A Detailed Critique of One Section of Steven Pinker's Chapter "Existential Threats" in *Enlightenment Now*

Technical Report 2, Version 1.2

Phil Torres
Project for Future Human Flourishing

Key findings:

—> The first quarter or so of the chapter contains at least two quotes from other scholars that are taken "completely" out of context—that is, their original meaning is either in tension or outright contradictory with respect to the meaning implied by their use in this chapter. In both cases, the quotes play integral rhetorical, and to some extent substantive, roles in the argument that Pinker aims to develop.

—> The chapter expends a great deal of energy attacking a small village of straw men, from the pessimism/optimism dichotomy that frames the entire discussion to the theoretical dangers posed by value-misaligned machine superintelligence. I argue that this tendency to knock down unserious or non-existent positions while ignoring or misrepresenting the most intellectually robust ideas does a disservice to the ongoing public and academic discussions about the various global-scale threats facing humanity this century.

—> Many citations appear to have been poorly vetted. For example, Pinker relies on numerous non-scholarly articles to make what purport to be scholarly assertions about a wide range of topics that fall outside his area of expertise. In some cases, Pinker makes these claims with considerable confidence, thus giving non-expert readers—some of whom may be responsible for shaping domestic and foreign policies—a false sense of their tenability.

—> Along these lines, many of the sources that Pinker cites to support his theses contain some facts, evidence, or ideas that undercut those theses. Rather than acknowledging that alternative views are also compatible with (or supported by) the evidence, though, Pinker preferentially selects the facts, evidence, and ideas that support his narrative while simply ignoring those that don't. This is part of a larger issue of "cherry-picked" data. Indeed, I argue that, when the facts are more comprehensively considered, the positions that Pinker champions appear far less defensible.

—> Overall, the assessment presented below leads me to conclude that it would be unfortunate if this chapter were to significantly shape the public and academic discussions surrounding "existential risks." In the harshest terms, the chapter is guilty of misrepresenting ideas, cherry-picking data, misquoting sources, and ignoring contradictory evidence.

"My new favorite book of all time." That's how Bill Gates has described Steven Pinker's most recent book *Enlightenment Now*. Since I was an admirer of Pinker's previous book *The Better Angels of Our Nature*, which I have cited (approvingly) many times in the past, I was eager to get a copy of the new tome. In particular, I was curious about Pinker's chapter on "existential threats," since this is a topic that I've worked on for years in both a journalistic and academic capacity, publishing numerous articles in popular media outlets and scholarly journals as well as two books on the topic (one of which Pinker mentions in *Enlightenment Now*). Thus, unlike world history, evolutionary psychology, and economics—all of which Pinker discusses with apparent erudition—this is a subject on which I have expertise and, consequently, can offer a thorough and informed evaluation of Pinker's various theses.

The present document does precisely this by dissecting individual sentences and paragraphs, and then placing them under a critical microscope for analysis. Why choose this unusual approach? Because, so far as I can tell, almost every paragraph of the chapter contains at least one misleading claim, problematic quote, false assertion, or selective presentation of the evidence.[1] Given (i) the ubiquity of such problems—or so I will try to show in the cooperative spirit of acquiring a better approximation of the truth[2]—along with (ii) the fact that *Enlightenment Now* will likely become a massively influential, if not canonical, book among a wide range of scholars and the general public, it seems important that *someone* takes the time to comb through the chapter on existential threats (again, my area of expertise) and point out the various problems, ranging from the trivial to the egregious, that it encounters.

To be clear, I think Pinker's overall contribution to culture, including intellectual culture, has been *positive*: humanity really has made measurable progress in multiple domains of well-being, morality, knowledge, and so on, and people ought to know this—if only to ward off the despair that reading the daily headlines tends to elicit. But I also believe that Pinker suffers from a scotoma in his vision of our collective existential plight: while violence has declined and our circles of moral concern have expanded, large-scale human activity and increasingly powerful "dual-use" technologies have introduced—and continue to introduce—a constellation of historically unique hazards that genuinely threaten our species' future on spaceship Earth. There is no contradiction here! Indeed, I have often recommended (before *Enlightenment Now*) that people read *Better Angels* alongside books like *The Future of Violence*, *Our Final Hour*, *Global Catastrophic Risks*, and *Here Be Dragons* to acquire a more *complete* picture of our (rapidly) evolving survival situation.[3] The major problem with Pinker's Enlightenment progressionism is thus one of incompleteness: he simply ignores (or misinterprets, in my view) a range of phenomena and historical trends that clearly indicate that, as Stephen Hawking soberly put it, "this is the most dangerous time for our planet."[4] Again, any perceived contradiction is illusory: the moment in history with the lowest rates of violence (etc.) also contains more

[1] Basic epistemology, of course, demands that one considers the *totality* of evidence, not just some shred that confirms one's prior or preferred beliefs. Pinker would obviously not disagree with this—but in practice, as we'll see, one could question the chapter's dedication to this paramount principle.

[2] To be clear, I don't "pull any punches" here—but nor is this intended to be, in any way, an unfriendly critique. I try to be candid with my criticisms—and with my intellectual disappointment with certain parts of Pinker's chapter—but I also don't see this document as the final word on these matters—not at all.

[3] Note that this statement entails (the true claim) that I have often recommended *Better Angels*. It is a book that I, in general, highly respect (although see below).

[4] For additional statements about this fact, see section 1 of this paper.

global risk potential than any other in the past 200,000 years.[5]

This is a general criticism of Pinker's progressionist project, in contrast to the more specific criticisms below. Another general complaint is that, with respect to the existential threats chapter, Pinker doesn't appear to be sufficiently conversant with the scholarly literature to put forth a strong, much less trenchant, criticism of (certain aspects of) the topic. Consistent with this are the following two facts: first, Pinker hardly cites *any* scholars within the field of existential risk studies; and second, the preface of the book suggests that Pinker didn't consult a *single* existential risk scholar while preparing the manuscript. If one wishes to present a fair, ideologically-neutral account of existential threats—especially if one's purpose is to knock the topic down to size—then surely it behooves one to seek the advice of actual experts and peruse the relevant body of the most serious scholarship.[6] This may sound harsh as stated but, as we will see, Pinker's chapter expends considerable energy fallaciously beating to death a small village of straw men.

Pinker not only ignores the scholarly literature on existential risks, though, he often relies upon popular media articles and opinion pieces in *Reason*, *Salon*, *Wired*, *Slate*, *The Guardian*, and *The New York Times* to support his claims. (The *Reason* citation in particular is deeply problematic, as we will explore below.) Not all of these media platforms are created equal, of course, and in fact the *Salon* article that Pinker cites (as additional reading) is one that I wrote about superintelligence more than two years ago. But given the *very* general audience that I had in mind while writing, it shouldn't have ended up in an "authoritative" book like Pinker's, or so I would argue.[7] (I—and plenty of others with *even more* competence on the topic—have numerous peer-reviewed articles, book chapters, etc. on superintelligence! A serious analysis of this ostensible risk, which Pinker purports to provide, should have cited these, and only these, instead.)

But the problems with Pinker's chapter are even more significant than this. The chapter also suffers from what I would describe as cherry-picked data, questionable citations, a few out-of-context quotes, and other scholarly infractions. One might argue that this is somewhat unsurprising given similar problems in *Better Angels*. For example, some investigative digging by Magdi Semrau, a communication science and disorders PhD student, finds that a single paragraph in *Better Angels*: (i) cites a non-scholarly book whose relevant citation (given Pinker's citation) is of a discredited academic article; (ii) bases a cluster of propositions, which Pinker presents as *fact*, on two sources: (a) a mere *opinion* expressed by a well-known anti-feminist whose employer is the American Enterprise Institute, a conservative think tank, and (b) an *op-ed* piece also written by an anti-feminist crusader, published in the non-scholarly, partisan magazine *City Journal*;[8] and (iii) references a survey from the Bureau of Justice Statistics but leaves out aspects of

---

[5] With perhaps the one exception of the Toba supereruption, which occurred ~75,000 years ago.

[6] This has nothing to do with the Courtier's Reply fallacy. Person A not knowing about X does not itself undercut A's arguments about X; but it could very well *explain* why A's arguments about X fail to hit any relevant targets.

[7] Indeed, the editors at *Salon* made the bad decision to include a picture of the Terminator with this article; I will explain below why this immediately undercut the article's credibility. Second, note that Pinker doesn't cite my article as, say, an example of "bad scholarship" or "fear-mongering in public." He cites it under the heading of "Robots turning us into paper clips and other Value Alignment Problems." It should not be there.

[8] *Please register the subtle point, which will no doubt be lost on some readers, that this has nothing to do with "feminism" itself.* The problem is that a proposition is presented as fact when its basis consists of the opinions of two individuals who have an ideological interest in persuading others to accept that proposition.

the survey that don't support the narrative being spun.[9] Furthermore, Semrau notes that Pinker includes data in a graph about homicide rates that is deeply flawed—and *known to be this* since 1998. Although Semrau has yet to organize these discoveries into a proper paper, they are sufficiently well-supported to warrant concern about the scholarly practices embraced in *Better Angels*—and thus *Enlightenment Now*.[10] Indeed, they are precisely the sort of tendentious (if that's not too loaded of a word) shortcuts that we will encounter many times below.

Given that Pinker's chapter on existential threats is quite long and the process of responding to each paragraph is tedious (although perhaps the tedium will be even worse for the reader!), I have here only reproduced part of this chapter. If readers find it particularly useful, then I would consider responding to the rest of the chapter as well.

