between micro-level consciousness (that had by the fundamental constituents of matter) and macro-level consciousness (that which attaches to more familiar entities like humans, dolphins, etc., only under certain combinations of micro-level consciousness), and it it likely that the former is neither necessary nor sufficient for being a welfare subject (electrons are not welfare subjects). For an overview of the varieties of panpsychism, see Chalmers (2015).

[13] What other reasons might there be to not colonize space if there is a microbial presence? One reason, articulated by Stoner (2017), is that we have a *pro tanto* reason to not disturb objects of scientific value. I have argued that this pro tanto reason is overridden by the expected value of space colonization as an x-risk mitigator (Gottlieb, 2019).

We do not know how wide-spread phenomenal consciousness is, but it is almost certainly not limited to humans. As such, when considering whether colonizing space will lead to welfare outcomes, our concern is appropriately not limited to humanity. With that said, that two entities are welfare subjects does not imply that their welfare should be weighted the same. Indeed, it might be that, while the capacity for phenomenal consciousness is the minimal threshold for being a welfare subject, factors beyond phenomenal consciousness—for example, the possession of self-concepts, the capacity for episodic memory—could be relevant to how much moral weight we give to the well-being of each kind of subject. Suppose it turned out that space colonization would drastically improve the welfare of humanity, but somewhat damage the welfare of *Korgs*, an (as of yet undiscovered) alien species from Alpha Centauri that, while possessive of phenomenal consciousness, lacked a self-concept. If so, then—again, depending on one's views about moral weight—it might be permissible to devalue the interests of korgs in favor of humanity. Devaluing the interests of Korgs does not necessarily imply ignoring them, but it would imply that there could be contexts where our interests win out.

8.2.3 Is the expected value of the future positive?

Recall claim (4) of our framework:

(4) In light of [the astronomical value thesis], whether we should focus on extinction risk mitigation (increasing the probability of there being a future at all) or on s-risk mitigation (increasing the probability of that future is good, assuming we don't go extinct), depends on whether the expected value of the future is positive or negative, which in turn depends in part on how much weight one gives to suffering versus happiness, and the ratio of future happiness to suffering.

If we take (4) and apply it to the case of space colonization (as it relates to x-risks and s-risks), we are asking whether we should focus on making a multiplanetary and interstellar future better if it happens, rather than working to ensure it happens simpliciter (cf. Tomasik, n.d.).

Following Althaus (2018), these questions can be answered by comparing a *normative bads-to-goods exchange rate* ("n-rate") and one's *empirically expected goods-to-bads ratio* ("e-ratio"). Say I have a n-rate greater than 1:1. Then, I believe that reducing bads is more important than increasing goods; if my ratio is at 1:1, then I believe these are of equal importance. Likewise, suppose I have an e-ratio greater than 1:1. Then, I believe that there will be more goods than bads (in expectation) in the future. While this glosses over some complications, the basic idea on Althaus's model is that, the more one's e-ratio exceeds one's n-rate, the higher one's expected value of the future, and the more one favors interventions that primarily reduce extinction risks. By contrast, the more one's n-rate exceeds one's e-rate, interventions that primarily reduce s-risks or otherwise improve the quality of the future without affecting its probability will look more appropriate.

How do you determine your n-rate? Althaus suggests the following thought experiment:

> *Imagine you could create a new world inhabited by x humans living in a utopian civilization free of involuntary suffering, and where everyone is extremely kind, intelligent, and compassionate. In this world, however, there also exist 100 humans who experience extreme suffering.*

We then can ask: what's the smallest value of x for which you would be willing to create this world? The higher x is, the higher your n-rate.

What sort of theory might justify a high n-rate, that is, an n-rate greater than 1:1? The most obvious example is *negative utilitarianism* ('NU'). Here are three variants. On *strong* NU, n-rates are (trivially) greater than 1:1 because only suffering counts. Happiness is neutral. On *lexical threshold* NU, n-rates might be greater than 1:1 because while suffering and happiness both count, there is some amount of suffering that no amount of happiness can outweigh. Finally, on *weak* NU, n-rates might be greater than 1:1 because while suffering and happiness both counts, suffering counts for more.

This is not the place to consider the merits of NU.[14] I am inclined to think that weak NU is the most plausible of the bunch, but absent a very low e-ratio, it's not likely to prejudice the space colonization question anyway. And indeed, on reasonable assumptions, the future will likely see higher e-ratios, perhaps in part because of space colonization.

Here's a familiar argument for why. Some suffering is caused by people who, by their very nature, are sadistic and Michael Caine's Alfred put it, take joy in "watching the world burn," for no specific reason. Other suffering comes from bad ideology. Here, those who cause suffering might believe that they are doing the right thing (motivated as they are by their ideologies), even though this belief is wrong. More still, some suffering is natural: that caused by earthquakes, tsunamis, tornadoes, etc. Finally, much suffering is simply a byproduct of misguided efficiency measures. In these cases, suffering is not caused by people who want to cause suffering, or by those motivated by ideology; rather, they just want some good, and often the easiest way to do this has suffering as a byproduct, of which they are often unaware.

A concrete example of this last category is nonhuman animal suffering, particularly that caused by the factory farm industry. For the most part, I doubt this system exists because people get enjoyment out of seeing cows, pigs, and chickens suffer. That is, the source of this suffering is not evil or sadism. Instead, we want to work as little as possible to get as much of what we want (in this case, meat that tastes good), and as it happens, the most economically efficient way to get this is via a technology (the factory farming

[14] For discussion, see Ord (2013) and Knutsson (2021).

industry writ large) that also just happens to cause a lot of suffering. If we could get what we wanted, and as much as we wanted, without the suffering, we would.

