Futures 100 (2018) 74–85

Contents lists available at ScienceDirect

Futures

journal homepage: www.elsevier.com/locate/futures

Space colonization and suffering risks: Reassessing the "maxipok rule"

Phil Torres

Project for Future Human Flourishing, 849 South 7th St., Apt. 4A, Philadelphia, PA, 19147, United States

ARTICLE INFO

Keywords:
Suffering risk
Existential risk
Space colonization
Security dilemma
Hobbesian trap
Space weapons

ABSTRACT

This article argues that, contra Bostrom (2003), every second of delayed space colonization could be immensely desirable; indeed, the longer the delay, the better, with the best outcome being no colonization at all. The argument begins by hypothesizing that expansion into space will generate a wide variety of distinct species, many having their own cultural, political, religious, etc. traditions. Next, the paper offers reasons for expecting catastrophic conflicts between different civilizations, both near and far, to be the default outcome. Third, it examines some strategies for mitigating conflict, including (i) the establishment of a "cosmic Leviathan" that is capable of imposing law and order within the cosmopolitical arena, and (ii) the implementation of policies of deterrence to prevent one civilization from attacking another. Both of these strategies appear problematic, though, due to (a) fundamental physical limitations on the speed of space travel and the transfer of information, and (b) the advanced weaponry that future civilizations will almost certainly have at their disposal. The conclusion is that colonizing our solar system, galaxy, and beyond will engender a Hobbesian predicament in which all actors are perpetually in fear of being destroyed—that is, when they aren't engaged in devastating wars with their neighbors.

1. Introduction

The *astronomical value thesis* states that the potential value of the future is astronomically large (Author, 2017). This is based in part on calculations of how long our civilization could last in the universe and the vast number of people who could occupy our future light cone. For example, whereas *Homo sapiens* has so far existed for about 2000 centuries, Earth could remain habitable for another 10 million centuries, or 1 billion years.[1] To put this in perspective, if we survive this long, contemporary humans could be a mere 0.0006 percent into writing our story—hardly a word into the prologue.[2] Mapping this onto the annual calendar, it means that humanity is slightly more than 3 min into the first hour of January 1. Now consider that the universe will remain habitable for *trillions* of years—placing us mere nanoseconds past the hour. As for the future population of humanity, Sagan (1983) argued that if our species survives for another 10 million years, we could expect some 500 trillion people to come into existence. A more recent estimate suggests that if Earth's population remains above 1 billion people with lifespans of "normal duration," a ten million billion, or 10^{16} people could inhabit the planet before the sun becomes a bloated red giant. If we colonize space, though, there could be upwards of a hundred thousand billion billion billion, or 10^{32} people. Even more, if whole-brain emulation (or mind-uploading) becomes feasible, entire planets could be converted into supercomputers that run simulations full of conscious beings. Within a single century, our local

E-mail address: philosophytorres@gmail.com.

[1] Following the Long Now Foundation, we could identify our current year as "0000002018" to emphasize the habitable time left on Earth.

[2] If we stipulate that the beginning of recorded history, about 4000 BCE, marks the origin of civilization. Also, here I gesture at Russell's chapter "Prologue or Epilogue?" in Human Society in Ethics and Politics (1954).

https://doi.org/10.1016/j.futures.2018.04.008
Received 1 November 2017; Received in revised form 13 April 2018; Accepted 20 April 2018
Available online 30 April 2018
0016-3287/

supercluster could house some 10^{38} lives—a truly astronomical figure (Bostrom, 2013).

Based on the astronomical value thesis, Bostrom (2003), following Ćirković (2002), proposes the "astronomical waste argument." The conclusion of this argument is that, insofar as one accepts a value theory that rejects time discounting of future lives and includes an aggregative evaluation function, we have two primary moral obligations. First, we should make "the objective of reducing existential risks … a dominant consideration whenever we act out of an impersonal concern for humankind as a whole" (Bostrom, 2003). In other words, we are obliged to follow the "maxipok rule," which instructs us to "maximize the probability of an 'OK outcome,' where an OK outcome is any outcome that avoids existential catastrophe" (Bostrom, 2002). An existential catastrophe is then (tacitly) defined as any event that prevents us from reaching a stable state of "technological maturity," which denotes "the attainment of capabilities affording a level of economic productivity and control over nature that is close to the maximum that could feasibly be achieved." There are at least four ways that this could happen, according to Bostrom: humanity could go extinct; civilization could permanently stagnate or collapse; civilization could reach technological maturity but in a flawed manner; and civilization could reach technological maturity but subsequently deteriorate (Bostrom, 2013; Torres, 2018a).

The second moral obligation is to colonize our Hubble volume as soon as possible. The reason is that every century of delayed colonization results in 10^{38} lives lost (as implied above), which equals approximately 10^{29} lives forever gone *every second*. Even if mind-uploading is impossible, our local supercluster could house 10^{23} biological humans, which "corresponds to a loss of potential equal to about 10^{14} potential human lives per second of delayed colonization" (Bostrom, 2003). While these figures could be off by orders of magnitude, accuracy is largely immaterial to the argument. As Bostrom (2003) writes, "what matters … is not the exact numbers but the fact that they are huge. Even with the most conservative estimate, assuming a biological implementation of all persons, the potential for one hundred trillion potential human beings is lost for every second of postponement of colonization of our supercluster."

Yet these two obligations are, or could be, in tension, since colonizing space requires the development of advanced technologies and the development of advanced technologies could increase the probability of an existential disaster, which current estimates suggest has a 19 to 50 percent chance of happening this century (see Torres, forthcoming). Thus, the astronomical waste argument also implores humanity to *prioritize* these desiderata: although the opportunity cost of delaying space colonization is staggering in terms of potential value forever lost, far worse than colonizing space later rather than sooner is failing to colonize it at all, which is why existential risk reduction must be "priority number one, two, three and four" (Bostrom, 2003).[3]

In this paper, I will argue that space colonization would likely have catastrophically negative outcomes—specifically, it could produce "astronomical amounts" of *suffering*, or what some theorists have dubbed an "s-risk," for "suffering risk," on the model of "x-risk," for "existential risk" (see Tomasik, 2017a).[4] It follows that, insofar as the maxipok rule mandates colonization, we should not abide by this heuristic. Indeed, if the following arguments are sound, then one might even view the maxipok rule as dangerous. Although I won't explore the issue further in the present paper, perhaps humanity should instead adopt a rule of thumb like the *maximin (or leximin) principle*, which asserts that one should choose the action with the best worst-case outcome, or the *minipnok rule*, which states that, paralleling Bostrom's definition above, whenever we act out of an impersonal concern for humankind as a whole we should try to minimize the probability of a "not-OK outcome," where a not-OK outcome is any outcome that fails to avoid a suffering catastrophe.[5]

The following sections explore, in order, technobiological evolution in space, causes of inter-civilizational conflict, mechanisms for enforcing peace, the advanced weaponry that could be available to future civilizations, and finally, what I argue will likely be the default outcome of colonization, namely, the condition of fear and violence that Thomas Hobbes described as "warre."

2. Descent with modification

As humanity expands into space and our population grows, the human lineage will undergo a process of radical *species diversification*. The result will be a vast multiplicity of distinct "species." There are two factors that will drive this process. First, we live in a *Darwinian world* whereby the mechanism of natural selection is constantly tweaking the genomes of organisms to ensure a satisfactorily good "fit" between the "features" of organisms and the "factors" of their environments.[6] The fact is that no planetary milieu, however terraformed, will be identical to that found on Earth, nor will any artificial environment constructed inside spacecraft like O'Neil cylinders. Such environments may be associated with different gravitational properties, atmospheric pressures and chemical compositions, seasonal variations, circadian and tidal patterns, flora and fauna, solar luminosities, and so on. These factors will amalgamate into selective environments that could influence differential reproduction rates, thereby modifying the frequencies of different alleles within spacefaring populations. Interworld transportation may initially result in some degree of gene

[3] Put differently: To realize astronomical amounts of value, we need to reach a stable state of technological maturity. To reach a stable state of technological maturity, we need to gain control over nature and become economically productive close to the maximum possible. To gain control over nature and become economically productive close to the maximum possible, we need to (a) avoid existential catastrophes, which, by definition, prevent us from reaching a stable state of technological maturity, and (b) colonize as much of our Hubble volume as soon as possible.