* * *

Pinker begins the chapter with:

But are we flirting with disaster? When pessimists are forced to concede that life has been getting better and better for more and more people, they have a retort at the ready. We are cheerfully hurtling toward a catastrophe, they say, like the man who fell off the roof and says "So far so good" as he passes each floor. Or we are playing Russian roulette, and the deadly odds are bound to catch up to us. Or we will be blindsided by a black swan, a four-sigma event far along the tail of the statistical distribution of hazards, with low odds but calamitous harm.

This gets the entire conversation off to a bad start. First, my reading of this chapter is that it's targeting, at least in part, the field of "existential risk studies," which has spawned a number of public discussions about biotechnology, synthetic biology, advanced nanotechnology, geoengineering, artificial intelligence, and so on. In fact, Pinker has elsewhere specifically attacked existential risk studies by calling its central concept (i.e., *existential risks*) a "useless category."

If this reading is correct, then Pinker's reference to "pessimists" is quite misleading. Many of the scholars who are *the most concerned* about existential risks are also pro-technology "transhumanists" and "techno-progressives"—in some cases, even Kurzweilian "singularitarians"—who explicitly hope, if not positively expect, technological innovation to usher in a techno-utopian future world marked by the elimination of all diseases, indefinite lifespans, "radical" cognitive and moral enhancements, mind-uploading, Dyson swarms, colonization of the galaxy and beyond, "radical abundance" (as Eric Drexler puts it), the creation of a type III (or higher) civilization (on the Kardashev scale), and so on. Indeed, most scholars working on existential risks unhesitantly *en-*

[9] As Semrau reminds us 24 tweets into the thread: "Again: *This is the source material for Steven Pinker's work*."
[10] Along these lines, the anthropologist Douglas Fry, whose work focuses on the "anthropology of war and peace, conflict resolution, nonviolence, [and] human rights," notes that Pinker fails to provide a *single citation* for the many claims made in this paragraph, which Fry claims is deeply flawed from an evidential perspective: "Foraging peoples can invade to gain territory, such as hunting grounds, watering holes, the banks or mouths of rivers, and sources of valued minerals like flint, obsidian, salt, or ochre. They may raid livestock or caches of stored food. And very often they fight over women. Men may raid a neighboring village for the express purpose of kidnapping women, whom they gang-rape and distribute as wives. They may raid for some other reason and take the women as a bonus. Or they may raid to claim women who had been promised to them in marriage but were not delivered at the agreed-upon time. And sometimes young men attack for trophies, coups, and other signs of aggressive prowess, especially in societies where they are a prerequisite to attaining adult status." See also Brian Ferguson's book chapter titled "Pinker's List: Exaggerating Prehistoric War Mortality." As one scholar, with whom I was in personal communication about his own work being used by Pinker, similarly asks, "How this guy managed to become a public intellectual in fields so far removed from his expertise is something to wonder at."

dorse the sort of Enlightenment progressionism for which Pinker evangelizes, even identifying such progress as a *reason* to take existing and emerging existential hazards seriously. I myself begin my book *Morality, Foresight, and Human Flourishing: An Introduction to Existential Risks* (hereafter, "*Morality*") with an affirmation of scientific, technological, and moral progress over time, especially since the Enlightenment; and Nick Bostrom, a leading transhumanist who more or less founded the field of existential risk studies (along with John Leslie), has literally written an ebullient article titled "Letter from Utopia" that describes the unfathomably blissful lives of future posthumans, who we could become if only we promote the values of technological progress and, in Bostrom's words, have "the opportunity to explore the transhuman and posthuman realms." One can be hopeful about a better future and still shout, "Oh my lord, there's a lion running toward us!"[11]

So, this is not an either/or situation—and this is why Pinker framing the issue as an intellectual battle between optimists and pessimists distorts the "debate" from the start. This being said, there no doubt are, as Pinker gestures at below, neo-Luddites, romantics, environmentalists, and people espousing certain moral theories (e.g., antinatalism[12]) who champion pessimistic views about humanity's past and/or future. But the large majority of individuals who are worried about existential risks don't fall within any of these categories. Rather, like the technocratic, idealist, neoliberal, space expansionist, visionary entrepreneur Elon Musk—who has repeatedly made anxious noises about the behemoth dangers of superintelligence—they see technology as a Janus-faced, double-edged sword (if readers don't mind mixed metaphors).

We should also mention that yes, indeed, we are playing Russian roulette to some extent, although the "deadly" odds are not necessarily "bound to catch up to us"! (I don't know of any prominent thinker in the field who believes this.) No species in our genus has ever before, in our ~2-million-year career on Earth, had to confront global-scale problems like anthropogenic climate change, the Anthropocene extinction, dual-use emerging technologies, and perhaps even computers whose problem-solving capabilities exceed that of the best humans in every cognitive domain. This is a historical fact, of course: we have no track record of surviving such risks. It follows that (i) given the astronomical potential value of the future (literally trillions and trillions and trillions of humans living worthwhile lives throughout the universe), and (ii) the growing ability for humanity to destroy itself through error, terror, global coordination failures, and so on, (iii) it would be extremely imprudent *not* to have an ongoing public and academic discussion about the number and nature of existential hazards and the various mechanisms by which we could prevent such risks from occurring. That's not pessimism! It's realism combined with the virtues of wisdom and deep-future foresight.

For half a century the four horsemen of the modern apocalypse have been overpopulation, resource shortages, pollution, and nuclear war.

[11] A particularly nice quote about this issue comes from David Denkenberger, who writes, "Optimists tend to ignore the risks, and I was guilty of that for years (as Mark Twain said, 'Denial is not just a river in Egypt.'). Pessimists tend to take the risk seriously, but don't think we can do anything about them. Very few people actually take the risks seriously and think we can do something about them, which is one reason why so little work gets done on them." Existential risk scholars tend to fall quite squarely within this "very few people" category.

[12] By this I mean that for antinatalists who believe that humanity will survive for centuries, millennia, or longer, and who also maintain that the morally best outcome for our species would be near-term, voluntary extinction, the future—so full of suffering as it will be—looks bleak. The result is a kind of ethico-futurological pessimism.

They have recently been joined by a cavalry of more exotic knights: nanobots that will engulf us, robots that will enslave us, artificial intelligence that will turn us into raw materials, and Bulgarian teenagers who will brew a genocidal virus or take down the Internet from their bedrooms.

A quick note about epistemology: it's crucial for readers to recognize that, when it comes to *evaluating* the legitimacy of a given risk, its "sounds crazy" quality is irrelevant. Consider the statements: "Over geological time, one species can evolve into another" and "if your twin were to board a spaceship and fly to Saturn and back, she would have aged less than you." Both sound—to naive ears "uncorrupted" by science—utterly absurd. Yet it is *epistemically reasonable* to accept them because the evidence and arguments upon which they're founded are strong. Thus, don't be fooled by the extent to which some emerging or anticipated future risks sound silly. Epistemology doesn't care about *what* a proposition says (content), it cares about *why* one might accept it (reasons).

The sentinels for the familiar horsemen tended to be romantics and Luddites. But those who warn of the higher-tech dangers are often scientists and technologists who have deployed their ingenuity to identify ever more ways in which the world will soon end.

To my ear, the second sentence makes it sound like devising new doomsday scenarios is a hobby: something done for the fun of it, for its own sake. That's not the case. As mentioned above, the future could contain immense amounts of moral, intellectual, scientific, etc. value; in *Morality*, I call this the "astronomical value thesis." It follows that one of the most important tasks that anyone could engage in is to increase, even if by minuscule increments, the probability that humanity avoids an existential catastrophe. This idea is formalized in Nick Bostrom's "maxipok rule," which essentially states that "the loss in expected value resulting from an existential catastrophe is so enormous that the objective of reducing existential risks should be a dominant consideration whenever we act out of an impersonal concern for humankind as a whole." Thus, *toward this end*, a relatively tiny group of scholars have indeed labored to identify as many existential risk scenarios as possible—not to scare people, declare that "we're all doomed," or give existential riskologists one more reason to lay awake at night with sweaty palms and dilated pupils, but to devise a regimen of effective strategies for *avoiding* an existential catastrophe. Given what's at stake, even a small reduction in overall existential risk could have an immense payoff.

In 2003, the eminent astrophysicist Martin Rees published a book entitled Our Final Hour *in which he warned that "humankind is potentially the maker of its own demise" and laid out some dozen ways in which we have "endangered the future of the entire universe." For example, experiments in particle colliders could create a black hole that would annihilate the Earth, or a "strangelet" of compressed quarks that would cause all matter in the cosmos to bind to it and disappear.*

Note that these statements are true: particle colliders could, in theory, destroy the earth, although this appears unlikely—but see this important article by Toby Ord, Rafaela Hillerbrand, and Anders Sandberg for complications.

Rees tapped a rich vein of catastrophism.

This short sentence strikes me as overly dismissive. Again, the entire point of existential risk studies—a nascent field of empirical and philosophical inquiry that receives a relative pittance of funding and has fewer publications than the subfield of ento-

mology dedicated to studying dung beetles—is to better understand the various hazards that could seriously and permanently affect the well-being of our species. That's it.

The book's Amazon page notes, "Customers who viewed this item also viewed Global Catastrophic Risks*;* Our Final Invention: Artificial Intelligence and the End of the Human Era*;* The End: What Science and Religion Tell Us About the Apocalypse*; and* World War Z: An Oral History of the Zombie War*." Technophilanthropists have bankrolled research institutes dedicated to discovering new existential threats and figuring out how to save the world from them, including the Future of Humanity Institute, the Future of Life Institute, the Center for the Study of Existential Risk, and the Global Catastrophic Risk Institute.*

(Note that "Center" in "Center for the Study of Existential Risk" should be "Centre.")