When it comes to e-ratios, the hope is that there will be some future technology that will get us what we want without the suffering. There is already a hint of this with the advent of *in vitro meat*. Chicken, pig, or cow meat that is grown in a lab without the chicken, the pig, and the cow checks both boxes at once: it gives us what we want without causing any suffering. For what we really want is just the pleasurable gustatory experiences that result from eating chickens, pigs, and cows; if we can get those experiences without causing any chickens, pigs, and cows to suffer, why wouldn't we? Indeed, this is plausibly why more people don't eat so-called fake meat like tofurkey. It doesn't cause pleasurable gustatory experiences. It causes unpleasant ones.

Now as things stand, *in vitro* meat has a long way to go along multiple dimensions.[15] But presumably the requisite technology will advance as times goes on, yielding not just a better product, but one that is cheaper and more efficiently produced than that from traditional factory farms. But the main point here is that space colonization will reasonably be a *driver* for such technological advances, given that (1) our desire for the gustatory experiences associated with eating meat will not go away anytime soon, if ever, and (2) the impracticality of attempting to raise chickens, pigs, and cows in a space colony. Of course, laziness-induced suffering is only one form of suffering—and factory farm suffering is only one form of laziness-induced suffering—and nothing said here implies that space colonization will reduce any other forms of suffering. Yet even if future technological progress only reduced factory farm suffering—an unlikely outcome anyway—that's not nothing, especially given that vast number of land animals killed per year in the United States alone is a small slice of the overall pie.[16]

So, there is cause for optimism. Granted, these considerations alone do not show that the expected value of the future is positive.[17] But they do give us some tools for figuring this out. A point to keep in mind as we proceed is this: if one's n-rate sufficiently exceeds one's e-ratio, and one thinks that space colonization will exacerbate s-risks, then one will not want to pursue space colonization even if one thinks it will help mitigate extinction risks.

[15] For a recent review of remaining technical and nutritional issues, see Chriki & Hocquette (2020).

[16] If we treat invertebrates like insects as welfare subjects, then the scale of suffering is orders of magnitude larger. This is not because of the number of insects eaten, but the number of insects that are killed as part of modern farming practices.

[17] For an argument that the expected value of extinction risk reduction is positive, see Brauner & Grosse-Holz (2018).

8.3 The promise and perils of space colonization

8.3.1 Set up

Such is our framework. What does it tell us about the merits of large-scale space colonization? Since the framework is pitched towards considering how space colonization will shape the future, including whether we have a future at all, it will be useful to pair our framework with several further distinctions.

Let an *existential catastrophe* be the destruction of humanity's long-term potential. An existential catastrophe is thus the realization of an x-risk: a risk that threatens the destruction of humanity's long-term potential (cf. Bostrom, 2013; Ord, 2020: 37). In this way, literal extinction is the most straightforward existential catastrophe; if we do not exist, there is no possibility for realizing our long-term potential. But literal extinction is not the only existential catastrophe. For example, an unrecoverable civilization collapse that falls short of literal extinction would count as an existential catastrophe, insofar as it permanently locks in bad values, and prevents us from achieving all the good (however "good" is defined) that we might have otherwise realized (Ord, 2020: 39). The range of x-risks is wide, with some natural, and others not. For example, asteroid and comet impacts are natural x-risks, as are gamma ray bursts and supernovae. Non-natural or anthropogenic risks include runaway climate change, unaligned artificial general intelligence, nuclear war, pandemics (especially engineered ones), nanotechnology, and perhaps totalitarian states.[18] As Bostrom (2013: 15–16) points out, it is important not to conflate our track record of surviving natural x-risks—humanity has survived these for hundreds of thousands of years—with our ability to survive anthropogenic risks. That anthropogenic risks are either relatively new—something like 75 years in the case of nuclear war—or arising from future technological breakthroughs, tells us that our longevity thus far should not be a cause for optimism.

This way of framing things is all fairly standard, but it does make the distinction with s-risks somewhat murky. In fact, s-risks could be a subcategory of x-risks. A *suffering catastrophe* is some event or process that involves suffering on an astronomical scale, for example, vastly exceeding the amount of suffering seen on Earth thus far. So, a suffering catastrophe is (as above) the realization of an s-risk: a risk where an adverse outcome would bring about suffering on an astronomical scale. Understood this way, an unrecoverable dystopia could be a suffering catastrophe so long as we assume that the loss of future value was either accompanied by or realized through suffering on an astronomical scale. Perhaps this is the default mode for dystopias, but we might imagine a scenario in which people don't literally suffer but that still counts as dystopic insofar what we value has been lost. For example, we might imagine a future in which all recreation, art, science, learning has been banished over time, with subsequent generations not realizing

[18] For a full accounting of x-risks (both natural and anthropogenic), see Ord (2020). Terrifyingly, Ord puts the chance of an existential catastrophe striking humanity in the next 100 years at 1 in 6.

what they have lost (i.e., because they have experienced these things, or even heard of them). This scenario is certainly dystopic and would count as an existential catastrophe, but it would not be a suffering catastrophe for future generations. On the other hand, if the existential catastrophe is literal extinction, and the extinction event occurs suddenly and in totality, then it won't be a suffering catastrophe, at least for those who go extinct.

What does an "astronomical scale" of suffering come to? This question is key, since in a sense it is trivial that space colonization will increase suffering since space colonization will greatly expand the number of welfare subjects, that is, being capable of suffering. And it is unlikely that there will be welfare subjects whose lives don't contain at least some suffering. Thus, absent an implausible position on what makes a life worth living or an implausible n-rate, we should not count the mere astronomical increase of welfare subjects as a suffering catastrophe in its own right. Rather, what should count as an "astronomical scale" of suffering is any scenario in which lives are on the whole vastly worse than they would have been otherwise.[19] To take a perhaps fantastical example, we might imagine a scenario in which, upon colonizing the solar system and nearby Alpha Centauri, we come in contact with an alien civilization. Instead of eradicating humanity (the realization of an extinction risk), these aliens enslave all of humanity for their ends, making us live lives replete with back-breaking labor, no joy, and near-constant physical agony.