[4] This paper is greatly indebted to the groundbreaking work of Deudney (in press), which provided the philosophical inspiration and groundwork for the present article.

[5] Bostrom (2013) argues that the maximin principle entails that we should party like there's no tomorrow because there exists no option that can completely eliminate the possibility of imminent human extinction, the worst-case outcome for our species. But as Sotala and Gloor (2017) correctly point out, there are indeed possible outcomes that could be much worse than extinction (e.g., s-risk outcomes), just as there are conditions of life that could be worse than death (e.g., being tortured, suffering from an aggressive type of brain cancer).

[6] See Laland and O'Brien (2010).

transfer between civilizations, but as future persons become increasingly spread out, parapatric speciation may yield to allopatric speciation. Although one might surmise that modern civilization (on Earth) has largely neutralized the effects of natural selection, this is probably false. Some geneticists even believe that human evolution by natural selection has *accelerated* in recent history, with examples being lactose tolerance and perhaps the exceptional intelligence of Ashkenazi Jews (Gibbons, 2007; Pinker, 2011b). Another evolutionary mechanism that could bring about biological evolution is genetic drift, whereby gene frequencies fluctuate randomly. This could be exacerbated by founder effects resulting from single or a small fleet of spacecraft transporting a relatively small number of individuals who ultimately yield a large population; in this case, the spacecraft would induce a "population bottleneck."[7]

Second, we live in a *Kurzweilian world* whereby the trajectory of our evolutionary development is increasingly within our own control.[8] In other words, in addition to the unintelligent design of natural mechanisms, we now have the option of intelligently designing our phenotypes to optimize our fit to increasingly artificial environments and realize organismal qualities that we value for positional or intrinsic reasons. The result is a process of cyborgization, or the fusing together of technology and biology, artifact and organism, resulting in posthuman properties like enhanced cognition and morality, indefinite lifespans, expanded emotional ranges, and so on. According to Clark (2003), humans are "natural born cyborgs" who have always used technology to substitute, modify, and enhance our phenotypes. In fact, the archeological record of rudimentary tools—the Oldowan toolkit—roughly coincides with the emergence of *Homo habilis*, or "man the maker." But the pace of cyborgization has undergone a rapid increase in recent decades, perhaps following Kurzweil's Law of Accelerating Returns (Kurzweil, 2005). The contemporary human is not merely fused with artifacts like shoes, clothes, and glasses, but computers, smartphones, automobiles, pacemakers, and neuroprosthesis, to name a few. At the extreme, technology could completely replace biology, an end achievable through whole-brain emulation or the creation of AGI software via direct programming, artificial evolution, or recursive self-improvement (see Bostrom, 2014). These wholly artificial beings could then reside in either the "real" world as embodied androids or simulated worlds like those described by Hanson in *The Age of Em* (2016).[9]

There are a few important consequences of Darwinian and Kurzweilian evolution that relate to the possibility of future species acquiring distinctive and unique cognitive-emotional architectures. First, consider the causal role that emotions play in driving behavior. Happiness, sadness, fear, anger, surprise, and disgust can all motivate us to act in various ways. It follows that posthuman species that develop qualitatively different emotional repertoires could be motivated to act in novel ways, some of which could utterly baffle us.[10] Weak analogues can be found among humans: for example, an atheist might find the religious fanatic's decision to blow himself up in a crowded market deeply perplexing. Yet all humans share the same fundamental neural structures and therefore basic emotional range. Thus, the difference between our emotionality and that of a posthuman species could be more analogous to the differences between, say, mice and chimpanzees, or chimpanzees and humans. A related issue concerns the "orthogonality thesis," which states that any set of final goals can in principle be combined with any level of intelligence (Bostrom, 2014). It follows that a posthuman species could become *superintelligent* and still be driven by goals that appear *irrational* to us (that is, in the *value* rather than *instrumental* sense of rationality; see Torres (2017a)).

As for cognition, a being can understand the world only insofar as it can represent the world, and it can represent the world only insofar as it can generate concepts that correspond to features of the world. This is important because a being can intentionally *manipulate* the world only insofar as it *understands* the physical laws and causal mechanisms by which it operates. Following Chomsky (1976), we can call puzzles that are in principle knowable relative to a mind-type M, even if not known, "problems" and puzzles that are in principle unknowable to M "mysteries." The problem-mystery distinction demarcates the *cognitive space* of M—a space within which are problems (to be) solved and outside of which are mysteries that will forever remain unknown. The point is that the cognitive space of each species is unique to that species, the product of contingent evolution (so far on Earth). It follows that future species with different cognitive architectures could have radically different cognitive spaces. These spaces could subsume, overlap, or entirely diverge from our human cognitive space. In each case, such beings could potentially devise theories that enable them to manipulate the world in ways that *we* find permanently inscrutable. To use a favorite example of mine, imagine a "chipmunk scientist" trying to figure out how the voice of someone in Tokyo can emerge from a plastic/metal device in Toronto. Since the chipmunk lacks the mental mechanisms needed to generate concepts like *microwave frequencies* and *geostationary satellites*, there is no amount of research or schooling that could ever enable the chipmunk to screech "Eureka! So that's how this works" (Torres, 2017a).

The point is that different species could have fundamentally different models of mind-independent reality, and this could enable

[7] I strongly suspect, though, that if our descendants haven't gained *total* control over their genomes by the time they venture beyond the solar system, they will sometime afterwords, more or less universally. Thus, the influence of Darwinian mechanisms and genetic drift will very likely shrink the further one peers into the deep future. (Thanks to a referee for emphasizing this point.)

[8] This could also be termed a "Huxleyan world." After all, long before Kurzweil, Julian Huxley wrote that "it is as if man had the biggest business of all, the business of evolution—appointed without being asked if he wanted it, and without proper warning and preparation. What is more, he can't refuse the job" (quoted in Agar, 2010).

[9] Incidentally, the possibility of creating cyborgs that are better suited to the space environment than trying to export our planetary environment into space was discussed in a 1960 article by Manfred Clynes and Nathan Kline called "Cyborgs and Space". For example, they write that "space travel challenges mankind not only technologically but also spiritually, in that it invites man to take an active part in his own biological evolution. Scientific advances of the future may thus be utilized to permit man's existence in environments which differ radically from those provided by nature as we know it," adding that "the task of adapting man's body to any environment he may choose will be made easier by increased knowledge of homeostatic functioning, the cybernetic aspects of which are just beginning to be understood and investigated."

[10] Along these lines, Roden (2015) argues that our "wide" descendants (i.e., posthuman progeny, broadly construed) could have radically different *phenomenological experiences*, where such experiences affect the moral status of future beings.

them to intervene on the universe in ways that are reciprocally unintelligible to each other. The result could be a rather confusing, unprecedented, and potentially catastrophic sort of "mutually asymmetrical warfare" (MAW), whereby each participant has access to weapons whose underlying causal mechanisms are cognitively closed to the others. Thus, one species might observe *something happening*—perhaps something harmful—with no way of figuring out how some species on the other side of the galaxy accomplished the feat, or vice versa. Unlike a technologically "advanced" civilization on Earth fighting a technologically "primitive" society—as was the case when Europeans reached the Americas—space wars between posthuman species with different mind-types would be more like a military confrontation between *Homo sapiens* and bonobos, except that in this terrestrial case the asymmetry is one-sided rather than mutual.[11]

Yet another kind of diversification that will occur as our descendants propagate through space is cultural or memetic. First, consider that—as discussed below—the cosmic speed limit of light will reduce the efficacy of communication between distantly located civilizations (Deudney, in press). This will, in turn, reduce inter-civilizational meme-flow and enable unique traditions of thought to take shape in quasi-isolated regions of spacetime. Given the possibility of radically different cognitive-emotional architectures, the cultural, political, governmental, religious, linguistic, intellectual, philosophical, scientific, technological, *and so on*, traditions that arise could be profoundly different from each other, and from the various traditions that have emerged during our own short history on this planet. In a phrase, colonizing space will have the exact opposite effect that globalization has had on Earth. Whereas the latter has homogenized the world in innumerable ways and, indeed, will ultimately yield a single race of brown-skinned humans, space colonization will generate unprecedented phylogenetic and ideological diversity. The global village will fracture into an astronomical number of distinct cosmic settlements.