How should we think about the existential threats that lurk behind our incremental progress? No one can prophesy that a cataclysm will never happen, and this chapter contains no such assurance. But I will lay out a way to think about them, and examine the major menaces. Three of the threats—overpopulation, resource depletion, and pollution, including greenhouse gases—were discussed in chapter 10, and I will take the same approach here. Some threats are figments of cultural and historical pessimism. Others are genuine, but we can treat them not as apocalypses in waiting but as problems to be solved.

This last sentence seems to knock down a(nother) straw man. I don't know of a single scholar in the field—and this is not from lack of familiarity—who believes that there are "apocalypses in waiting." Even when Bostrom writes that we should recognize the "default outcome" of machine superintelligence as "doom," he's saying that *unless we solve the control problem, then almost by definition the consequences will be existential, so let's allocate the necessary resources to solve the control problem, please?* And he provides an entire book of rather nuanced, sophisticated, and philosophically formidable arguments to support this conclusion.[13] (We will return to this issue later.)

The reigning view among existential risk scholars is thus precisely what Pinker advocates: secular apocalypses like nuclear winters, engineered pandemics, and superintelligence takeovers are seen as *problems to be solved*. Since one can't solve these problems without doing the relevant research—or communicating with the public so that they vote for political leaders who understand and care about the relevant challenges—the fledgling "interdiscipline" of existential risk studies was born! Yet Pinker writes that:

At first glance one might think that the more thought we give to existential risks, the better. The stakes, quite literally, could not be higher. What harm could there be in getting people to think about these terrible risks? The worst that could happen is that we would take some precautions that turn out in retrospect to have been unnecessary.

Note that the phrases "thought we give to existential risks" and "getting people to think about these terrible risks" are ambiguous. By "we" and "people," is Pinker referring to (Group A) scientists, philosophers, policymakers, and other specialists, or (Group B) the general public comprised of

[13] Note that nowhere in this chapter—as we will see—does Pinker address the various issues that Bostrom outlines, such as the instrumental convergence thesis. It is perplexing that some intellectuals believe that they have demolished concerns about superintelligence without engaging in the actual arguments for concern.

non-experts? This is important to disambiguate because there could be quite distinct reasons for promoting the concept of existential risks to one group but not the other (or vice versa, or neither). Conflating the two is thus problematic, as the very next paragraph illustrates:

> *But apocalyptic thinking has serious downsides. One is that false alarms to catastrophic risks can themselves be catastrophic. The nuclear arms race of the 1960s, for example, was set off by fears of a mythical "missile gap" with the Soviet Union. The 2003 invasion of Iraq was justified by the uncertain but catastrophic possibility that Saddam Hussein was developing nuclear weapons and planning to use them against the United States. (As George W. Bush put it, "We cannot wait for the final proof—the smoking gun-that could come in the form of a mushroom cloud.") And as we shall see, one of the reasons the great powers refuse to take the common-sense pledge that they won't be the first to use nuclear weapons is that they want to reserve the right to use them against other supposed existential threats such as bioterror and cyberattacks.*[2] *Sowing fear about hypothetical disasters, far from safeguarding the future of humanity, can endanger it.*

If Pinker means to include the public in this passage, one could argue that what matters isn't *that* the public is warned about "hypothetical disasters" but *how* they are warned. After all, as mentioned above, the public is responsible for deciding who ends up with the political clout to catalyze societal change—indeed, this is one reason that (the now-disgraced[14]) Lawrence Krauss once told me in an interview about the Doomsday Clock:

> As responsible citizens, we can vote. We can pose questions to our political representatives. And that's a major factor. Politicians actually are accountable, and if lots of people phone them with questions or issues, politicians will listen. The second thing is that we all have access to groups, although some of us have bigger soapboxes than others. School groups, church groups, book clubs—we can all work to educate ourselves and our local surroundings, on a personal basis, to address these issues. The last thing anyone should feel is completely hopeless or powerless. We certainly affect our daily lives in how we utilize things, but also we affect our community in various ways. So, we have to start small, and each of us can do that. And, of course, if you're more interested [in working to reduce the threat of a catastrophe], *you can organize a local group and have sessions in which you educate others about such issues*. The power of voting and the power of education—those are the two best strategies.[15]

Furthermore, George Bush's politically-motivated and often mendacious exclamations about Saddam—some of which were based on cherry-picked intelligence—are quite unlike the rather "clinical" warnings of scholars like Lord Martin Rees, Nick Bostrom, Stephen Hawking, Anders Sandberg, Jason Matheny, Richard Posner, Max Tegmark, and countless climatologists, ecologists, biotech-

[14] It's worth noting here that Pinker thanks the "skeptic" Michael Shermer in the preface of *Enlightenment Now*, despite years of credible allegations of sexual harassment, assault, and even rape. To my eye, this undercuts Pinker's avowed commitment to Enlightenment values—in particular, to the important value of respecting women and *believing them* when they report patterns of sexual misconduct. As Sean Carroll writes, "if the Enlightenment is your thing (and it should be!), nothing should outrage you more about our current society than the fact that women/minorities/LGBTQ are systematically discriminated against."

[15] Italics added. It's also worth noting that, as Olle Häggström points out, ignoring potential existential risks "seems to fly straight in the face of one of Pinker's most cherished ideas during the past decade or more, namely that of scientific and intellectual openness, and Enlightenment values more generally. … surely the approach best in line with Enlightenment values is … to openly discuss the problem and to try to work out whether the risk is real."

nologists, synthetic biologists, nanotechnologists, and other experts. One might also wonder why Pinker ignores those instances when dire warnings actually did gesture at some real hazard. For example, many observers made (what critics at the time could have described as) "alarmist" or "hyperbolic" claims about the march of Nazi Germany in the 1930s; yet Neville Chamberlain conceded lands to Hitler on the (false) assumption that this would mollify him and the US didn't enter the war until 1941b (after the attack on Pearl Harbor). In other words, if only such warnings had been heeded, World War II might not have left some 80 million people in the muddy grave. Furthermore, concerns about the catastrophic effects of ozone depletion during the 1980s led to the Montreal Protocol of 1987, which effectively averted what most experts agree would have been a disastrous state of affairs for humanity.

So, one could easily retort that "apocalyptic thinking *can also have serious upsides*" by citing instances in which shouts about death and doom either did or probably could have obviated major calamities, if only they were taken seriously.[16] In his *Global Catastrophic Risks* chapter about millennialist tendencies, James Hughes examines a number of historical cases that lead him to a similar conclusion, namely, that

> millennialist energies can overcome social inertia and inspire necessary prophylaxis and force recalcitrant institutions to necessary action and reform. In assessing the prospects for catastrophic risks, and potentially revolutionary social and technological progress, can we embrace millennialism and harness its power without giving in to magical thinking, sectarianism, and overly optimistic or pessimistic cognitive biases? … I believe so: understanding the history and manifestations of the millennial impulse, and scrutinizing even our most purportedly scientific and rational ideas for their signs, should provide some correction for their downsides.[17]

A second hazard of enumerating doomsday scenarios is that humanity has a finite budget of resources, brainpower, and anxiety. You can't

---

[16] I am reminded here of Pinker listing some predictions of doom that didn't come true—presumably in an attempt to embarrass those who have warned about future catastrophes—but ignoring all the predictions of utopia that similarly never came to pass. As the Princeton historian David Bell points out, "Pinker fails to acknowledge how very closely his own radical optimism echoes some of the wilder—and more misguided—pronouncements about the human future from the Enlightenment itself. 'The human species … is capable of … unbounded improvement … mankind in a later age are greatly superior to mankind in a former age.' This is not Pinker, but Joseph Priestley, writing in 1771. 'No bounds have been fixed to the improvement of the human faculties…the perfectibility of man is absolutely indefinite.' This time, the words come from the Marquis de Condorcet, in 1793–94. Even as Rousseau denounced progress, and Diderot and Voltaire cast a skeptical eye toward it, many other philosophes confidently predicted the end of war, the eradication of disease, and the worldwide spread of liberty. That few of these things have been fully realized after more than two centuries should, perhaps, have given Pinker pause. … A few months after writing his paean to human perfectibility, Condorcet committed suicide in prison during the Reign of Terror."

[17] It may also be worth mentioning that a central topic within existential risk studies is that of "information hazards," or "risks that arise from the dissemination or the potential dissemination of true information that may cause harm or enable some agent to cause harm." The subtypes of *idea hazards* ("a general idea, if disseminated, creates a risk"), *attention hazards* ("the mere drawing of attention to some particularly potent or relevant ideas or data increases risk, even when these ideas or data are already 'known'"), and *evocation hazards* ("there can be a risk that the particular mode of presentation used to convey some content can activate undesirable mental states and processes") are especially relevant. Pinker does not *seem* to be aware of these phenomena, although perhaps I'm wrong. An example of the latter might be Musk saying that "with artificial intelligence, we're summoning the demon"—a colorful statement that could easily be interpreted as mere hyperbole, thus leading some to dismiss the more serious arguments for why superintelligence is dangerous—although one could also argue that such eye-popping rhetoric is necessary to capture people's attention and persuade them to take the risk seriously (see below). The point is that many existential risk scholars take information hazards very seriously and, as a result, are quite cautious about *which* nuggets of information are conveyed to the public and *how*.

worry about everything. Some of the threats facing us, like climate change and nuclear war, are unmistakable, and will require immense effort and ingenuity to mitigate. Folding them into a list of exotic scenarios with minuscule or unknown probabilities can only dilute the sense of urgency.