With these points in place, we can map the situation using the following matrix (Fig. 8.1):

Let the status quo be the baseline level of x-risks (including s-risks and extinction risks) that we would have if we did not colonize space. Let the *chronic distance* between two points be the time it takes to travel between them. Different possible colonization

	Change Relative to Status Quo			
Chronic Distance		Extinction Risk	Non-Extinction X-Risk	Non-Extinction X-Risk/S-Risk
↑	Interstellar colonization	↑↓ —	↑↓ —	↑↓ —
	Intra-solar colonization	↑↓ —	↑↓ —	↑↓ —
	Martian colonization	↑↓ —	↑↓ —	↑↓ —

Fig. 8.1 ***Space colonization and X-Risks.***

[19] Recall that, in different people-different number choices (such as when we opt for space colonization), we can understand "lives on the whole being vastly worse" not in terms of being worse for those people in the scenario, but worse in an impersonal sense. Again, see fn 2.

efforts have different chronic distances: starting from Earth, "Martian Colonization" has a shorter chronic distance than a colonization effort that covers the entirety of the solar system ("Intra-Solar Colonization"), and far shorter than one that reaches to an exoplanet in the Alpha Centauri system ("Interstellar Colonization"). Chronic distances are a function not just of metric distances but also technological factor and contextual factors; if I have to traverse 10 miles via a bike while its snowing, doing so will take me longer than if I drove my car in pleasant weather.

With Kohler (2019), we can think of chronic size as a key determinant on the size of sovereign political organizations. This matters for the question at hand if one thinks that the relationship between space colonization and x-risks is in part of function of potential anarchy and the inability to craft a cosmic Leviathan (Torres, 2018; Duedney, 2020: 237).[20] Now at the outset, any space-based colonies will likely be extensions of Earth-based nations. They won't be independent. If they decide to become independent, a central question will be how this independence is achieved (will it be by violent means?), and what happens after the independence is achieved. So, independence *per se* is not bad, but as chronic distances between independent space settlements grow, there is an increased risk of value drift (itself an x-risk) and an increased level of near-intractability to cosmic collective actions problems (Kohler, 2019).

It is hard to give a precise estimate of the upper limit on the chronic size of a government, but based on a historical analysis of British, Mongol, Roman, and Russian empires, Kohler (Ibid.) suggests that:

> *[A] reasonable upper size limit for a government at about 3 months for one-way communication, resulting in 6 months latency. The reasonable maximum time to transport a significant amount of goods and humans in the form of an army at about 6 months, resulting in 9 months latency.*

Now space is big—very big. So, if we take Kohler's analysis for granted, the *governable universe* is probably around 6 lightmonths from the center, where "governable universe" means that maximum portion of space where governing and collective action is possible. This makes, for example, a colony on Mars fall well within the range of such a governable universe, but, say, a colony near Promixa Centauri well without—Promixa Centauri being the star closest to the Sun at 4.2 lightyears away.[21] Placing the seat of an interstellar government in say, the inner Kuiper belt—a problematic proposal in its own right—wouldn't make a difference, as that's "only" 50 AU from the Sun.

[20] Is the anarchy under consideration only top-down, or is there a possibility of bottom-up anarchy too? Given the inherent instability of (at least early) space colonies, the latter is less likely. That is, while chronic distances might make inter-settlement policing and governance difficult if not impossible, strict *intra-settlement* policing and governance will perhaps be the default, at least at the outset, given the realities of space environments. We already expect the curtailment of some individual freedoms in space colonies (Cockell, 2015; Milligan, 2015; Schwartz, 2016).

Of course, as noted, the upper limit on the chronic size of government is also a function of technology. We cannot travel faster than the speed of light, but one might wonder whether even an achievement of 25% of the speed of light might help. Unfortunately, it would not help much. This is so even with exotic technologies like an Alcubierre drive which, though not enabling faster-than-light travel through space, permits faster-than-light travel by causing the space ahead of the ship to contract while expanding the space behind, letting the ship go along for the ride. As Kohler (Ibid.) notes, even if we managed to exceed the speed of light by 50% via an Alcubierre drive (or something else), this would still be insufficient for a bistellar government between our sun and Proxima Centauri. A one-way trip would take "only" about 32 months.

In any event, this should be enough background to begin our analysis. How does space colonization—in its various guises—fare when it comes to x-risk mitigation, exacerbation, and creation? We'll begin by looking at the promise of space colonization, and then turn to the perils, with a brief detour to non-welfarist considerations along the way.

8.3.2 The promise

Let's begin with the upside. There is an intuitive case for space colonization that trades on space colonization (in the requisite sense) creating *catastrophe independent* pockets of humanity, where a "pocket" of humanity is independent from another pocket if a catastrophic event in one pocket doesn't necessarily cause a catastrophic event in another (cf. Abney, 2019; Green, 2019). The case goes like this:

B1. For a catastrophic event E to cause either civilizational collapse or literal extinction, every pocket of humanity must be impacted by E.

B2. By making the pockets of humanity catastrophe independent, we vastly lower the likelihood that every pocket of humanity is impacted by E.

B3. Space colonization would make the pockets of humanity catastrophe independent.

B4. Therefore, space colonization will vastly lower the likelihood that every pocket of humanity is impacted by E.

There are several points to make about this argument. First, for most any two pockets P_1 and P_2, P_1 and P_2 won't be catastrophe independent for every catastrophic event. For example, if on pocket P_1 (e.g., Earth), a high-energy physics experiment results in a supermassive black hole, this might also cause a catastrophic event on pocket P_2 (e.g., the Moon). However, this is where the distance between colonies actually helps;

[21] Mars is within the chronic window of 6 months even with current technology, although this is complicated by launch windows. In 2003, Mars and Earth were 34.8 million miles apart, although the average distance between them is closer to 140 million miles. For some perspective, it took the Mariner 7 128 days to reach Mars for its 1969 flyby.

presumably, an interstellar colonization effort will provide a greater degree of catastrophe independence than a colonization effort that is limited to the moon and Mars, save some fantastical false vacuum technology, or a nearby supernova.