3. Why would space be peaceful?

Given this snapshot of the far future, the question arises: why would space be peaceful? Here we will propose a theory of cosmic conflict, or an explanatory account of why civilizations might be driven to engage each other in violent confrontations. To begin, consider "an analysis of the incentives for violence that is as good as any today," as Pinker (2011a) describes it, namely, that proposed by Hobbes. This analysis identifies causes of conflict within a condition of *anarchy*, i.e., a situation in which there is no ruling power—and hence no *hierarchy*—that can impose law and order on individuals, a fact that will become relevant in Section 4. Hobbes (1982) writes:

> So that in the nature of man, we find three principal causes of quarrel. First, competition; secondly, diffidence; thirdly, glory. The first maketh men invade for gain; the second, for safety; and the third, for reputation. The first use violence, to make themselves masters of other men's persons, wives, children, and cattle; the second, to defend them; the third, for trifles, as a word, a smile, a different opinion, and any other sign of undervalue, either direct in their persons or by reflection in their kindred, their friends, their nation, their profession, or their name.

Let's call actors motivated by gain—or, more usefully, "malign intentions"—*Machiavellian actors* and those motivated by diffidence or fear *Tuckerian actors*.[12] (We will discuss actors motivated by credibility—Hobbes's third cause—in Section 5)

But Hobbes's analysis is incomplete; there are actors whose motivations do not clearly fall within his tripartite framework. Consider religious zealots who pursue an action for the sole reason that they believe God has commanded them to do so. In fact, history is replete with dangerous apocalyptic movements that perpetrated violent acts in accordance with eschatological narratives in which they saw themselves as active participants. The doomsday cult Aum Shinrikyo, for example, released sarin in the Tokyo subway system in an explicit effort to initiate World War III, or Armageddon, and the Islamic State attempted to lure coalition forces to northern Syria—specifically, to a small town called Dabiq—because this is where they believed that Armageddon would occur. Similarly, Christian dispensationalists in the US—sometimes dubbed the "Armageddon lobby" (Haija, 2006; Torres, 2017a)—have shaped US foreign policy in the Middle East based on their particular premillennial beliefs about how the eschaton will unfold. None of these groups were motivated by fear (in Hobbes's sense), nor were they seeking personal gain, at least not as a final goal. Although intelligence and religiosity are negatively correlated among humans on Earth, the orthogonality thesis reminds us that it is possible for some future posthuman species to acquire exceptionally high levels of instrumental rationality while nonetheless adhering to some religious and/or eschatological theses that would appear—to *us*—to be epistemically and/or morally absurd (see Torres, 2017b, 2018d; Zuckerman, Silberman, and Hall, 2013).

Borrowing from the "agential risk" framework that I have elsewhere outlined, we can further identify actors who are driven by "pro-mortalist" ethical systems, extreme environmentalist ideologies, or "idiosyncratic" beliefs/desires about how the world is and ought to be (Torres, 2017a). For example, "radical negative utilitarians" (RNUs) believe that one's behavior is morally good *only insofar* as it reduces the suffering of sentient beings. Since the ultimate way to reduce suffering is to eliminate that which can suffer—preferably through some painless, instantaneous process, which may become possible, as we will see below—it follows that the annihilation of all life in the universe constitutes the best possible outcome. The point is that radical NUs exist today on Earth, and there's no reason to believe that this moral stance couldn't spread to or spontaneously emerge within other civilizations, thereby

[11] Yet another potentially catastrophic possibility is that posthuman AIs lack consciousness, and that these AIs become dominant in the universe. See Schneider (2016).

[12] Named after Albert Tucker, the mathematician who coined the term "prisoner's dilemma." The prisoner's dilemma will enter the picture later on when we discuss the Hobbesian trap. Note also that these categories are *not* mutually exclusive: one can have malign intentions and also be fearful of others, thus attacking them for primarily self-preservational reasons.

posing an existential threat to everyone. One can similarly imagine some form of "cosmic environmentalism" that sees the colonization of space by descendants of *Homo sapiens* as destroying the "natural beauty" of the universe—especially if destructive conflicts become the norm. There could arise sentiments according to which our progeny (their contemporaries) are "Posthumanpox" that have spread like a virus, destroying everything that it comes into contact with.[13] As a virus, this Posthumanpox must be eliminated *in toto*, just as some radical environmentalist groups have advocated the complete extermination of human beings on Earth through the use of advanced technologies. Given the potentially vast number of future civilizations, it stands to reason that *at least some* will develop contrarian views that cast our descendants' collective exploitation of negentropy in a negative light.

And finally, there could be civilizations led by individuals who harbor a death wish for posthumanity, as it were, due to some pathological quirk in their psychological make-up. This is also not implausible given that, as I have elsewhere documented (Torres, 2018), there have been numerous people—often of high intelligence—throughout history who have both (a) engaged in horrific violence, and (b) expressed omnicidal fantasies in either public or private. There is, indeed, strong evidence that if such individuals were to gain access to a "doomsday machine" of some sort they would have sadistically, suicidally, and gleefully used it to annihilate their conspecifics. As Sagan (1994) notes, referring to the menacing possibility of redirecting asteroids toward Earth, there really are madmen in the world:

> We are sometimes told that this or that invention would of course not be misused. No sane person would be so reckless. This is the "only a madman" argument. Whenever I hear it (and it's often trotted out in such debates), I remind myself that madmen really exist. Sometimes they achieve the highest levels of political power in modern industrial nations. This is the century of Hitler and Stalin, tyrants who posed the gravest dangers not just to the rest of the human family, but to their own people as well. In the winter and spring of 1945, Hitler ordered Germany to be destroyed—even "what the people need for elementary survival"—because the surviving Germans had "betrayed" him, and at any rate were "inferior" to those who had already died. If Hitler had had nuclear weapons, the threat of a counterstrike by Allied nuclear weapons, had there been any, is unlikely to have dissuaded him. It might have encouraged him. Can we humans be trusted with civilization-threatening technologies?

The very same question can and must be asked about our posthuman descendants—indeed, it may be all the more urgent given the cognitive-emotional diversification of lifeforms during the deep space diaspora. The picture that emerges from such considerations is one in which there will exist at least some, and potentially many, civilizations that are inclined toward violence. Some will engage in violence for imperialistic reasons—for gain—while the impetus for others will be religious, apocalyptic, pro-mortalist, anti-posthumanist, environmentalist, or "psychopathological" in nature. The existence of Machiavellian actors will, in turn, give others a strong incentive to engage in *preventive* or *preemptive* strikes against potential predators. To quote Levy and Thompson in *Causes of War* (2010), "a preventive war is motivated by the perception of a rising adversary, a shift in power, and by the fear that once the adversary is stronger it will attempt to exploit its advantage through coercion or war … and is driven by 'better-now-than-later' logic." In contrast, "preemption involves a military attack in response to the virtual certainty that the adversary is about to strike and by the motivation of gaining the advantages of striking first."