The problem with this passage is the word—also used above—"exotic." The fact is that most serious analyses of "dual-use" emerging technologies, both from intelligence agencies and the academic community, conclude that they could carry *far more profound risks* to the long-term survival of humanity than climate change or nuclear war (the two biggest *existing* risks). Why? One reason is that—as we'll discuss at the end of this document—such technologies are simultaneously becoming more *powerful* and *accessible*. The result is that a growing number of lone wolves and terrorist organizations are gaining the technological capacity to wreak ever-more devastating harm on civilization.[18]

Put differently, consider what Leó Szilárd famously wrote after he successfully initiated a chain reaction with uranium in 1939: "We turned the switch and saw the flashes. We watched them for a little while and then we switched everything off and went home. That night, there was very little doubt in my mind that the world was headed for grief." This captures precisely what many scholars who study *anthropogenic* existential risks in particular feel: unless humanity seriously examines how malicious agents could misuse and abuse (i) *emerging artifacts* like CRISPR/Cas-9, base editing, digital-to-biological converters, "slaughterbots," advanced AI systems, SILEX (i.e., "separation of [uranium] isotopes by laser excitation"), and (ii) *future anticipated technologies* like autonomous nanobots, nanofactories, and "stratospheric sulfate aerosol deposition" techniques (for the purpose of geoengineering), then the world may be headed for grief. These are not "exotic" dangers in the sense that Pinker seems to mean: they concern dual-use technologies currently being developed and some that appear very likely, if not almost certain, to be developed in the foreseeable future.

Perhaps the only risk discussed in the literature that could aptly be described as "exotic" is the possibility that we live in a computer simulation and it gets shut down. Yet even this scenario is based on a serious philosophical argument—the "simulation argument," of which one aspect is the "simulation hypothesis"—that has not yet been refuted, at least to the satisfaction of many philosophers. To my eye, the word "exotic" is far too facile, and it suggests (to me) that Pinker has not seriously perused the body of scholarly work on existential dangers to humanity. (For a comprehensive list of risk scenarios that are taken seriously by the community, see my book *Morality* and this report by the Global Challenges Foundation.)

Recall that people are poor at assessing probabilities, especially small ones, and instead play out scenarios in their mind's eye. If two scenarios are equally imaginable, they may be considered equally probable, and people will worry about the genuine hazard no more than about the science-fiction plotline. And the more ways people can imagine bad things happening, the higher their estimate that some thing bad will happen.

This is why cognitive biases are so strongly emphasized within the field. Indeed, there's an entire chapter dedicated to this topic in the seminal *Global Catastrophic Risks* edited collection, and I begin *Morality* with a section in Chapter 1 titled "Biases and Distortions," about the many ways that bad mental software can lead us to incorrect conclu-

[18] It is also worth mentioning here that a "minuscule" probability is not the same as an "unknown" probability. In the former case, there is no reason for concern, whereas this is not true for the latter.

sions—including conclusions that the overall risk to human survival is high *or* low.

> *And that leads to the greatest danger of all: that people will think, as a recent* New York Times *article put it,"These grim facts should lead any reasonable person to conclude that humanity is screwed."*[3]

This is a somewhat odd article to cite here. First of all, it's a short review of the journalist Dan Zak's book *Almighty: Courage, Resistance, and Existential Peril in the Nuclear Age*. It's *not* an article about "existential threats" in general. Second, by "screwed," the author of the review, Kai Bird, *isn't* saying that humanity is destined to go extinct or civilization is bound to collapse next year; he's merely referring to the use of one or more nuclear weapons. And third, the larger point that he's making is simply that a nuclear weapon being detonated appears to be inevitable given that (i) "a quarter-century after the end of the Cold War, nine nations possess some 16,000 nuclear warheads; the United States and Russia each have more than 7,000 warheads" and "four countries—North Korea, Pakistan, India and Israel—have developed nuclear arsenals and refuse to sign the Treaty on Non-Proliferation of Nuclear Weapons," and (ii) a nuclear bomb could be smuggled into New York City by only "three or four men." To support the latter claim, Bird quotes Robert Oppenheimer who, in response to a question about whether this is possible, avers "of course it could be done." As Bird puts it—and this statement is almost certainly true, as simple arithmetic affirms—"the odds are that these weapons will be used again, somewhere and probably in the not-so-distant future." It's hard to see how Pinker's "greatest danger of all" statement follows from this citation, since the book review isn't about multiple risk scenarios but the specific risk of nuclear conflict (which is indeed serious).[19]

Here we should also reiterate that the large majority of "technodoomsters"—Pinker's coinage—who are nervous about existential risks does not believe that humanity is "screwed," at least not in the sense that our extinction is certain within the coming decades or centuries (or even before some 10^{40} years in the future, at which point all protons in the universe will have decayed). I can't think of a single notable scholar who holds this view. There are a few conspiratorial, fringe figures like Guy McPherson who've made such claims but, as such, these individuals are *not at all* representative of the far more modest, tentative "mainstream" positions within the field of existential risk studies. Indeed, perhaps the most radical estimate from a respectable scholar comes from Lord Martin Rees, who proposed the conditional argument that *unless* humanity alters the developmental trajectory of civilization in the coming decades, *then* it may be that "the odds are no better than fifty-fifty that our present civilisation on Earth will survive to the end of the present century."[20] This is not a fatalistic declaration that we're "screwed." Rather, Rees's warning is more like a doctor telling a patient: "If you don't make certain lifestyle changes right away, then there's a roughly 50 percent chance that you'll perish"—in contrast to, "No matter what you do at this point, you're a goner, sucker!" Alerting others that humanity is in great

[19] Bird himself has affirmed to me that Pinker takes this quote out of context.

[20] Indeed, the very next sentence states: "Our choices and actions could ensure the perpetual future of life (not just on Earth, but perhaps far beyond it, too). Or in contrast, through malign intent, or through misadventure, twenty-first century technology could jeopardise life's potential, foreclosing its human and posthuman future. What happens here on Earth, in this century, could conceivably make the difference between a near eternity filled with every more complex and subtle forms of life and one filled with nothing but base matter."

danger is not tantamount to declaring that all hope is lost.

If humanity is screwed, why sacrifice anything to reduce potential risks? Why forgo the convenience of fossil fuels, or exhort governments to rethink their nuclear weapons policies? Eat, drink, and be merry, for tomorrow we die! A 2013 survey in four English-speaking countries showed that among the respondents who believe that our way of life will probably end in a century, a majority endorsed the statement "The world's future looks grim so we have to focus on looking after ourselves and those we love."

Few writers on technological risk give much thought to the cumulative psychological effects of the drumbeat of doom.

First, the "drumbeat of doom" is misleading, for reasons discussed above (and below). Second, it's not entirely true that few writers have thought hard about the relevant psychological effects. For example, in *Morality*, I repeatedly emphasize the astronomical value thesis (a reason for hope), endorse what Paul Romer calls "conditional optimism" (which Pinker also cites in *Enlightenment Now*), and quote Cormac McCarthy's witticism that "I'm a pessimist but that's no reason to be gloomy!" Furthermore, I underline the dangers associated with what Jennifer Jacquet calls the "anthropocebo effect," which refers to "a psychological condition that exacerbates human-induced damage—a certain pessimism that makes us accept human destruction as inevitable."

Even more, in one of the *founding documents* of the field, Bostrom conjectures that the depressing nature of the topic may be one reason that it has received so little scholarly attention—to which he adds that "the point [of existential risk studies] is not to wallow in gloom and doom but simply to take a sober look at what could go wrong so we can create responsible strategies for improving our chances of survival." Other scholars within the field, including Owen Cotton-Barratt, Toby Ord, and Max Tegmark, have highlighted the importance of "existential hope," which brings to the foreground a sense of how good things could turn out (if we play our cards right).[21] These are just a few of many examples that could be adduced to corroborate the claim that many "writers on technology risk" have thought very much, or very deeply, about mental effects of cogitating doomsday scenarios.

As Elin Kelsey, an environmental communicator, points out,"We have media ratings to protect children from sex or violence in movies, but we think nothing of inviting a scientist into a second grade classroom and telling the kids the planet is ruined. A quarter of (Australian) children are so troubled about the state of the world that they honestly believe it will come to an end before they get older."[5]

Here, Pinker cites a website called "Ocean Optimism," which, on a separate page, supports the last sentence above by citing the paper "Hope, Despair, and Transformation," which itself cites a report titled "Children's Fears, Hopes, and Heroes." (Why not cite the original? It's not clear.) The original source of this data also notes that "44% of children are worried about the future impact of climate change," and "43% of children worried about pollution in the air and water." At least from one perspective, this is quite encouraging: young people are taking the very real dangers of environmental degradation seriously. While fear can be

[21] This is somewhat imprecise: existential hope is presented as the hope that one or more "eucatastrophes" occur in our future light cone, where a "eucatastrophe" is any event that significantly increases the expected value of the future.

crippling—it is, at times, the "brother to panic"—it can also be a great motivator.

According to recent polls, so do 15 percent of people worldwide, and between a quarter and a third of Americans.[6] *In* The Progress Paradox*, the journalist Gregg Easterbrook suggests that a major reason that Americans are not happier, despite their rising objective fortunes, is "collapse anxiety": the fear that civilization may implode and there's nothing anyone can do about it.*

Is there any empirical data to support this thesis, though? So far as I can tell, the answer is "no." Thus, one might wonder: What if worrying about the end of the world provides the necessary impetus to recycle more, fly less, donate to the Future of Life Institute, plant a tree, stop using plastic bags, earn a degree in philosophy, ecology, or computer science, educate others about the "intertwined" promise and peril of technology, and vote for political leaders who care about the world beyond the next election cycle?[22] In my own case—that is, on a personal level—realizing (i) how much potential future value there is to lose by succumbing to an existential catastrophe, and (ii) the extent to which the existing and emerging risks to humanity are historically unprecedented, are what *inspired* me to dive into and contribute to the literature.[23] For me, futurological fear has been the greatest driver of intellectual activism to ensure a good future for humanity.