Second, suppose we see x-risk mitigation as a *knapsack problem* (Baum & Barrett, 2018). On this view, the overall goal of minimizing x-risk is best conceptualized as a matter of picking the right range of projects, relative to their potential impact, given a fixed budget. This engenders a worry. For even if space colonization reduces x-risks, it may just be too expensive relative to that benefit, given the fixed budget, whatever it is. So, while B3 may be true, this does not mean we should pursue space colonization for x-risk mitigation.

Space colonization is indeed expensive. But help is on the way. One important factor is cost to LEO. Lowering the cost to LEO is a minimal necessary condition on space colonization. The near-full reusability of SpaceX's rockets has led to impressive inroads here, reducing launch costs by a factor of 20 via the Falcon 9 (Jones, 2018), with an ambitious target of $10/kg for the still in-development Starship. If the actual costs are even remotely in the vicinity—keep in mind that NASA's space shuttle was priced at $54,500/kg—this would go some way towards significantly reducing overall costs.

In addition, even assuming x-risk mitigation is a knapsack problem, and even assuming that our primary motivation for space colonization is x-risk mitigation, this does not mean that all of the financing for space colonization (*qua* x-risk mitigator) will come from the same fixed budget as other x-risk mitigation measures. As Ord (2020: 394) reminds us, given the inherently inspirational nature of space colonization, this is unlikely anyway. Indeed, it might be—as Ord also points out—that space colonization actually increases our total budget for x-risk mitigation measures.

Note also that there is nothing in this argument which assumes that we must begin colonizing space now. For all the argument says, perhaps it is better to wait till we are a little wiser, or till we have developed yet more advanced technology that makes colonization easier and cheaper, such as space elevator or a better form of propulsion. Whether we should begin colonizing space anytime soon depends on (of course) whether we should colonize space at all, but also—and especially relevant given the present emphasis—whether we have reason to believe that we are at what Parfit (2011) called the *hinge of history*, that is, an especially critical part of human history.[22] If, for example, we think that we are in a period of especially high levels of extinction risk, perhaps due to the advent of some novel and novelly dangerous technology, then we are more pressed for

[22] See MacAskill (2021) for an extended analysis of this notion, and an argument that we likely are not living at the hinge of history. As MacAskill (Ibid.) notes, for the hinge of history thesis to be true, it would not only have to be the case that we live during an exceptionally high period of x-risks, but that, during this period, we are also especially able of doing something to eliminate these risks. Thus, if one thinks that space colonization is a good x-risk reducer, and one also thinks that we live at the hinge of history, then there is extra pressure to invest now in the technology that would make large-scale space colonization feasible.

time. Yet this is not essential to the longtermist perspective; one can agree that how we shape the future is of paramount importance, and that much of any future value resides in world-paths where we colonize space, yet still countenance *patient longtermism*. The time for such action will come, but perhaps it is not yet here, and it might be better to invest present resources for the future.[23]

Third, and finally: whether B3 is true depends on the E we are interested in. The risk of some E's might be correlated across the pockets or across the space settlements. Ascertaining the level of risk correlation in the abstract is difficult, given that the strength of these correlations will often depend on what these pockets are like and the relationships between the pockets. For example, as noted, there might be lower risk correlations across pockets that are sufficiently distant from one another—so risk correlation will look different for colonization efforts that are limited to Mars from those that consume the whole solar system, or are interstellar. One plausible exception to this is unaligned AGI. As Ord (2020: 194) remarks, if it impacts Earth, it will likely impact every off-Earth colony. However, all this shows is that space colonization is not sufficient to mitigate all x-risks. No one thought that anyway.[24] When thinking about the relationship between space colonization and x-risks, we should view space colonization as removing the uncorrelated portion of our aggregate risk (cf. Ord, 2020: 394).

This leads us to a preliminary, conditional result. If space colonization is a necessary component for mitigating extinction and civilizational collapse portions of the x-risk space, and our e-ratio exceeds our n-rate, space colonization will look especially attractive. However, we can go further. We have not yet considered how space colonization bears on *s*-risks specifically, save our earlier point that colonization might serve as a driver for humane technologies like *in vitro* meat. And while in the next section we'll consider how space colonization might exacerbate and create s-risks (along with other x-risks), it's also possible that space colonization will reduce s-risks, independent of its bearing on extinction and civilizational collapse.

Suppose you think, like Cowen (2018: 30, 40–48), that insofar as we care about positively shaping future well-being (and so reducing future suffering), the one goal we should advance above all is maximizing the sustainable rate of economic growth, understood in terms of *wealth plus*. According to Cowen, this entails "the traditional measures of economic value found in GDP statistics, but…also leisure time, household

[23] Being patient in this regard might help colonization efforts anyway, as advanced technology will presumably make the process less time-consuming and more efficient. Recall the lower-bound estimate for simply creating a self-sufficient Mars colony (see fn. 3).

[24] Schwartz et al. (2021) put this point nicely: "…space settlement should not be pursued as a replacement for attempts to mitigate threats to humanity's existence on Earth. If we conceive of the duty to extend human life as one that attaches to humanity *in toto* (that is, to all humans, rather than a proper subset), then space settlement is one tool among many that must be pursued to satisfy this duty."

production, and environmental amenities." If that's right, we might consider whether space colonization facilitates our maximizing economic growth. Consider asteroid mining. According to Asterank, the most valuable target is *Davida*, a c-type (carbonaceous) asteroid with a semi-major axis of 3.165 AU—more than three times greater than the distance between the Earth and the Sun.[25] Davida has a diameter of 326 km, and is estimated to be worth greater than 100 trillion USD. By comparison, global GDP was approximately 84.54 trillion USD in 2020. This example is of course extreme—both in terms of the value and the size of the asteroid itself—but its extremeness underscores that accessing anything remotely like this requires a measure of infrastructure that would likely be impossible without some form of space colonization making the retrieval of high value asteroids easier and more efficient.[26]

8.3.3 The perils

8.3.3.1 A brief detour on nonwelfarist considerations

In asking whether we should colonize space, we have focused on the impact of colonizing space would have on welfare. This is not strictly an anthropocentric focus, since we have not assumed that the only welfare subjects are humans. Still, one might insist that nonwelfarist considerations not only carry weight—a point consistent with this essay's framework—but that they carry significant weight, so much that colonization efforts could be stopped in their tracks.