Even more, the motivation to strike first need not involve a Machiavellian actor at all; it could involve two or more Tuckerian actors with no malicious inclinations whatsoever. The crucial idea here is what international relations scholars refer to as the *security dilemma*, whereby, in sum: anarchy generates uncertainty about the present and future intentions of other actors; this leads to fear, resulting in the accumulation of weapons arsenals, etc. for "defensive" purposes; this increases the fear of other actors uncertain of one's true intentions, thereby producing a spiral effect, or vicious positive feedback cycle, that can foment conflict, as other actors increase their own arsenals for "defensive" purposes as well (see Tang, 2009). In other words, two peaceable civilizations could end up warring due merely to a spiral of escalating militarization given a lack of mutual trust. A related concept is *Schelling's dilemma*, also known as the "Hobbesian trap," whereby one actor engages in a first strike against a second actor due to a fear of being imminently attacked by the first actor. Again, neither might harbor malign goals (although one could), yet they engage in war for purely game theoretic reasons. The classic illustration of this involves a robber with a gun who breaks into a house intending only to steal jewelry; the owner wakes up and confronts the robber with a gun. Neither wishes to shoot the other, yet each fear that they will be shot if they don't shoot first. The result is tragedy.

There is another version of this situation that doesn't pertain to each actor's intentions with respect to others. Rather, it arises from a combination of (a) fallibility, and (b) technological capability. For example, civilization A might decide, after sufficient deliberation, that civilization B poses no malign threat; yet it might also worry that B is not responsible enough to possess its technological power. Perhaps B is conducting high-powered physics experiments that could produce a dangerous black hole or some other catastrophic phenomenon that would affect A. If efforts by A to convince B not to run such experiments fail, it could be in A's preservational self-interest to invade, conquer, and/or destroy B. Thinking about this situation in the context of a galaxy of potentially *billions* of civilizations, it could be in any given civilization's best interest to annihilate all other civilizations in the universe, just in case they were to cause a galactic- or cosmic-scale disaster *by accident*. Put differently, error as well as terror could fuel inter-civilizational conflicts.

Even more, the security dilemma/Hobbesian trap predicaments could be exacerbated by potential difficulties in interspecies communication, which would further vitiate the trust needed for civilizations not to attack each other. First, the Quinian "indeterminacy of translation" suggests that contact between civilizations could fail to convey the intended meaning, possibly leading to trouble (see Jebari & Olsson-Yaouzis, in press). Second, if two species come to have different cognitive spaces or emotional

[13] I coin the term "Posthumanpox" on the model of Foreman's (1991) term "Humanpox."

repertoires, this could make understanding the other fundamentally impossible, thereby fueling suspicions about the beliefs, desires, and capacity for deception of the proverbial "Other." Indeed, the lack of common "ontological ground" could also lead to breakdowns of empathy: trying to understand how an action X makes another species "feel" would be like a human trying to understand "what it's like to be a bat." More dangerously, it might not even be clear to species A that species B can have *conscious experiences* of pain in the first place. "So," A might reason, "why would it be unethical to harm species B?" Species in such situations are not merely *aliens to* each other but, more significantly, *alienated from* each other.

Yet another issue worth mentioning is that future space weapons could not only enable civilizations to obliterate each other, but phenomena like mind-uploading and life-extension could enable captors to inflict "eternal punishment"—that is, until the entropic death of the universe[14]—on those captured, thus greatly increasing the stakes of conflict. For example, civilization A might not only worry about the aggressive, expansionist proclivities of civilization B, but fear that if it were to resist B's demands and subsequently succumb to its military advances, the surviving individuals of A would be cast into an artificial perdition of interminable suffering. This could give A an even greater incentive to launch a first strike against B—to eliminate the dual threats of dying in war and living in hell.[15]

To summarize so far: expansion into space will generate phylogenetic and ideological diversity that could yield profoundly disparate types of civilizations. The species who comprise these civilizations could have entirely different normative preferences, moral tendencies, and even scientific institutions. Some will almost certainly be violence-inclined, thus giving others an incentive to strike first. Even more, diversity with respect to cognition, emotionality, and language will undercut the mutual trust needed for otherwise irenic civilizations to avoid spirals of militarization or defect in prisoner's dilemma predicaments. Thus, a colonized cosmos would be an arena poised and spring-loaded for violence. But is there a way to prevent conflict from breaking out?

4. How could space be peaceful?

In *Leviathan*, Hobbes argues that when instrumentally rational, self-interested individuals find themselves in anarchic conditions, they will band together through a "social contract" to establish a supreme governing body that has a "monopoly of the legitimate use of physical force within a given territory" (to quote Weber, 1919). In exchange for giving up some personal liberties, this governing body—which Hobbes called "the Leviathan"—will provide the citizens with security. Consequently, war is replaced with law, anarchy with hierarchy. Moving up a level of organization from the state to the international system, one finds an isomorphic situation with respect to Hobbes's "state of nature." Here governments (and their institutions) can be seen as "individuals" in an anarchic realm that consists of all other states. It follows that one way to establish peace among states is to implement a *global Leviathan* that takes the form of a "world government," "singleton," or "supersingleton" (see Bostrom, 2005; Torres, 2018b). Moving yet another level up from the geopolitical to the cosmopolitical realm, the same conclusion follows: to replace war with law, civilizations should band together and establish a *cosmic Leviathan* that provides security at the minor cost of some civilizational freedoms.

Unfortunately, this appears unpromising. Let's begin by reflecting on the inscrutable vastitude of space and how this would affect a cosmic Leviathan's ability to coordinate, regulate, and punish the actions of Machiavellian and Tuckerian actors. While potentially habitable exoplanets cluster around common solar bodies, the distance between solar systems can be immense. The super-Earth Gliese 581d, for example, is approximately 20 light-years from Earth, meaning that an electromagnetic signal sent as of this writing, in 2018, wouldn't reach it until 2038. A spaceship traveling at one-quarter the cosmic speed limit—perhaps employing some form of nuclear pulse propulsion—wouldn't arrive until 2098, and a message to simply affirm that it had arrived safely wouldn't return to Earth until 2118. And Gliese 581 is relatively close as far as exoplanets go: the Andromeda Galaxy is some 2.5 million light-years away and the Triangulum Galaxy about 3 million light-years. Even more, there are some 54 galaxies in our Local Group, which is about 10 million light-years wide, within a universe that stretches some 93 billion light-years across; and recall that the universe is metrically expanding at an accelerating rate. (See Fig. 1.)

The point is that the laws of physics, as we know them, impose significant constraints on the travel of spacecraft and information-carrying beams, which would make a cosmic Leviathan extremely dissimilar to the Leviathans under which we live on Earth.[16] Timeliness is necessary for states to satisfy their half of the social contract, and the hard limits to travel and communication would render attempts to provide civilizational security across galaxy clusters, superclusters, and so on, *untimely*. Imagine the futility of the state if one has to wait two weeks for an emergency 911 call to reach the operator or for the police to show up at the scene of a bank robbery. The social contract would fall apart, and for this very reason it is unlikely to ever be "signed" by a large number of spacefaring civilizations to begin with. Another problem with the cosmic Leviathan proposal is that it would require the approval of its member civilizations for its legitimacy, and approval would require the government to adequately represent the interests, beliefs, desires, and so on, of not only trillions and trillions and trillions of different *individuals*, but upwards of billions and billions and billions of different *species*. It is difficult to imagine how a single, centralized entity could do this, especially if some of the interests, etc. of member species are in tension or outright contradictory, as will no doubt be the case.

The discussion so far has largely assumed a neorealist framework—the most dominant theory among international relations scholars—with respect to the cosmopolitical realm. Yet some scholars in the liberal tradition emphasize other possibilities for peace.

[14] Or, speaking very speculatively, beyond, if future beings discover a way to migrate into another universe.