In fact, Easterbrook argues that "collapse anxiety is essential to understanding why Americans do not seem more pleased with the historically unprecedented bounty and liberty in which most live." But nowhere does he provide an argument or evidence for "collapse anxiety" *being essential*. For example, after listing a number of desirable trends pertaining to life expectancy, health, education, and comfort, he simply asserts that "collapse anxiety hangs over these achievements, engendering subliminal fear that prosperity will end."

It's also worth noting that Easterbrook—again, someone who Pinker cites approvingly, and whose book Google amusingly categorizes as "Self-Help"[24]—claims that "if a collapse were coming, its signs ought to be somewhere. That is not what trends show. Practically everything is getting better." But this is demonstrably *false* with respect to, say, climate change and global biodiversity loss, both of which have fueled only the sixth mass extinction event in life's 3.8-billion-year history. These phenomena are not "getting better" in any sense, and indeed their catastrophic effects, some of which are already irreversible, will almost certainly linger for millennia or longer. (Some biologists have even suggested that the Anthropocene extinction will be our greatest legacy on Earth.) There are also ominous trends with respect to the growing power and accessibility of dual-use emerging technologies, as discussed more below. Suffice it to say that Easterbrook appears to suffer from the same scotoma that I claimed above has led Pinker to embrace an overly roseate picture of where we are and, more importantly, where we're going.

---

[22] Again, see James Hughes's statements above about millennialist tendencies. Easterbrook himself notes that "some amount of never-ending anxiety may be rational—keeping us on guard. … some awful collapse may happen, of course. We can't be sure the arrive op progress will remain pointed forward" (although see below).

[23] And, as I write in a footnote to the postscript of *Morality*, my ultimate hope is precisely what Lewis Mumford, the technology critic, expresses in this passage: "I would die happy if I knew that on my tombstone could be written these words, 'This man was an absolute fool. None of the disastrous things that he reluctantly predicted ever came to pass!'," although in my own case the term "predicted" should be replaced with "meticulously studies and cautiously warned about."

[24] In contrast, Amazon places it in three subcategories: "Class" (sociology), "Happiness" (mental health), and "Personal Transformation" (self-help).

Of course, people's emotions are irrelevant if the risks are real.

Wait! A moment ago Pinker was arguing that *alerting* people of certain risks was itself bad because it can lead to nihilism ("The world's future looks grim so we have to focus on looking after ourselves and those we love") and "collapse anxiety," which distorts one's perception of just how good contemporary life is. Of course, "people's emotions" are irrelevant to whether some proposition about existential risk is *true* or not, since truth is a mind-independent property. But, as we discussed above, people's emotions are extremely relevant to the paramount task of motivating people to care about these issues, voting for the right political candidates, and so on. Thus, it's precisely *when the risks are real* that understanding the emotional responses of humans to global-scale danger is the most relevant. Again, as Bostrom writes, the point isn't to wallow in gloom and doom, it's to *do something*—yet doing something requires good psycho-emotional management.

But risk assessments fall apart when they deal with highly improbable events in complex systems. Since we cannot replay history thousands of times and count the outcomes, a statement that some event will occur with a probability of .01 or .001 or .0001 or .00001 is essentially a readout of the assessor's subjective confidence. This includes mathematical analyses in which scientists plot the distribution of events in the past (like wars or cyber attacks) and show they fall into a power-law distribution, one with "fat" or "thick" tails, in which extreme events are highly improbable but not astronomically improbable.[7] The math is of little help in calibrating the risk, because the scattershot data along the tail of the distribution generally misbehave, deviating from a smooth curve and making estimation impossible. All we know is that very bad things can happen.

Which is, I would urge, sufficient for allocating modest resources to investigate "bad things," including *speculative* "bad things," to ensure that they don't occur. (Recall that Pinker acknowledges above that "the stakes, quite literally, could not be higher.") It's also worth noting that many scholars who study existential risks don't believe that the probability of a global-scale disaster is 0.01 or lower. For example, and "informal" survey of experts conducted by the Future of Humanity Institute at Oxford University yielded a median estimate for human extinction before 2100 of 19 percent. This is pretty typical of estimates by scholars with genuine expertise on the topic, as I discuss in section 1 of my paper "Facing Disaster." Indeed, just as one is more likely to die from a meteor than a lightning strike, such estimates suggest that people are far more likely to die from an existential catastrophe than either. Insofar as *genuine expertise* should be headed—and I believe that it would be an act of anti-intellectualism to ignore such experts—Pinker is wrong that we're dealing with "highly improbable events."

That takes us back to subjective readouts, which tend to be inflated by the Availability and Negativity biases and by the gravitas market (chapter 4).[8]

First, it's worth reemphasizing that existential risk scholars are, on the whole, acutely aware of the confounding effects of cognitive biases—including the availability and negativity biases. As Pinker himself has noted in the past, one of the benefits of knowing about bad modes of intellection is that this knowledge can itself serve as a bulwark against problematic thinking.[25] Second, does

[25] I cannot find the citation to this; it may have been an interview with Robert Wright.

Pinker provide any evidence to support the claim that "subjective readouts" by scientists—in our case, existential risk scholars—"tend to be inflated by the Availability and Negativity biases"?[26] The citation provided in footnote 8 is this: "Overestimating the probability of extreme risks: Pinker 2011, pp. 368-73," where "Pinker 2011" refers to *Better Angels*. I encourage readers to investigate these pages for themselves to try and identify what's relevant to the sentence above, which starts a new paragraph in *Enlightenment Now*. Indeed, there is not a *single* mention of the availability or negativity biases on these pages.[27] Pinker does mention "power law" phenomena, but only says the following:

> (a) "Terrorist attacks obey a power-law distribution, which means they are generated by mechanisms that make extreme events unlikely, but not astronomically unlikely."
> (b) "Combine exponentially growing damage with an exponentially shrinking chance of success, and you get a power law, with *its disconcertingly thick tail*. Given the presence of weapons of mass destruction in the real world, and religious fanatics willing to wreak untold damage for a higher cause, a lengthy conspiracy producing a horrendous death toll is *within the realm of thinkable probabilities*." And…
> (c) "In practice, as you get to the tail of a power-law distribution, the data points start to misbehave, scattering around the line or warping it downward to very low probabilities. The statistical spectrum of terrorist damage reminds us *not to dismiss the worst-case scenarios*, but it doesn't tell us how likely they are."[28]

So, citation 8 appears to be misplaced at the end of Pinker's sentence—and again, I would argue that "unlikely" extreme events and events with "horrendous death toll[s]" that are "within the realm of thinkable probabilities" should be enough to fund precisely those organizations, focused on existential risks, that Pinker seems to denigrate. (After all, Pinker tells us that *existential risk* is a "useless category.")

Third, and just as importantly, there are numerous cognitive biases that Pinker conspicuously ignores to make his case—biases that can lead one to *underestimate* the probability of a global catastrophe or human extinction. For example, the "observation selection effect" occurs when one's data is skewed by the fact that gathering such data is dependent upon the existence of observers like us. In other words, there are some types of catastrophes that are incompatible with the existence of certain observers, meaning that observers will always find themselves in worlds in which those types of catastrophes have not previously occurred—a fact that could lead such observers to underestimate the probability of those catastrophes. As Milan Ćirković puts the point, "people often erroneously claim that we should not worry too much about existential disasters, since

---

[26] My own view is that, given the extent to which cognitive biases are emphasized by existential risk scholars, these biases *don't* seriously affect, *if at all*, the estimates proposed by hard-headed scholars. Indeed, there may even be reason to think that the estimates of catastrophe are *lower* than they should be as scholars overcompensate for the potential distorting effects of bad patterns of thought.

[27] In fact, *Better Angels* doesn't once mention the negativity bias, and only three times, in the context of discussing the "hemoclysms" of the twentieth century, does it mention the availability bias.

[28] Italics added. To be clear, Pinker says that, with respect to nuclear terrorism in particular, "the point is not that nuclear terrorism is impossible or even astronomically unlikely. It is just that the probability assigned to it by anyone but a methodical risk analyst is likely to be too high," where "too high" here means "with certainty" or "more probable than not." But this claim is not based on "subjective readouts" being compromised by the availability and negativity bias, or the so-called "gravitas market."

none has happened in the last thousand or even million years. This fallacy needs to be dispelled." Other biases especially relevant in this context include the disjunction fallacy,[29] overconfidence, progress trap delusions, brain lag, and what Günther Anders calls "apocalyptic blindness," which

> determines a notion of time and future that renders human beings incapable of facing the possibility of a bad end to their history. The belief in progress, persistently ingrained since the Industrial Revolution [contra Pinker], causes the incapability of humans to understand that their existence is threatened, and that this could lead to the end of their history.

In my view, a fair (rather than tendentious) presentation of these issues would note—as I do in *Morality*—the various biases that can push one in either direction of over- or under-estimating the likelihood of doom.

Those who sow fear about a dreadful prophecy may be seen as serious and responsible, while those who are measured are seen as complacent and naive. Despair springs eternal. At least since the Hebrew prophets and the Book of Revelation, seers have warned their contemporaries about an imminent doomsday.