Consider what Cockell says about microbes:

> *Microbes have intrinsic worth equal to, if not greater than, that of any other species. They have instrumental worth so great that they are vital to our survival and the survival of all other multicellular organisms on Earth.*
>
> ***(2004: 149)***

Cockell makes two claims here—one about intrinsic worth, and one about instrumental worth. Let's focus on the former. Perhaps Cockell thinks otherwise, but I'll just state flat-footedly here that microbes are not welfare subjects. Thus, they do not count insofar as we are evaluating outcomes in terms of the welfare in these outcomes. So, if they have intrinsic worth all the same—or moreover, as Cockell shockingly suggests, greater intrinsic worth than humans—it must be for some other

[25] Asterank is a database of asteroids, including information about mass, composition, and orbital parameters. Asterank uses JPL's Small Body Database and the Minor Planet Center as resources. Per Asterank, asteroid value estimates are function of asteroid mass and spectral type. Profit and ROI calculations are a function of accessibility and value, with mining costs factored in.

[26] On at least one analysis, an approach in which asteroid mining is conducted only via multiple small spacecraft (capable of reaching near-Earth asteroids within a range of ~.03 AU), breaking even financially would require one hundred spacecraft and their successful operation for over 5 years. See Calla et al. (2018). Asteroid mining would no doubt aid space colonization as well, as any serious space colonization effort will require that we harvest resources in space.

reason. But Cockell does not provide any other reason, and having intrinsic worth is not a brute fact. There must be some property F which, by virtue of having F, microbes have intrinsic worth. We know what F is in the case of welfare subjects; it is the property of having the capacity for consciousness. (For comparison, Kant appealed to rational capacities.) Absent some other grounds of intrinsic worth for microbes, there is little reason to take Cockell's claim seriously. If we care about the microorganisms in our gut flora, it's not because they have value for their own sake, but because they are important for digestive health. What exactly is the moral difference between microorganisms in our gut flora and those (such as there are any) on Mars?[27] It's hard to see how there could be any. True, microorganisms have instrumental value, but things with merely instrumental value don't have rights; their value is *value for us*, and in that respect, we can to some extents do what we want. Certainly then, there is little reason to let the presence of microbes impede colonization efforts ad infinitum.

Now notice that this does not imply that the presence of microbes might give us some reason to colonize with more caution than we would otherwise. For example, if we found microbes on Mars, their presence would be a monumental scientific discovery, and insofar as it is prudent to delay colonization in order to adequately study these microbes, perhaps we should do so. Of course, what counts as prudent will depend on the urgency of the risks we face. If we know that an asteroid is going to obliterate all life on Earth in 20 years, no measure of scientific interests should delay colonization efforts (Gottlieb, 2019). We must also realize that scientific interests degrade over time (Smith, 2016: 197). At some point, we will have learned all there is to know (if there is anything to know) about Martian microbes, and when that time comes, it's not clear why we should not colonize Mars if it is in our best interests, forthcoming asteroid or not.

I think much of the same applies to the idea that sterile worlds have a right to remain sterile, and free from the human contamination that is inevitable from colonization. Indeed, it's unclear what "right" could even mean here, since it's (at best) unclear how a sterile world could be a rightsholder. This does not mean that all of our actions in a sterile world are ipso facto morally permissible. Just as it would be morally grotesque, and failure to exercise proper environmental virtue, to fill the floor of the Grand Canyon with a giant shopping mall, it would be morally grotesque, and failure to exercise proper environmental virtue, to ring Olympus Mons with a giant roller coaster. Just as the moral grotesqueness of the former is not contingent on the Grand Canyon being

[27] See Smith (2016) for an extended argument against what he calls "Mariomania," the idea that Martian life, even if of the simplest variety, has moral worth at least on par with that of humanity. Cockell (2004) does not explicitly state that that, if microbes exist on Mars, we ought not colonize Mars, and more recent work (Cockell, 2011) seems to push against that conclusion; but the mere possibility that they have "intrinsic worth equal to, if not greater than, that of any other species" seems to at the very least cast doubt on the moral tenability of colonization.

sterile (it's not), neither is the moral grotesqueness of the latter so contingent. Yet there is nothing about space colonization as such that requires or encourages such actions.

8.3.3.2 Welfarist considerations

Back now to welfarist considerations. What are the perils of space colonization here? Could space colonization have catastrophically negative outcomes? Could it—instead of safeguarding the future and flourishing of sentient beings—create a future rife with astronomical amounts of suffering? The affirmative position is taken by Daniel Duedney:

> *Given the many obstacles that render successful space colonization at scale very improbable, it is extremely imprudent for humanity to view space colonization as a hedge against catastrophe or extinction. With low probabilities of success and significant moral hazards and opportunity costs, space colonization should not be considered a viable response to severe Earth problems and troubles.*
>
> ***(2020: 337).***

While space advocates propose a variety of ways space expansion might alleviate or escape existing risks, they given almost no attention to whether the expansion might generate new risks or help re-activate already regulated ones…[T]here are six major ways in which the realization of the space expansionist agenda for solar orbital space generate or activate catastrophic and existential risk. Taken in combination these arguments provide a strong basis for putting ambitious space expansion on the list of megathreats (Ibid.: 357).