[15] A related possibility involves one civilization coercing another by threatening to simulate suffering beings. If the latter civilization adheres to an ethics according to which suffering is morally bad, then this could motivate it to fulfill the former civilization's wishes.

[16] See also Tomasik (2017b), which discusses similar issues in another context.

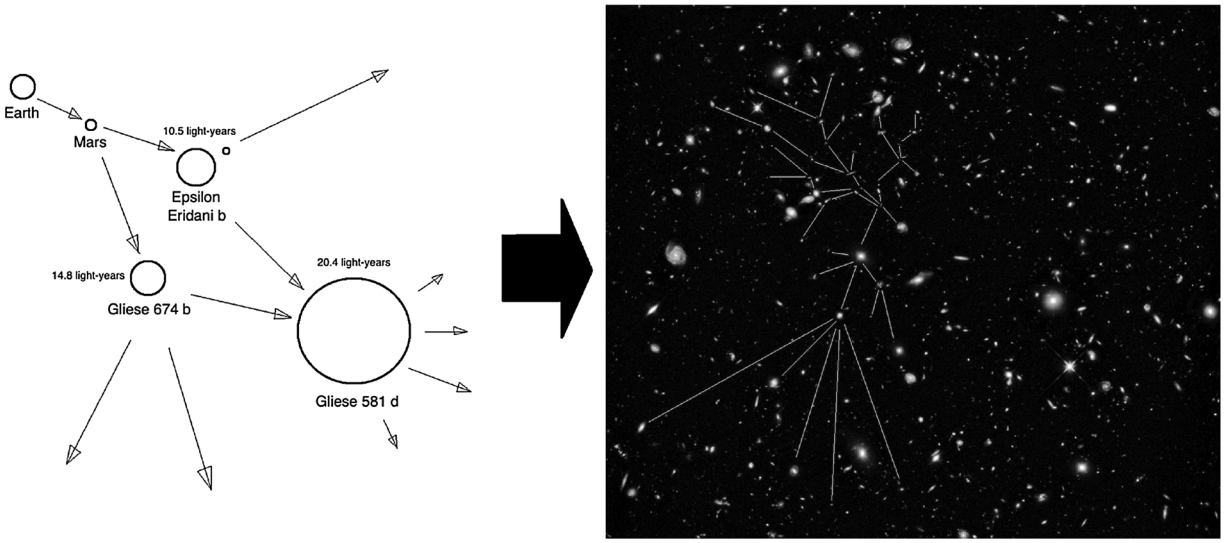

Fig 1. The immense vastness of space will make a cosmic Leviathan infeasible.

For example, there is robust evidence that democracies almost never fight each other, an idea dubbed the "democratic peace theory." The two primary reasons proposed for this finding are, first, that democracies share a special "trust" and adhere to similar norms; this is the *normative* explanation. And second, that democratic leaders are more accountable to their citizens than autocratic states; this is the *institutional* explanation. Note here that democracies aren't necessarily less warlike, only that in dyadic configurations where both actors are democratic, war is improbable. A related idea goes by the name "capitalist peace theory," according to which states with capitalist economies rarely go to war. One reason is trade interdependence: it might not pay to invade another country if that country is supplying the invading force with useful commodities. Along these lines, Friedman (2005) has offered the "Dell Theory of Conflict Prevention," which states that "no two countries that are both part of a major global supply chain, like Dell's, will ever fight a war against each other as long as they are both part of the same global supply chain."

The question is whether such theories offer hope that inter-civilizational war will be rare, and the answer appears to be negative. For one, there is no particular reason to believe that future civilizations will be democratic, although even if *all* of them are, at the extreme, the differences between species—their cognitive spaces; emotional repertoires; cultural, political, religious, etc. traditions; and so on—would surely overwhelm the common value that "electing political leaders through popular vote" is the best form of government. Second, the political fallout of launching a war provides leaders an incentive not to fight only if there exists the possibility of injurious retaliation or engagement; as we will see below, though, some futuristic weapons would very likely preclude both. And third, the (i) distances between solar systems, (ii) ability to mine asteroids, and (iii) development of artifacts like nano-factories would make material trade between civilizations unnecessary. Thus, it appears unlikely that two or more civilizations would become part of a major cosmic supply chain, to echo Friedman's phrasing. This suggests that establishing a "Kantian-type peace" within the cosmopolitical realm will be unworkable. The liberal tradition offers no more hope than a cosmic Leviathan.

5. Space-Age weaponry and the balance of terror

Yet there is another strategy for neutralizing the Hobbesian trap, namely, a *policy of deterrence*, also known as a "balance of terror" or, during the Cold War, "mutually-assured destruction" (MAD). This asserts that "if you strike me, I will most assuredly strike back with equal or greater force, and if I strike you it will *only* be because you struck me first."[17] Deterrence is only effective when one's adversaries genuinely believe the statement, "I will most assuredly strike back." This returns us to Hobbes's third cause of conflict from Section 3: glory, honor, or *credibility*. To establish credibility and, therefore, dissuade potential attackers, one has reason to engage in confrontations with others and, in doing so, to demonstrate one's capacity for violence. The question is whether policies of deterrence implemented by civilizations throughout the cosmos would be sufficient to obviate war. To answer this question, let's begin by considering the unsettling range of weapons that will likely be available to our spacefaring progeny; we will then explore how these weapons could enhance or mitigate the effectiveness of deterrence.

5.1. Weapons of total destruction (WTDs)

There are a variety of "kill mechanisms" that one civilization could use to obliterate another. In relatively close propinquity,

[17] Or, alternatively, defensive actors could threaten to employ a "scorched Earth" tactic against Machiavellian actors out for material gain. That is, the former could issue credible threats that if it or its territory is attacked/plundered, it will destroy all the natural resources in the region, thus making an attack otiose.

chemical and biological weapons could offer a means of targeted violence, since the deleterious effects of such weapons might be limited to a particular species (Deudney, in press). For example, the toxicity of a chemical X might be low for a species A but lethal to a species B. This could enable A to use X on B without fear of X harming A—a concern that has dissuaded some terrorists from employing chemical weapons. The same goes for a pathogenic germ Y: since pathogens often only harm single species, biological weapons could be used without the perpetrators worrying about becoming sick. With respect to artificial intelligences, there could be viral malware that affects only certain types of software; in this case, such viruses could be transferred not at the velocity of a sneeze but at the speed of light, traversing astronomically large stretches of space to devastate colonies of artificial-substrate beings.

Another possibility involves weaponizing "minor planets" like asteroids. This hints at the *deflection dilemma* discussed by Sagan (1994), among others, whereby the very same technology that could deflect an asteroid away from Earth could also be used to redirect one toward it. The resultant "planetoid bombs" could be launched in the direction of target civilizations at extremely high velocities and inflict far greater destruction than all the nuclear arsenals on Earth combined (see Cole & Cox, 1965; Deudney, in press). Even more, asteroids are extremely numerous in the solar system and have a wide range of sizes, with estimates of 1.1 to 1.9 million that have greater-than-1-kilometer diameters in the asteroid belt between Mars and Jupiter. (A 1-kilometer impactor striking Earth would likely annihilate humanity by causing an impact winter.) Thus, asteroids constitute an abundant source of easily obtainable, civilization-ending weaponry—a particularly worrisome fact given that the technological capabilities to redirect asteroids will likely emerge at an early stage in our diaspora "out of Earth," as it were (see Deudney, in press).