Sure, but the epistemological foundation of religious prophesies could not be more different than the epistemological foundation of scientific warnings about climate change, the Anthropocene extinction, nuclear conflict, and even engineered pandemics and misaligned superintelligence. I stress this point repeatedly in my books *Morality* and, especially, *The End*. It's important because some non-experts might inadvertently conflate these two categories simply because the message presented—"Be wary!"—sounds vaguely similar. Unfortunately, by mentioning "prophets," the "Book of Revelation," and "seers," Pinker contributes to this problem. Indeed, Pinker frequently vacillates—both above and below in his chapter—between talking about the existential warnings of scientists and the apocalyptic logorrhea of religionists. This further muddles the discussion.

Forecasts of End Times are a staple of psychics, mystics, televangelists, nut cults, founders of religions, and men pacing the sidewalk with sandwich boards saying "Repent!"[9] The storyline that climaxes in harsh payback for technological hubris is an archetype of Western fiction, including Promethean fire, Pandora's box, Icarus's flight, Faust's bargain, the Sorcerer's Apprentice, Frankenstein's monster, and, from Hollywood, more than 250 end-of-the-world flicks.[10] As the engineer Eric Zencey has observed, "There is seduction in apocalyptic thinking. If one lives in the Last Days, one's actions, one's very life, take on historical meaning and no small measure of poignance."[11]

Here footnote 11 provides the following citation: "Quoted in Ronald Bailey, 'Everybody Loves a Good Apocalypse,' *Reason*, Nov. 2015." First, Bailey is a former climate-denying libertarian who edited the 2002 book *Global Warming and Other Eco Myths: How the Environmental Movement Uses False Science to Scare Us to Death*, although he has more recently acknowledged that climate change is real but "will be solved through economic growth." Second, Bailey doesn't provide a citation to Zencey's quote,[30] mean-

[29] That is to say, human extinction is a *disjunctive phenomenon*, meaning that it can be caused by numerous distinct causal concatenations. Incidentally, the section of *Better Angels* that Pinker cites in footnote 8 goes into some detail about the conjunction fallacy, which can lead to incorrect estimates of the likelihood of different scenarios. But neither there nor in *Enlightenment Now* does Pinker mention this fallacy's evil twin, the disjunctive fallacy.

[30] Although this is not the case in a *different* article by Bailey, namely, "DiCaprio's The 11th Hour: We are the Most Important Generation in History."

ing that Pinker references a secondary source that quotes a scholar without providing a primary source citation to verify the accuracy of the quote. As it happens, I reached out to Zencey—not an "engineer," but a political economist[31]—and asked him about the quote. His response was illuminating:

> I appreciate your effort to nail down the source, and I especially appreciate the opportunity to set the record a great deal straighter than it has been. That quotation has bedeviled me. It is accurate but taken completely out of context. … You'd be doing me a service if you set the record straight.

The original source of the quote is a highly contemplative 1988 article in *The North American Review* titled "Apocalypse and Ecology."[32] In it, Zencey claims that, in response to catastrophic environmental degradation, he once anticipated a "coming transcendence of industrial society," a kind of "apocalyptic *redemption*" that would usher in an epoch marked by "the freedoms we would enjoy if only political power were decentralized and our economy given over to sustainable enterprises using renewable fuels and minimizing resources."[33] This was ultimately an *optimistic* form of apocalypticism; as Zencey puts it, "we were optimists, filled with confidence in the power of education." Yet in our exchange, Zencey is quite explicit that

> too many people use that quotation [about "apocalyptic thinking"] to make it seem that I line up *against* the idea that we face an ecological apocalypse. If on reading the essay ["Apocalypse and Ecology"] you think I wasn't sufficiently apocalyptic about the damage humans are doing to the ecosystems that are their life-support system, I can only plead that in 1988 we knew far less than we know now about how rapidly our ecological problems would foreclose upon us, and I wanted the ecology movement to reach an audience, not leave itself vulnerable to being apparently disproven in the short run.[34]

All of this being said, it's worth noting once more—at the risk of belaboring this point to death—that the sort of "apocalyptic thinking" referenced by Bailey and Pinker does not characterize the kind of concern ubiquitous among existential risk scholars.

> *Scientists and technologists are by no means immune. Remember the Y2K bug?*[12]

This is yet another suspicious citation, provided in footnote 12. It points to a 407-word-long *New York Times* article titled "Revisiting Y2K: Much Ado About Nothing?," which also includes a video by the Retro Report. Much of what Pinker says below draws directly from, and parallels, this short article and video, including the quotes of Bill Clinton and Jerry Falwell, and the reference to bolts in a bridge—almost as if Pinker is simply copying (in his own words) this information. He continues:

---

[31] With a PhD in political philosophy and history of science (personal communication).

[32] Via personal communication, Zencey adds that this article "was published in June of (let me emphasize) *1988*—thirty years ago. *Of course* I would put things differently now. The ecological apocalypse *is* playing out in human historical time. You *can* read about it in your daily paper (if you know how to decode the information and see the signs). If you are ecologically literate, you *can* see the signs of species loss and climate change (the two most obvious symptoms, but by no means the only symptoms) outside your window in the morning."

[33] The alternative is, in Zencey's view, "technological fascism."

[34] To clarify, Zencey tells me that "the one part of that that I would tweak [about the original article] is the part about how 'the scale of time in which that change will happen is most likely to be larger and longer than an individual human life.' What seemed unlikely then was, in hindsight, all too probable. As I've said elsewhere … , one of the most amazing things about industrial culture's use of fossil fuels is that it gave humans the (literal) power to bring the geologic time scale and the human time scale into congruence."

In the 1990s, as the turn of the millennium drew near, computer scientists began to warn the world of an impending catastrophe. In the early decades of computing, when information was expensive, programmers often saved a couple of bytes by represent ing a year by its last two digits. They figured that by the time the year 2000 came around and the implicit "19" was no longer valid, the programs would be long obsolete. But complicated software is replaced slowly, and many old programs were still running on institutional mainframes and embedded in chips. When 12:00 A.M.on January 1, 2000, arrived and the digits rolled over, a pro gram would think it was 1900 and would crash or go hay wire (presumably because it would divide some number by the difference between what it thought was the current year and the year 1900, namely zero, though why a program would do this was never made clear). At that moment, bank balances would be wiped out, elevators would stop between floors, incubators in maternity wards would shut off, water pumps would freeze, planes would fall from the sky, nuclear power plants would melt down, and ICBMs would be launched from their silos.

And these were the hardheaded predictions from tech-savvy authorities (such as President Bill Clinton, who warned the nation, "I want to stress the urgency of the challenge. This is not one of the summer movies where you can close your eyes during the scary part").

To be clear, Clinton may have been an "authority who was tech-savvy," but he wasn't a "savvy authority of tech," which makes it odd, in my view, to cite him in the context of *evaluating* the Y2K warnings.

Cultural pessimists saw the Y2K bug as comeuppance for enthralling our civilization to technology. Among religious thinkers, the numerological link to Christian millennialism was irresistible. The Reverend Jerry Falwell declared, "I believe that Y2K may be God's instrument to shake this nation, humble this nation, awaken this nation and from this nation start revival that spreads the face of the earth before the Rapture of the Church." A hundred billion dollars was spent worldwide on reprogramming software for Y2K Readiness, a challenge that was likened to replacing every bolt in every bridge in the world.

As a former assembly language programmer I was skeptical of the doomsday scenarios, and fortuitously I was in New Zealand, the first country to welcome the new millennium, at the fateful moment. Sure enough, at 12:00 A.M. on January 1, nothing happened (as I quickly reassured family members back home on a fully functioning telephone). The Y2K reprogrammers, like the elephant-repellent salesman, took credit for averting disaster, but many countries and small businesses had taken their chances without any Y2K preparation, and they had no problems either. Though some software needed updating (one program on my laptop displayed "January 1, 19100"), it turned out that very few programs, particularly those embedded in machines, had both contained the bug and performed furious arithmetic on the current year.

This is half true, according to the Retro Report video. Therein, the narrator states that "not everything needed to be fixed, including most embedded chips." Of course, "most" chips not needing to be fixed does not entail "very few" needing to be fixed. The latter term suggests a small minority whereas the former merely denotes a non-majority (of problematic chips).

The threat turned out to be barely more serious than the lettering on the sidewalk prophet' s sandwich board.

It is perplexing how Pinker arrives at this conclusion—especially given the citation of footnote 12. Indeed, this may be a particularly striking example of Pinker's preferential elevation of data that supports his narrative while quietly ignoring those that don't. For example, the *New York Times* article asks: "Was it all just a huge goof that faked out

even the president?," to which it responds, "no, according to this week's Retro Report video. While a lot was overblown—we spent an estimated $100 billion to combat Y2K—there were also legitimate concerns." The article adds that "among other things, Retro Report concludes that the financial markets were able to reopen quickly after 9/11 thanks to lessons learned from the work on Y2K." As one might expect, Pinker doesn't mention either (a) the legitimacy issue, or (b) the longer-term benefits of having taken the Y2K threat seriously.

The video also contains a number of statements that Pinker chooses to ignore in this discussion. For example, with respect to dodging a global disaster, Paul Saffo laments that "you never get credit for the disasters you avert, especially if you're a programmer and nobody understands what you're doing to begin with." Similarly, John Koskinen, who led Clinton's "Council on Year 2000 Conversion," asserts that "we have sort of a lack of confidence that things can get done [in America]. People did not grasp the magnitude of the *effort*. The *easier* thing to keep in your mind was, 'All that noise about it and nothing happen, it must have just been a hoax.'" The narrator of the video then notes that "the Senate's final report on Y2K found that government and industry did successfully avert a crisis at an estimated cost of 100 billion dollars," although it adds that such efforts may have overspent by 30 percent. The point is this: The article and video that Pinker cites are quite balanced, whereas Pinker's presentation of the topic is not, which indicates, to me, that Pinker has an agenda.