According to Duedney (Ibid.: 358) and rephrasing a bit, these "six major ways" include:

S1. Malefic geopolitics: space colonization will create new opportunities for increasingly violent war.

S2. Natural threat amplification: space colonization will exacerbate natural x-risks.

S3. Restraint reversal: space colonization will exacerbate anthropogenic x-risks via contributing to various forms of deregulation.

S4. Hierarchy enablement: space colonization will exacerbate the risk of totalitarian rule.

S5. Alien generation: space colonization will exacerbate extinction risk due to species radiation and extreme cultural and ideological diversification—creating 'aliens' in both literal and figurative sense.

S6. Monster multiplication: space colonization will exacerbate the risk of (as of yet) unknown "monsters" via the concurrent development of potentially hazardous technology.

Duedeny is not the only one to argue along these lines. For instance, Torres (2018) also emphasizes Malefic Geopolitics and Alien Generation (albeit with different terminology)

in his argument that space colonization would produce astronomical amounts of suffering.[28] While for simplifying purposes I'll organize my discussion around Duedney's presentation, I suspect that what follows will apply equally to Torres'.

Let's consider Natural Threat Amplification first. Asteroid strikes are a natural x-risk, and depending on their size, a natural extinction x-risk specifically. It is a natural x-risk because the rate at which they collide with Earth is typically unconnected with humanity's activities. Duedney fears that, as we colonize space, these risks will be exacerbated and become non-natural. The idea here is that asteroid deflection technology—on its own a good thing—can be made dangerous in ways not possible prior to the advent of spaceborne societies. As Duedney notes, this technology can also be used to direct asteroids toward Earth from strategic ground that is as 'high' as it gets (Duedney, 2020: 359).

The response to this point is not, of course, to say that we should abandon asteroid deflection technology. Such technology will be a component of any x-risk mitigation plan. Moreover, space colonization would plausibly enhance the effectiveness of this technology. Here's an analogy. To protect the homeland, a country does not simply arrange their defenses tightly around the circumference of their nation. Instead, they disperse their defenses strategically, expanding the scope of their defensive capabilities. Consider now something like the recently successful DART ("Double Asteroid Redirection Test"') program, whose target was the binary near-Earth system Didymos (comprised of a primary body 780 meters across, and a secondary body approximately 160 meters across). We can compare the effectiveness of a DART-like program that launches exclusively from Earth with a DART-like program that launches from various points throughout the solar system. The latter program will be able to reach more targets and reach these targets faster, and is only possible with at least some colonization infrastructure. Of course, most asteroid threats have long lead times, but in principle they need not. So even if space colonization would increase the threat of intentional asteroid strikes, it's not at all clear that space colonization would increase the risk of asteroid strikes *period*, since any increase in the former needs to be compared to the potential decreases in non-intentional asteroid strike risk.

Now, would space colonization in fact increase the threat of intentional asteroid strikes, be it on Earth or elsewhere? It's not obvious if we can disentangle this question from Malefic Geopolitics and Alien Generation. It's more natural to read this matter as an exemplification of Malefic Geopolitics and Alien Generation; a more general increased tendency towards war, and the creation of humans with very different values from us and wholly alien post-humans, may very well lead to Earth being destroyed by asteroid bombs.

While we should be careful in putting too much weight on fictional scenarios, it remains useful, if only as a hypothetical, to consider in this context the 'Belters' in *The*

[28] For a response to Torres, see Ćirković (2019).

Expanse series of novels written by James S. A. Corey.[29] Belters are persons born in the asteroid belt colonies, or on the moons of the outer planets, like Europa. Belters are phylogenetically different from humans born on Earth, and they are at war with Earth-born humans throughout much of *The Expanse*. Part of the issue here is, again, chronic distances; even with advanced propulsion technology, these outer colonies are just so far that solar-system wide governance is all but impossible in *The Expanse*. But the issue comes to more than this, since the Belters don't want to be part of a government with Earth at the center. Their values are just too different. The distance, it seems, makes the Earthers seem less human to the Belters, and vice-versa. And eventually—spoiler alert!—things come to a head; the Belters, and their "free navy," bombard Earth with asteroids.

So, I suspect that there really may be an issue here. The real question, however, is how this compares to the status quo. Take Alien Generation. Presumably, speciation requires off-Earth colonization, but the Alien Generation hypothesis is not limited to this; nanotech and cybernetic modification will happen irrespective of whether we pursue space colonization, and it's unclear why space colonization would make these things more dangerous. Indeed, space colonization might make them less dangerous insofar as it would enable largely insulated research facilities. Metric (and so chronic) distances help here.

On the other hand, even if chronic distances provide some insulation from the treats of Alien Generation, the Restraint Reversal hypothesis suggests that such distances inherent in space colonization will prompt us to take on more risk, and remove the restraints on potentially dangerous technologies that would be in place if we were Earth-bound. Duedney (2020: 360) cites artificial superintelligence ("ASI") as an example. This choice is curious. Now it's plausible that ASI can help colonization efforts, especially interstellar ones. And as we saw, risks from AI (including AGI and ASI) are likely correlated across settlements. Still, it's unclear why space colonization would make us relinquish any legal constraints we might otherwise have on ASI development. For one thing, precisely because ASI is a potential threat irrespective of one's location, off-Earth colonists should know that their off-Earth status provides them no comfort. Second, despite the role of ASI as an x-risk, a well-aligned ASI could potentially mitigate x-risks, including s-risks (Bostrom, 2013). As such, there is reason to think that ASI, if developed at all, will be developed either prior to or alongside space colonization (as part of the "knapsack"), thus making restraint reversal in space moot. The cat will likely already be out of the bag.