Other futuristic space weapons include military drones that either initiate attacks or engage in clandestine surveillance of other civilizations. Such drones could hide themselves from counter-surveillance detectors by employing metamaterial invisibility cloaks and propagate themselves through the von Neumann process of self-replication, that is, by converting raw materials into clones of themselves. There is also the possibility of using "heliobeams," or "sun guns," to destroy targets by concentrating large amounts of solar radiation via a concave mirror on a satellite. Even more catastrophic are direct-energy weapons (DEWs) like lasers and particle-beams that use highly focused energy to superheat their targets. In fact, the US government has already developed weapons of this sort—they are science fact rather than fiction—although future breakthroughs could enable them to become immensely more destructive. If this is the case, they will offer yet another mechanism for wreaking unprecedented harm (see Deudney, in press). Along these lines, Sandberg (in press) and Sandberg, Armstrong, and Cirkovic (2016) suggests that technologically advanced civilizations could potentially use gravitational waves to create black holes. Generating waves of sufficient intensity would be energetically inefficient, according to current physics, but they have the advantage that they can interact with dark matter objects, unlike electromagnetic-energy weapons.

Even more, the universe appears to be in a "metastable" energy state. This suggests that one could tip it into a more stable, lower-energy state, perhaps by concentrating huge quantities of energy in tiny regions of spacetime, as occurs in some high-powered physics experiments. In other words, a particle collider could be weaponized to intentionally nucleate a "vacuum bubble," or sphere of "true vacuum" spreading in all directions at the speed of light and destroying everything with which it comes into contact. Who might weaponize a particle collider? First, there could be actors who use the threat of a vacuum bubble for blackmail purposes. Second, there could be madmen (like Hitler) who create a vacuum bubble to avoid defeat. That is to say, a predatory actor could hold the following preference ordering: (i) triumphant victory over, say, its Local Group, (ii) total annihilation of the universe, and (iii) defeat. Third, particle colliders would also be the ideal WTD for RNUs, since it would enable them to obliterate not only all extant life in the universe but the very *potential* for life to arise—and it would do this without inflicting any suffering whatsoever.[18] Another possibility is that Tuckerian actors create a vacuum bubble for the purely defensive reason of eliminating all potential attackers in the universe. As Sandberg (2017) speculates, it might be possible for "certain configurations of matter, energy, black holes, etc. [to] induce a post-transition structure that can act as an assembler." This "assembler" would enable "some information [to] be transmitted into the new state," thus making it possible for a civilization to "survive," in some sense, the universe settling into a lower-energy configuration. On the other side of this transition, the "structure" can recrudesce into a new daughter civilization with the certitude that it is completely alone and, therefore, safe.

Finally, it is crucial to note that future beings—some of whom may have hugely augmented cognitive capacities—will almost certainly invent new weapons that are more powerful and effective than anything we could imagine. Such weapons could enable civilizations—or perhaps lone wolves, of which there could be, once again, trillions and trillions and trillions—to cause unprecedented injury to other civilizations. Consider the following passage from Bostrom (2013):

> One can readily imagine a class of existential-catastrophe scenarios in which some technology is discovered that puts immense destructive power into the hands of a large number of individuals. If there is no effective defense against this destructive power, and no way to prevent individuals from having access to it, then civilization cannot last, since in a sufficiently large population there are bound to be some individuals who will use any destructive power available to them.
>
> Scale this up from the individual level to the cosmopolitical level and the same conclusion follows: Life in the universe cannot last.

[18] Yet another possibility is that Tuckerian actors destroy the entire universe out of fear. Sandberg (2017) refers to this as the "game theory of triggering vacuum decay." Here's the catch: it might be possible for "certain configurations of matter, energy, black holes etc. [to] induce a post-transition structure that can act as an assembler," thus enabling "some information [to] be transmitted into the new state." This could make surviving a vacuum day in some sense possible (although probably with large costs), and therefore it could make inducing a vacuum decay advantageous if one is seriously worried about malicious actors preemptively attacking.

5.2. Policies of deterrence

With this brief sketch of space weaponry in mind, let's consider the deterrence predicament beginning with the colonization of Mars and expanding outward from there. As colonies on the fourth rock from the sun become increasingly Earth-independent, they will begin to develop their own culture, political systems, religious traditions, and perhaps even technologies. The Darwinian and Kurzweilian mechanisms will also engender new forms of martian posthumans that nontrivially differ from Earth-bound (post) humans. If "morphological freedom" is granted to martian citizens, then there could emerge a general phylogenetic trajectory of the entire population in addition to more specific ontogenetic trajectories resulting from individual phenotype modifications. (The same goes for populations on Earth.) As Deudney (in press) notes, geopolitical theory predicts that groups exhibiting greater differences are more likely to engage in conflict, and since differences are likely to evolve between the populations of Earth and Mars, one should expect tensions to rise. There could, for example, be competition for astronomical resources (such as asteroids and comets), leading to disagreements about inter-planetary policies and practices. Domestic affairs that one side sees as worrisome—e.g., the election of a demagogic strongman with xenophobic tendencies, or the collapse of some global regulatory organization—could also lead each to question the trustworthiness of the other, thus planting the seeds for a security dilemma whereby each militarizes space, for "defensive" reasons, in response to the other militarizing space, and so on.

One might surmise here that a balance of terror could establish bipolar stability, just as MAD did during the Cold War. Yet this appears implausible given the weapons mentioned above. For example, if one side could release self-replicating nanobots that cripple the target civilization before it can retaliate, the result would be a terror *im*balance that, under certain circumstances, would make a first strike game theoretically rational. In fact, Kurzweil outlines a scenario in which ecophages destroy the entire biosphere of Earth within ~90 min. This would involve a two-stage attack: first, a small population of nanobots would spread around the globe, and second, at an "optimal" time this population would begin to self-replicate at an exponential pace. To put this in perspective, signal delays between Earth and Mars range from 4 to 24 min, depending on where each planet is in its orbit, and travel times range from 150 to 300 days. Add to this the inevitable lag of bureaucracies and the outcome is a serious credibility-of-deterrence problem. Even more, some future genius could invent a far more effective way of weaponizing nanobots in the next 100 years, at which point humanity will probably have established martian colonies.[19] Related scenarios involving designer pathogens that initiate "engineered global pandemics" or planetoid bombs capable of obliterating whole metropolises—or perhaps an entire ecumenopolis, if one exists—could also be imagined, although I will leave this task for the reader.

But the situation is far worse than this, because ecophages, pathogens, and asteroids won't pose the greatest risks to inter-planetary peace: heliobeams, DEWs, and gravity waves not only could inflict catastrophic damage on their targets but they could do this at or near lightspeed. In a flash, one civilization could cripple the other's key military and/or civilian infrastructure, thus rendering it unable to effectively respond to an attack. Furthermore, since the speed of light imposes an upper bound on information transfer, there could be, in principle, no early-warning systems to alert the target civilization that an attack has commenced, which would severely compromise its ability to initiate defensive measures. One might here wonder: perhaps the attackee could overcome this defensive vulnerability by stationing counterstrike military drones throughout the solar system. They could be programmed to launch a coordinated attack if they fail to receive a "no-strike" signal that is ordinarily sent to them every few minutes. Thus, the destruction of key military infrastructure would result in the cessation of this signal and therefore the initiation of a counterstrike. But this too appears otiose since a first strike using, say, DEWs could simply target these drones as well. The result is that threats of retaliation from each civilization would be literally *in*-credible and the balance of terror would collapse.

Here we should also not overlook the potential for *accidents* to cause conflicts when inter-civilizational tensions are sufficiently high. The disturbing historical fact is that "pure dumb luck" played a critical role in preventing nuclear war from occurring during the Cold War. Individuals like Vasili Arkhipov and Stanislav Petrov more or less single-handedly averted nuclear holocausts, and an interpretation error in 1995 led Boris Yeltsin to become "the first Russian president to ever have the 'nuclear suitcase' open in front of him" (Cirincione, 2013). Although intelligence is negatively correlated with accident proneness, and presumably our (post)human descendants will be cognitively enhanced to some extent, even a small probability of error could make disaster almost certain (see Torres, forthcoming). For example, imagine that a mere 500 people have access to a "button" that, if pushed, would initiate a catastrophic first strike against the other civilization. If each of these individuals has a mere 0.01 chance per decade of accidentally pushing this button, the result is a staggering 99.3 percent probability that, within 10 years, the strike will occur. So, perhaps Earth and Mars—whose civilizations could potentially coexist for another 10 million centuries, until the sun burns out—won't be quite as lucky as the US and Soviet Union were for the slightly more than four decades between 1947 and 1991.