The Great Y2K Panic does not mean that all warnings of potential catastrophes are false alarms, but it reminds us that we are vulnerable to techno-apocalyptic delusions.

Again, the quotes above suggest that, at least from one legitimate perspective, this wasn't a "techno-apocalyptic delusion," although (i) there is ongoing debate about how necessary certain measures were (i.e., there isn't a settled view about whether Y2K slipped from *alarm* to *alarmism*[35]), and (ii) there were many conspiracy theorists, religious fanatics, gun-loving survivalists, and so on, who exploited the "dread factor" of Y2K for their own purposes.

How should we think about catastrophic threats? Let's begin with the greatest existential question of all, the fate of our species. As with the more parochial question of our fate as individuals, we assuredly have to come to terms with our mortality. Biologists joke that to a first approximation all species are extinct, since that was the fate of at least 99 percent of the species that ever lived. A typical mammalian species lasts around a million years, and it's hard to insist that Homo sapiens *will be an exception. Even if we had remained technologically humble hunter-gatherers, we would still be living in a geological shooting gallery.[13] A burst of gamma rays from a supernova or collapsed star could irradiate half the planet, brown the atmosphere, and destroy the ozone layer, allowing ultraviolet light to irradiate the other half.[14] Or the Earth's magnetic field could flip,*

[35] Indeed, as James Hughes writes, "an example of a more successful channelling of techno-apocalyptic energies into effective prophylaxis was the Millennium Bug or Y2K phenomenon. … The date 1 January 2000 was as unremarkable as all predicted millennial dates have been, but in this case, many analysts believe potential catastrophes were averted due to the proactive action from governments, corporations, and individual consumers (Special Committee on the Year 2000 Technology Problem, 2000), motivated in part by millennial anxieties. Although the necessity and economic effects of pre-Y2K investments in information technology modernization remain controversial, some subsequent economic and productivity gains were probably accrued (Kliesen 2003). Although the size and cost of the Y2K preparations may not have been optimal, the case is still one of proactive policy and technological innovation driven in part by millennial/apocalyptic anxiety."

exposing the planet to an interlude of lethal solar and cosmic radiation.

I must say, the last sentence is a bit odd given that Pinker consistently selects the most *optimistic* set of data or data interpretations to support his case (at least in this chapter), yet this is a more pessimistic account of what might happen. For example, some scientists contend that a magnetic field flip would not do much more than cause power outages, interfere with radio communications, and possibly damage satellites. But even the worst case scenario isn't that bad: it could render the ozone vulnerable to coronal mass ejections (CMEs), resulting in ozone holes that increase the rate of skin cancer.[36] The other risk that Pinker identifies—i.e., gamma-rays—is so improbable that it's hardly worth mentioning, especially in a chapter that references far more likely risk scenarios, from climate change to global pandemics.

An asteroid could slam into the Earth, flattening thousands of square miles and kicking up debris that would black out the sun and drench us with corrosive rain. Supervolcanoes or massive lava flows could choke us with ash, $C0_2$, and sulfuric acid. A black hole could wander into the solar system and pull the Earth out of its orbit or suck it into oblivion. Even if the species manages to survive for a billion more years, the Earth and solar system will not: the sun will start to use up its hydrogen, become denser and hotter, and boil away our oceans on its way to becoming a red giant.

Technology, then, is not the reason that our species must someday face the Grim Reaper. Indeed, technology is our best hope for cheating death, at least for a while.

This is utterly perplexing. First, the propositions "technology could save us" and "technology could destroy us" are *not mutually exclusive*. Which is to say, both could be true at the same time, just as the propositions "this AK-47 could save me" and "this AK-47 could kill me" could simultaneously be true. As mentioned above, one of the primary causes for existential concern about emerging technologies is that all, or nearly all, are *dually usable* for both morally good and bad ends, and this "dual-use" property appears to be an intrinsic feature of such artifacts (i.e., the "promise and peril" of advanced technology is a package deal; to neutralize either is to eliminate both.)

Second, Pinker is correct that "technology is our best hope for cheating death," but only if we're talking about natural "kill mechanisms" like asteroid/comet impacts and—perhaps—supervolcanic eruptions (although see below).[37] We don't know how to guard against gamma ray bursts, supernovae, or black holes, and while venturing into space could enable us to survive the death of our solar system, there are strong reasons for believing that space expansionism is a sour recipe for immense suffering, if not total annihilation. This being said, no one questions that using technology to eliminate risks from nature is a good thing—it is, in fact, a common refrain from existential risk scholars—but nor does anyone who seriously study the future of humanity believe that the greatest dangers to our collective survival are

[36] For example, in a book chapter titled "Influence of Supernovae, Gamma-Ray Bursts, Solar Flares, and Cosmic Rays on the Terrestrial Environment," Arnon Dar notes that "past reversals [of Earth's magnetic field] were not associated with any major extinction according to the fossil record, and thus [another reversal is] not likely to affect humanity in a catastrophic way."

[37] To be precise, certain advanced technologies could "save us" from the dangers posed by certain other advanced technologies. This is true. The point being made, though, concerns the overall *net* existential risk that humanity will be exposed to: in the case of natural risks, technology is unambiguously positive in its net effects; in the case of anthropogenic risks, it's unclear whether humanity will come out ahead or not—technology is both the "summoned demon" and our potential savior.

natural phenomena. Rather, it's large-scale human activity, dual-use emerging technologies, and value-misaligned superintelligence that, without question, constitute the most urgent global-scale hazards.

As long as we are entertaining hypothetical disasters far in the future, we must also ponder hypothetical advances that would allow us to survive them, such as growing food under lights powered with nuclear fusion, or synthesizing it in industrial plants like biofuel.[15]

For the sake of clarity, the citation provided in footnote 15 is to *Feeding Everyone No Matter What: Managing Food Security After Global Catastrophe*, by David Denkenberger and Joshua Pearce. Although the structure of Pinker's sentence might imply that Denkenberger and Pearce endorse growing *food* under lights that are powered by nuclear fusion, this is not the case. Rather, Denkenberger and Pearce claim that growing food under such lights would be too inefficient and pricey; a configuration of this sort would only be practicable for "high value" commodities like *drugs*. This citation is also rather odd (and misleading) because Denkenberger takes the topic of existential risks very seriously—indeed, his organization ALLFED notes that "there is an estimated 10 percent chance of a complete loss of food production capability this century." Furthermore, in response to my criticisms of Pinker's understanding of technological risk, Denkenberger tells me, "I agree that [Pinker, in this chapter] is being too dismissive of the risks from technology." Note once more the causal link between worrying about existential risks and working to mitigate them. In a sense, "collapse anxiety" is precisely what motivates Denkenberger's important research.

Even technologies of the not-so-distant future could save our skin. It's technically feasible to track the trajectories of asteroids and other "extinction-class near-Earth objects," spot the ones that are on a collision course with the Earth, and nudge them off course before they send us the way of the dinosaurs.[16] *NASA has also figured out a way to pump water at high pressure into a supervolcano and extract the heat for geothermal energy, cooling the magma enough that it would never blow its top.*[17]

The term "figured out" is rather misleading. What NASA did was propose a "thought experiment" for how a supervolcano might be defused. But, importantly, this is not practically feasible right now and won't be for the foreseeable future (e.g., we are barely able to dig sufficiently deep into the ground); doing this would require truly *huge* amounts of water; and there remains a chance that pumping water into a supervolcano could actually *trigger* a supereruption, which could bring about a catastrophic volcanic winter.[38]

Our ancestors were powerless to stop these lethal menaces, so in that sense technology has not made this a uniquely dangerous era in the history of our species but a uniquely safe one.

Once again, this is true *if we're talking about natural risks*. With respect to the growing swarm of anthropogenic risks discussed above, *just the opposite is the case*.

For this reason, the techno-apocalyptic claim that ours is the first civilization that can destroy itself is misconceived.

Which "techno-apocalypticists" make this claim? Pinker doesn't provide a citation,

[38] For a comprehensive scholarly survey of strategies for supervolcanic risk mitigation, see "Interventions that May Prevent or Mollify Supervolcanic Eruptions" by David Denkenberger and Robert Blair. An accessible article about the NASA proposal, written by a volcanologist at Denison University, can be found here.

which is unsurprising, since no reasonable person would claim that "ours" (whatever that means exactly) is the very first civilization in human history that's capable of destroying itself. After all, asserting this would entail denying that the Mayan, Roman, and Easter Island civilizations ever collapsed, which is patently absurd. If by "[our] civilization" one means the McLuhanian "global village" in which all 7.6 billion contemporary humans live, then yes, since 1945—*but not before*—we have indeed possessed the historically unique ability to wreak planetary-scale harm that could bring about the implosion of all major human societies around the world.