What of Monster Multiplication? To some extent, this is no doubt true. Space colonization will result in as of yet unforeseen risks simply because, in colonizing space, we expand the scope of what is (technologically, not nomologically or metaphysically) possible. Kovic (2021) calls these "aberration risks," and as an example, has us imagine a scenario in which Earth has 10 billion people living roughly as good of lives as they

[29] "James A. Corey" is the joint pen name for Daniel Abraham and Ty Franck.

live today, while 2 billion Venusians suffer from chronic respiratory issues due to a failed terraforming effort, and 500,000 Martians are even worse off, suffering from chronic muscular and skeletal disease, along with DNA damage, due to constant radiation exposure. This scenario counts as a case of Monster Multiplication since the health risks of a (e.g.,) failed terraforming are risks that are not possible on a world-path that does not include colonization.

Of course, if a scenario such as Kovic's came to be, that would be an unhappy result. But what is the lesson here, exactly? If the idea is that we should expect similar scenarios to become common in a colonized world-path—or, more generally, that the inhabitants in a colonized world-path will have much worse lives on average than those in noncolonized world-paths—then this tells against colonization only insofar as one has a sufficiently high n-rate. So long as the lives of the Martians and Venusians are still worth living, most reasonable moral theories (outside of strict person-affecting views) will say that the world-path in which Mars and Venus is colonized is still the better of the two.

I'll wrap up this portion of our discussion with a brief note about freedom and other liberties in colonized space. It is often recognized that the harsh realities of life in space will place severe limits on everyday liberties that, as citizens of Earth, we frequently take for granted (see fn. 20). Is this a potential s-risk? I am doubtful. Many but not all of the things we intuitively take to be part of our "freedoms" or one of our "rights" are to some extent contingent—only in place because of circumstantial facts about the world as we find it.[30] What this means is that we should not import our Earth-bound conceptions of freedoms and rights to space-based life. This is certainly so at the outset, when colonies are just getting off the ground, but may even hold (albeit to a lesser extent) as colonies mature. My point is this: when it comes to suffering from the loss of individual liberties and rights, using Earth as our yardstick is a substantive and perhaps false move.

8.4 Conclusion

I have purposely avoided giving an explicit answer to our titular question. If the question "should we colonize space" were given not a moral bent, but one where we were just asked to envision the future, we personally find most exciting, my answer would be a resounding "yes." As for the moral question that has been our focus here, if pressed, I'd say "yes, we should colonize space." However, this answer depends on several assumptions—moral and empirical—that may or may not be correct. And my credence in those claims is not high enough to feel especially confident about the moral question. However, my hope is that the framework provided here will at least tell us which assumptions we need to make to figure out the moral question (or more modestly, one take on the moral question). Thus, more work is needed, but knowing where that work lies is a good first step.[31]

[30] This is one of the lessons of Mulgan's (2011) "Ethics for a Broken World."

[31] For helpful comments and discussion, I thank Keith Abney, Les Johnson, Saja Parvizian, Ken Roy, and Kelly Smith.

References

https://forum.effectivealtruism.org/posts/CmNBmSf6xtMyYhvcs/descriptive-population-ethics-and-its-relevance-for-cause#Normative_and_expected_ratios__aka_exchange_rates_and_future_optimism_, Access date: October 2, 2021.

https://medium.com/@KevinKohlerFM/cosmic-anarchy-and-its-consequences-b1a557b1a2e3, Access date: 10/4/21.

Abney, K., 2019. Ethics of colonization: arguments from existential risk. Futures 110, 60–63.

Adams, F.C., Laughlin, G., 1999. The Five Ages of the Universe: Inside the Physics of Eternity. Free Press, New York, NY.

Althaus, D.., 2018. Descriptive population ethics and its relevance for cause prioritization.

Arrhenius, G., 2000. An impossibility theorem for welfarist axiologies. Eco. and Philos. 16 (2), 247–266.

Baum, S., Barrett, T., 2018. Towards an Integrated Assessment of Global Catastrophic Risk. In: Garrick, B.J. (Ed.), *Proceedings of the First International Colloquium on Catastrophic and Existential Risk*, Garrick Institute for the Risk Sciences, University of California. Los Angeles, pp. 41–62.

Beckstead, N., 2013. On the overwhelming importance of shaping the far future. PhD thesis, Rutgers University. Available at https://rucore.libraries.rutgers.edu/rutgers-lib/40469/PDF/1/play/, Access date: 8/19/21.

Beckstead, N., 2014. Will we eventually be able to colonize other stars? Notes from a preliminary review. https://www.fhi.ox.ac.uk/will-we-eventually-be-able-to-colonize-other-stars-notes-from-a-preliminary-review/. Access date: 8/19/21

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.

Brauner, J., Grosse-Holz, F.M., The expected value of extinction risk reduction is positive. https://forum.effectivealtruism.org/posts/NfkEqssr7qDazTquW/the-expected-value-of-extinction-risk-reduction-is-positive.

Broome, J., 1999. Ethics Out of Economics. Cambridge University Press, Cambridge, UK.

Calla, P., Fries, D., Welch, C., 2018. Asteroid mining with small spacecraft and its economic feasibility. https://arxiv.org/abs/1808.05099. Access date: 10/10/21

Cannon, K.M., Britt, D.T., 2019. Feeding one million people on mars. New Space 7 (4). https://doi.org/10.1089/space.2019.0018.

Ćirković, M.M., 2019. Space colonization remains the only long-term option for humanity: a reply to Torres. Futures 105, 166–173.

Cockell, C.S., 2004. The rights of microbes. Interdisciplinary Science Reviews 29, 141–150.

Cockell, C.S. (Ed.), 2011. Ethics and extraterrestrial life. In: Landfester, U. (Ed.), Humans in Outer Space: Interdisciplinary Perspectives. Springer, New York, pp. 80–101.

Cockell, C.S., 2015. Human Governance Beyond Earth: Implications for Freedom. Springer, Switzerland.

Cowen, T., 2018. Stubborn Attachments. Stripe Press, San Francisco, CA.