The final step in the present argument is to project this bi-planetary predicament into the vast reaches of outer space. Consider the *billions and billions and billions* of populations that could come to occupy a universe with 10 trillion galaxies and 10^{24} stars, each with its own traditions, boasting of weapons that could destroy entire galaxies or even the entire universe, and embedded in a cosmo-political system of lawless anarchy. There is no supreme governing system to provide security and no policies of deterrence to reliably prevent first strikes. It is hard to imagine how such a predicament could avoid constant and catastrophic wars between civilizations both near and far. Indeed, theorists like Waltz (1979) have argued that multipolar state configurations are less stable and more prone to conflict than bipolar configurations. The reason is that uncertainty increases with the number of actors, and as uncertainty increases, so does distrust of everyone else's intentions. Hence, the more civilizations there are in the universe, the greater the incentive for Tuckerian actors to preventively or preemptively strike their neighbors—or to induce a vacuum bubble in the hope that

[19] Note that Kurzweil's scenario is somewhat contentious among experts. At the very least, it represents an extreme case.

an "assembler" on the "other side" can enable some form of post-transition survival. The point is that the future will be marked by *radical multipolarity*, and this will greatly increase the probability of violence. Yet the difficulty of establishing Earth-independent colonies on Mars without catastrophic wars—as outlined above—suggests that our descendants might not make it beyond the solar system. In fact, Deudney (in press) argues that attempts to colonize space could constitute the Great Filter that explains why we see no evidence of intelligent aliens crying out for cosmic companionship in a universe slowly sinking into thermodynamic equilibrium.[20]

6. Additional considerations

Before concluding, let's consider three additional issues that are relevant to the present thesis. First, this paper focuses on one of a few foreseeable space-colonization scenarios. Another possibility, which is endorsed by some scholars at the Future of Humanity Institute (FHI), is that humanity creates a *singleton* controlled by a friendly superintelligence before we propagate beyond the solar system.[21] I see two problems with this: (i) we will almost certainly establish martian colonies before we leave the solar system and, as subsection 5.2 notes, tensions will likely emerge between martian and earthian civilizations as they become increasingly independent; this could make a joint martian-earthian singleton difficult to establish.[22] And (ii) it is unclear why a superintelligence that facilitates the posthuman colonization of space wouldn't encounter the same insurmountable challenges that led us to dismiss, in Section 4, the feasibility of a "cosmic Leviathan." How exactly could a superintelligence enforce law and order when physical limitations like the speed of light severely problematize the coordination of far-away entities? One might try to circumvent this issue by arguing that a singleton could take the form of some immutable software that governs the behavior of all future beings and must be embedded within every technological civilization, spacecraft, and so on. This would overcome the communication challenge associated with the spatial vastness of the universe, since no communication between instances of the program would be necessary. Yet this too seems problematic. Consider that if humanity spreads beyond the solar system in 100 years, then we will need to have this software in its final form within a century. Doing this would require solving the philosophical problem of determining which values should guide *all future beings for the rest of time* (since the software is immutable, a necessary condition to overcome the communication challenge), as well as the technical problem of ensuring with virtual certainty that this software will remain regulatorily efficacious even after millions of years of unimaginable future development (i.e., it can't be the case that future breakthroughs enable hackers to disable the software). Perhaps there could be periodic software updates, but this brings us back to the formidable question of what central decision-making body would decide which updates to make, how this body could represent the interests of so many diverse species, and so on. In my view, this proposal does not offer a promising solution to the security problems outlined above.

Second, a shortcoming of the present analysis that we should flag concerns the immense *uncertainty* about what future technology might look like. There are, obviously, epistemic limitations to anticipating the kinds of defensive technologies that could be invented centuries, millennia, or billions of years from today. Perhaps some distant civilizations develop highly effective and "invincible" counterstrike weapons that, as such, can deter all preemptive attacks. Just as Bostrom (2013) asks us to imagine a ball that, once removed from the urn of innovation, cannot be placed back into that urn and that brings with it almost certain doom, so too could there be "eucatastrophic" inventions that greatly increase inter-civilizational and inter-species peace (see Cotton-Barratt & Ord, 2015). Still, I would push back against this point. On the one hand, the historical fact (on Earth) is that offensive technologies have typically antedated defensive technologies, thus resulting in periods of excessive vulnerability. This chronology of offensive followed by defensive technologies is likely to occur in space as well, especially given that distant civilizations could develop completely novel forms of weaponry in secret, thereby making it impossible for target civilizations to develop effective defenses in time. The resulting window of vulnerability, however short, would pose an unacceptable threat if the relevant offensive technologies have the capacity to *annihilate* their targets. An existential catastrophe can, by definition, only occur once in a species' lifetime.[23]

Finally, we should note that the present paper compliments the conclusions of several other scholars who have approached the topic from different angles. Perhaps most notably, Brian Tomasik worries that terraforming Earth-like planets or spreading life via "directed panspermia" (as Claudius Gros, who founded the Genesis Project, advocates) could significantly increase the total amount of suffering in the universe—an especially bad outcome if one espouses a "suffering-focused" ethics (Tomasik 2016, 2017a, 2017b). There could also be massive simulations running on exoplanets that have been converted into computronium in which billions of sentient simulants suffer immense agony. Given the huge number of future beings who could exist if we do colonize space, it stands to reason that *someone somewhere* would run such simulations (perhaps from within a simulation), create new biospheres in which wild animals are subject to Darwinian misery, and so on. As Tomasik (2017a) speculates, "if I had to make an estimate now, I would give ~70% probability that if humans choose to colonize space, this will cause more suffering than it reduces on intrinsic grounds." The result could be an s-risk. Thus, the present paper offers a complimentary reason for rejecting the normative ideology of *space*

[20] One way for individual civilizations to escape this situation is to travel to the furthest reaches of space as soon as possible. This could enable them to ride the wave of the expanding universe such that it would be impossible not only for other spacecraft chasing but for a lightspeed signal to ever reach them. See Armstrong (2012). Another issue that was brought up during an Existential Risk to Humanity workshop at Chalmers University is that advanced technologies could enable future posthumans to live indefinitely long lives, indefinitely long lives could significantly increase one's aversion to risk, and an increase in one's aversion to risk could lower the probability of potentially lethal conflict. One possible response is that indefinitely long lives could increase the prevalence of boredom, and as Soren Kierkegaard once claimed, "boredom is the root of all evil" (Kierkegaard, 1987). Thus, the desire for a thrill, for the moments of "flow" that can only come by dancing along the threshold of annihilation, could motivate otherwise "immortal" posthuman civilizations to nonetheless engage in violent campaigns.

[21] Personal communication.

[22] I am here assuming that we establish a martian colony that becomes increasingly Earth-independent before we succeed in creating a friendly superintelligent machine.

[23] That is, on Bostrom's (2013) definition.

expansionism.

7. Conclusion

Let's now return to the topic of Section 1, i.e., the astronomical waste argument. According to Bostrom, our first priority is to reduce existential risk, because an existential catastrophe would prevent us from reaching a stable state of technological maturity and technological maturity is necessary to realize astronomical value. Furthermore, to reach technological maturity, we will need to colonize space. It follows that utilitarians (in particular) should prioritize existential risk reduction while also advocating for the colonization of space *as soon as possible*. Baum (2016) echos this sentiment when he argues that, if one accepts consequentialism, "space colonization should proceed with caution, but ultimately should proceed at immense scale."