Chalmers, D., 2015. Panpsychism and Panprotopsychism. In: Alter, T., Nagasawa, Y. (Eds.), Consciousness in the Physical World: Perspectives on Russellian Monism. Oxford University Press, New York, pp. 246–276.

Chriki1, S., Hocquette, J.P., 2020. The myth of cultured meat: A Review. Frontiers in Nutrition 07. https://doi.org/10.3389/fnut.2020.00007.

Deudney, D., 2020. Dark Skies: Space Expansionism, Planetary Geopolitics, and the Ends of Humanity. OUP, Oxford, New York, NY.

Gottlieb, J., 2019. Space colonization and existential risk. J. Am. Philos. Assoc. 5 (3), 306–320.

Gottlieb, J., 2022. Discounting, Buck-Passing, and Existential Risk Mitigation: The Case of Space Colonization. Space Policy 60. https://www.sciencedirect.com/science/article/abs/pii/S0265964622000121.

Greaves, H., 2017. Population axiology. Philos. Compass 12 (11), e12442. https://doi.org/10.1111/phc3.12442.

Greaves, H., MacAskill, W., 2021. The case for strong longtermism GPI Working Paper No. 5-2021. https://globalprioritiesinstitute.org/hilary-greaves-william-macaskill-the-case-for-strong-longtermism-2/</. Access date: 9/1/21..

Green, B.P., 2019. Self-preservation should be humankind's first ethical priority and therefore rapid space settlement is necessary. Futures 110, 35–37.

Jones, H., 2018. The recent large reduction in space launch cost. Infor. Int. Conference on Environ. Systems (ICES). https://ntrs.nasa.gov/citations/20200001093. Kaczmarek, P., Beard, S., 2020. Human extinction and our obligations to the past. Utilitas 32 (2), 199–208. Access date: 8/1/21.
Kovic, M., 2021. Risks of space colonization. Futures 126. https://doi.org/10.1016/j.futures.2020.102638.
Knutsson, S., 2021. The world destruction argument. Inquiry 64 (10), 1004–1023. http://doi.org/10.1080/0020174X.2019.1658631.
MacAskill, W., 2021. Are we living at the hinge of history? In: McMahan, J. et al. (Eds.), Ethics and Existence: The Legacy of Derek Parfit. Oxford University Press, Oxford (forthcoming).
Milligan, T., 2015. Rawlsian deliberation about space settlement. In: Cockell, C. (Ed.), *Human Governance Beyond Earth*. Space and Society. Springer, Cham.
Mogenson, A. 2021. Moral demands and the far future. Philos. Phenomenol. Res. 103 (3), 567-585.
Mogensen, A.L., 2020. Moral demands and the far future. Philos. Phenomenol. Res. https://doi.org/10.1111/phpr.12729.
Mulgan, T., 2011. Ethics for a Broken World: Imagining Philosophy after Catastrophe. Acumen Publishing, New York, NY.
Ng, Y-K., 1989. What should we do about future generations? Impossibility of Parfit's Theory X. Economics and Philosophy 5 (2), 135–253.
Nagel., T., 1974. What is it like to be a bat? The Philos. Rev. 83 (4), 435–445.
Narveson, J., 1973. Moral problems of population. The Monist 57 (1), 62–86.
O'Neil, G.K., 1976. The High Frontier: Human Colonies in Space. Apogee, Burlington, VM.
Ord, T., 2013. Why I'm Not a Negative Utilitarian. http://www.amirrorclear.net/academic/ideas/negative-utilitarianism/. Access date: 9/6/21.
Ord, T., 2020. The Precipice: Existential Risk and the Future of Humanity. Hachette Books, New York, NY.
Parfit, D., 1984. Reasons and Persons. Oxford University Press, Oxford.
Parfit, D., 2011. On What Matters. Vol. 1. Scheffler, S. (Ed.), Oxford: Oxford University Press.
Rawls, J., 1971. A Theory of Justice. Belknap Press, Cambridge, Mass.
Schwartz, J.S.J, 2016. Lunar labor relations. In: Cockell, C. (Ed.), Dissent, Revolution, and Liberty in Space. Springer, Switzerland, pp. 41–58.
Schwartz, J.S.J., Wells-Jensen, S., Traphagan, J.W., Weibel, D., Smith, K., 2021. What do we need to ask before settling space? J. Br. Interplanet. Soc. 74 (4), 140–149.
Smith, K.C., 2016. The curious case of the martian microbes: mariomania, intrinsic value and the prime directive Schwartz & Tony Milligan. In: James, S.J. (Ed.), The Ethics of Space Exploration. Springer, pp. 195–208.
Stoner, I., 2017. Humans Should Not Colonize Mars. J. Am. Philos. Assoc. 3 (3), 334–353.
Snyder-Beattie, A., Ord, T., Bonsall, M.B., 2019. An upper bound for the background rate of human extinction. Sci. Rep. 9, 11054.
Tännsjö, T., 2002. Why we ought to accept the repugnant conclusion. Utilitas 14 (3), 339–359.
Tarsney, C., Thomas, T., 2020. Non-additive axiologies in large worlds GPI Working Paper 9-2020. https://globalprioritiesinstitute.org/christian-tarsney-and-teruji-thomas-non-additive-axiologies-in-large-worlds. Access date: 8/4/21.
Tomasik, B. (n.d.). Risks of astronomical future suffering. Foundational Research Institute. https://foundational-research.org/risks-of-astronomical-future-suffering/. Access date: 8/3/21.
Torres, P., 2018. Space colonization and suffering risks: reassessing the "maxipok rule". Futures 100, 74–85.
Wolf, E., Toon, O.B., 2015. The evolution of habitable climates under the brightening Sun. J.f Geophys. Res.: Atmos. 120, 5775–5794.
Zuber, S., et al., 2021. What should we agree on about the repugnant conclusion? Utilitas. http://doi.org/10.1017/S095382082100011X.