Yet a closer look at what I have argued are the *most probable* results of colonizing the "last great frontier" suggests that doing so would yield a state of Hobbesian "warre" in which civilizations wallow in perpetual anxiety—*existential* anxiety—when they aren't actively engaged in confrontations with their neighbors. The argument that I present thus invites a Gestalt switch: rather than peering up at the firmament and pondering how much of our cosmic endowment of negentropy is being lost that could realize some form of positive "value," one should instead ponder how much negentropy is being lost that could realize an s-risk, or a condition marked by astronomical amounts of pain, misery, dread, fear, and suffering. In a phrase, *every second of delayed colonization should be seen as immensely desirable*, and the longer the delay, the better. This is not a conclusion that I find particularly appealing, yet I see no obvious flaws in the above arguments.

Acknowledgements

Thanks to Brian Tomasik and David Brin for helpful comments on this paper, and to attendees of the "Existential Risk to Humanity" workshop held at Chalmers University from September 1 to October 31.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Agar, N. (2010). *Humanity's end: Why we should reject radical enhancement.* Cambridge, MA: MIT Press.
Armstrong, S. (2012). *Von Neumann probes, dyson spheres, exploratory engineering, and the fermi paradox.* FHIOxfordhttps://www.youtube.com/watch?v=zQTfuI-9jIo.
Baum, S. (2016). The ethics of outer space: A consequentialist perspective. In S. J. Schwartz, & T. Milligan (Eds.). *The ethics of space exploration* (pp. 109–123). .
Bostrom, N. (2002). Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology, 9*(1).
Bostrom, N. (2003). Astronomical waste: The opportunity cost of delayed technological development. *Utilitas, 15*(3), 308–314.
Bostrom, N. (2005). What is a singleton? *Linguistic and Philosophical Investigations, 5*(2), 48–54.
Bostrom, N. (2013). Existential risk prevention as global priority. *Global Policy, 4*(1), 15–31.
Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies.* Oxford, UK: Oxford University Press.
Chomsky, N. (1976). Problems and mysteries in the study of human language. In A. Kasher (Ed.). *Language in focus: Foundations, methods, and systems*. Boston, MA: D. Reidel Publishing Company.
Cirincione, J. (2013). *Nuclear nightmares: Securing the world before it is too late.* New York, NY: Columbia University Press.
Ćirković, M. (2002). Cosmological forecast and its practical significance. *Journal of Evolution and Technology,* (12), 1–13. http://www.jetpress.org/volume12/CosmologicalForecast.pdf.
Clark, A. (2003). *Natural-born cyborgs: Minds, technologies, and the future of human intelligence.* Oxford, UK: Oxford University Press.
Clynes, N., & Kline, N. (1960). *Cyborgs and space. Astronautics* [September 1960].
Cole, D., & Cox, D. (1965). *Islands in space: The challenge of the planetoids.* New York, NY: Chilton Books.
Cotton-Barratt, O., & Ord, T. (2015). *Existential risk and existential hope: Definitions. FHI technical report*https://www.fhi.ox.ac.uk/Existential-risk-and-existential-hope.pdf.
Deudney, D, (in press). Dark Skies: Space Expansionism, Planetary Geopolitics, and the Ends of Humanity. Oxford: Oxford University Press.
Foreman, D. (1991). *Confessions of an eco-warrior.* New York, NY: Crown Publishers, Inc.
Friedman, T. (2005). *The world is flat: A brief history of the twenty-first century.* New York, NY: Farrar, Straus, and Giroux.
Gibbons, A. (2007). *Human evolution is speeding up. Science*http://www.sciencemag.org/news/2007/12/human-evolution-speeding.
Haija, R. (2006). The armageddon lobby: Dispensationalist Christian Zionism and the shaping of US policy towards Israel-Palestine. *Holy Land Studies, 5*(1), 75–95.
Hanson, R. (2016). *The age of em: Work, love, and life when robots rule the earth.* Oxford, UK: Oxford University Press.
Hobbes, T. (1982). *Leviathan.* New York, NY: Penguin Classics.
Jebari, K., Olsson-Yaouzis, N. (in press). A Game of Stars: On Why Active SETI Is a Very Bad Idea.
Kierkegaard, S. (1987). *Either/Or.* Princeton, NJ: Princeton University Press.
Kurzweil, R. (2005). *The singularity is near: When humans transcend biology.* New York, NY: Viking.
Laland, K., & O'Brien, M. (2010). Niche construction theory and archaeology. *Journal of Archaeological Method and Theory, 17*(4), 303–322.
Levy, J., & Thompson, W. (2010). *Causes of war.* Oxford, UK: Blackwell Publishing.
Pinker, S. (2011a). *The better angels of our nature: Why violence has declined.* New York, NY: Penguin Books.
Pinker, S. (2011b). *Jews, genes, and intelligence.* YIVO Institute for Jewish Researchhttps://www.youtube.com/watch?v=1GexZF5VIMU.
Roden, D. (2015). *Posthuman life: Philosophy at the edge of the human.* Oxon: Routledge.
Russell, B. (1954). *Human society in ethics and politics.* New York. NY: Routledge.
Sagan, C. (1983). Nuclear war and climatic catastrophe: Some policy implications. *Foreign Affairs, 62*(2), 257–292.
Sagan, C. (1994). *Pale blue dot: A vision of the human future in space.* New York, NY: Random House.
Sandberg, A., Armstrong, S., & Cirkovic, M. (2016). That is not dead which can eternal lie: The aestivation hypothesis. *Journal of the British Interplanetary Society, 69*(11), 406–415.
Sandberg, A. (2017). *Game theory of triggering vacuum decay.* Presentation.
Sandberg, A. (in press). [Draft of unnamed forthcoming book.].
Schneider, S. (2016). *It may not feel like anything to Be an alien. Nautilus*http://cosmos.nautil.us/feature/72/it-may-not-feel-like-anything-to-be-an-alien.
Sotala, K., & Gloor, L. (2017). Superintelligence as a cause or cure for risks of astronomical suffering. *Informatica, 41*, 389–400.
Tang, S. (2009). The security dilemma: A conceptual analysis. *Security Studies, 18*(3), 587–623.
Tomasik, B. (2016). *Will space colonization multiply wild-animal suffering?* Foundational Research Institutehttp://reducing-suffering.org/will-space-colonization-

multiply-wild-animal-suffering/.
Tomasik, B. (2017a). *Risks of astronomical future suffering.* Foundational Research Institutehttps://foundational-research.org/risks-of-astronomical-future-suffering/.
Tomasik, B. (2017b). *Gains from trade through compromise.* Foundational Research Institutehttps://foundational-research.org/gains-from-trade-through-compromise/.
Torres, P. Morality, Foresight, and Human Flourishing: An Introduction to Existential Risks, 2017a, Pitchstone Publishing; Durham, NC.
Torres, P. (2017b). Agential risks and information hazards An unavoidable but dangerous topic. *Futures, 95*, 86–97.
Torres, P. (2018a). *What are existential risks and how should we model them?* Unpublished Manuscript. https://bit.ly/2Hheh6L.
Torres, P. (2018b). Superintelligence and the future of governance: on prioritizing the control problem at the end of history. In R. Yampolskiy (Ed.). *Artificial intelligence safety and security*. New York, NY: Taylor & Francis Group.
Torres, P. Facing Disaster: The Great Challenges Framework. forthcoming, Foresight. Pre-publication manuscript: https://bit.ly/2HH1yd3.
Torres, P. (2018d). Who would destroy the world? Omnicidal agents and related phenomena. *Aggression and Violent Behavior, 39*, 129–138.
Waltz, K. (1979). *Theory of international politics.* Long Grove, IL: Waveland Press, Inc.
Weber, M. (1919). *The vocation lectures.* Indianapolis, IN: Hackett Publishing Company, Inc 2004.
Zuckerman, M., Silberman, J., & Hall, J. (2013). The relation between intelligence and religiosity. *Personality and Social Psychology Review, 17*(4), 325–354.