As Ozymandias reminds the traveler in Percy Bysshe Shelley's poem, most of the civilizations that have ever existed have been destroyed. Conventional history blames the destruction on external events like plagues, conquests, earthquakes, or weather. But David Deutsch points out that those civilizations could have thwarted the fatal blows had they had better agricultural, medical, or military technology:

> *Before our ancestors learned how to make fire artificially (and many times since then too), people must have died of exposure literally on top of the means of making the fires that would have saved their lives, because they did not know how. In a parochial sense, the weather killed them; but the deeper explanation is lack of knowledge. Many of the hundreds of millions of victims of cholera throughout history must have died within sight of the hearths that could have boiled their drinking water and saved their lives; but, again, they did not know that.*
>
> *Quite generally, the distinction between a "natural" disaster and one brought about by ignorance is parochial. Prior to every natural disaster that people once used to think of as "just happening," or being ordained by gods, we now see many options that the people affected failed to take—or, rather, to create. And all those options add up to the overarching option that they failed to create, namely that of forming a scientific and technological civilization like ours. Traditions of criticism. An Enlightenment.*[18]

It's not entirely clear how any of these points are relevant: yes, *obviously*, Deutsch is right that ignorance will always leave whole populations vulnerable to natural disasters like epidemics, but this doesn't change the fact that undergirds many dire warnings about the threat of advanced technology, namely, that knowledge—like its "phenotypic expression," technology—can *itself* be dually usable. As such, it can be use to defend against deadly epidemics *and* synthesize unnaturally lethal, incurable, and contagious designer pathogens (with long incubation periods like that of HIV). There is no tension between these claims. In fact, Deutsch himself is rather less blithe about the various global risks facing humanity than Pinker is, a fact that Pinker decides not to mention. For example, after noting that technology will effectively eliminate some catastrophic dangers, Deutsch observes that

> problems are inevitable. We shall always be faced with the problem of how to plan for an unknowable future. *We shall never be able to afford to sit back and hope for the best.* Even if our civilization moves out into space in order to hedge its bets, as Rees and Hawking both rightly advise, a gamma-ray burst in our galactic vicinity would still wipe us all out. Such an event is thousands of times rarer than an asteroid collision [about which, Deutsch states earlier in the chapter, the average person has a "far larger chance of dying" from "than in an aeroplane crash"], but when it does finally happen we shall have no defence against it *without a great deal more scientific*

> *knowledge* and an enormous increase in our wealth.[39]

Of course, the acquisition of "a great deal more scientific knowledge" about *risks in general* is precisely what existential risk studies hopes to acquire! Deutsch then continues:

> But first we shall have to survive the next ice age; and, before that, other dangerous climate change (both spontaneous and human-caused), and weapons of mass destruction and pandemics and all the countless unforeseen dangers that are going to beset us. Our political institutions, ways of life, personal aspirations and morality are all forms or embodiments of knowledge, and all will have to be improved if civilization—and the Enlightenment in particular—is to survive every one of the risks that Rees describes [in *Our Final Hour*] and presumably many others of which we have no inkling.

So, Deutsch appears to have a more cautiously "optimistic" view than Pinker—indeed, a view that very much aligns with the general attitude that motivates scholarly work on existential hazards (although he nonetheless claims, contra Rees and Hawking, that the present moment is not uniquely dangerous[40]).

Prominent among the existential risks that supposedly threaten the future of humanity is a 21st-century version of the Y2K bug. This is the danger that we will be subjugated, intentionally or accidentally, by artificial intelligence (AI), a disaster sometimes called the Robopocalypse and commonly illustrated with stills from the Terminator movies.

I would argue that this egregiously misrepresents the threat of "superintelligence," i.e., a hardware-based general intelligence algorithm that can, by definition, outperform even the best humans in every cognitive domain. First of all, there are hardly any analogical points-of-contact between Y2K and superintelligence; we will return to this momentarily. Second, Pinker is wrong that this threat is "sometimes called the Robopocalypse"—that is, if he's talking about the community of scientists and philosophers who are focused on creating "value-aligned" superintelligence. Perhaps there are some *journalists* who would use this term—although virtually all the top Google results link to a novel and movie by the same name, in addition to some articles written by Pinker—but it has literally no traction among AI experts. The reason is obvious: no one is worried about an army of *robots* bringing about the *apocalypse*, which is what the term clearly implies. Third, it's true that many popular media articles about superintelligence feature an image of the Terminator. But this is, one should note, a perpetual source of trichotillomania-inducing annoyance among scholars of AI safety. As Eliezer Yudkowsky—one of the leading figures in the community—states in a podcast interview, "I think that at this point all of us on all sides of this issue are annoyed with the journalists who insist on putting a picture of the Terminator on every single article they publish of

[39] Italics added to emphasize that this is, once again, the fundamental aim of existential risk studies.
[40] Personal communication. Of course, I believe that Deutsch is generally wrong about this. Or, to be more precise, it may be that the present moment—i.e., the past and next few decades—could be the best time for any human to be alive, *ever*. This view essentially fuses the two broadest trends here discussed, namely, (i) human well-being has improved over the centuries in numerous important respects, and (ii) the overall *global risk potential* of the world has steadily risen since 1945 and will almost certainly continue to rise, perhaps meteorically, later this century, as climate change and biodiversity loss and increasingly powerful emerging technologies place civilization under unprecedented survival pressures. Thus, we may find ourselves at the *most desirable intersection of these two trends*, where both the past and future are worse in some critical way than the unprecedentedly good *now*.

this topic."[41] So, for those readers who care about being *informed* about this issue, when you think about superintelligence, *don't* think about the Y2K scare, Robopocalypse, or Terminator movies—these are nothing more than *red herrings*.

As with Y2K, some smart people take it seriously. Elon Musk, whose company makes artificially intelligent self-driving cars, called the technology "more dangerous than nukes." Stephen Hawking, speaking through his artificially intelligent synthesizer, warned that it could "spell the end of the human race."[19] But among the smart people who aren't losing sleep are most experts in artificial intelligence and most experts in human intelligence.[20]

First, it's completely irrelevant that Musk's company and Hawking's synthesizer uses AI. This strikes me as similar to someone saying, with an annoying dose of sardonic snark, that "it's really ironic that Joe McNormal works at the nuclear power plant yet *also* claims that nuclear missiles are dangerous. Pfft!" Second, it's worth noting that footnote 20 cites a survey conducted by Vincent Müller and Nick Bostrom in 2014. Yet Pinker fails to mention that this survey also yields a median probability estimate for "human-level machine intelligence" (HLMI) of 10 percent by 2022, 50 percent by 2040, and 90 percent by 2075. It also reports a median estimate of 75 percent that superintelligence will follow within 30 years of the first HLMI. Thus, a child born today has a fairly good chance of living to see the first human-level AI and perhaps even the first superintelligence. Indeed, if a HLMI is created using what David Chalmers calls an "extendible method," which would enable the HLMI to improve itself by modifying its own code,[42] the result could be machine superintelligence within a relatively short period of time. (We will discuss this more below.) On the other hand, there are reasons for thinking that a HLMI might automatically *be* superintelligent, given its substrate, the software that it would have available to it, and so on. As Rob Miles explains in an educational video titled "What can AGI do? I/O and Speed," a HLMI that's running on the silicon (or carbon nanotube) hardware could process information roughly a *million times faster* than the human brain—meaning that if it takes the average student about eight years to earn a PhD, a computer-run HLMI could accomplish this in less than five minutes. In this sense, the HLMI would constitute a *quantitative* superintelligence. But it could also have direct access to a broad range of "narrow" AI systems, such as calculators, that already vastly surpass the best humans in circumscribed cognitive domains like mathematics. This would further enable the HLMI to effectively exceed the human level of cognitive performance.

Finally, it's crucial to note that people with expertise on the *technical issues* of building AI systems, training neural nets, and so on, won't automatically have expertise on the *philosophical issues* that lead theorists like Nick Bostrom and Stuart Russell to worry about value-misaligned superintelligence. The difference is, roughly speaking, one between an engineer who designs the landing gear or hydraulic system of airplanes and an ethicist who ponders the moral implications of flying airplanes into skyscrapers.[43] Inci-

[41] Again, my editor at *Salon* was guilty of doing just this, much to my dismay!
[42] Note here that self-improvement, or "cognitive augmentation," constitutes an *instrumental value* that virtually any competent AI system would pursue, whatever its final goals happen to be.
[43] As I will hopefully make clear below, this analogy is misleading because the experts are not worried about "malevolent" or "evil" superintelligence (with one small exception). Thus, a better analogy would involve an ethicist who ponders the moral implications of planes flying themselves into buildings that happen to get in the way of reaching their destinations.

dentally, in footnote 20, Pinker writes that "AI experts who are publicly skeptical" that "high-level AI pose[s] the threat of 'an existential catastrophe'" include "Kevin Kelly" and "Stuart Russell." These are peculiar names to include here because (a) Kelly may be a "technology expert," as Pinker elsewhere describes him, but he's not an expert of AI, and (b) Russell is one of the most prominent computer scientists who's deeply concerned about the existential risks associated with superintelligence![44] Yet again, one is left baffled by Pinker's citational practice—was it an accident that Pinker classifies Russell as an AI "skeptic," or is Pinker unaware of Russell's actual position on the issue? If it's the latter, then one should question whether Pinker is sufficiently knowledgeable about AI to confidently pontificate, in an influential public forum, about how misguided AI safety efforts are.

The Robopocalypse is based on a muzzy conception of intelligence that owes more to the Great Chain of Being and a Nietzschean will to power than to a modern scientific understanding. [21]

As we will see, no one in the AI safety research community is worried that a superintelligence will destroy humanity because of its Nietzschean "will to power." Furthermore, the "Great Chain of Being" metaphor suggests a linear hierarchy of "higher" and "lower" levels of intelligence, which is not how, in the present context, one should think about intelligence. As Yudkowsky writes in a book chapter on the topic (which Pinker elsewhere cites),

> the term "Artificial Intelligence" refers to a vastly greater *space of possibilities* than does the term "Homo sapiens." When we talk about "AIs" we are really talking about *minds-in-general*, or optimization processes in general. Imagine a map of mind design space. In one corner, a tiny little circle contains all humans; within a larger tiny circle containing all biological life; and all the rest of the huge map is the *space of minds-in-general*. The entire map floats in a still vaster space, the *space of optimization processes*.

In other words, we should conceptualize *qualitatively* different intelligences[45] or, more generally, optimization processes, in terms of differently-sized-and-shaped regions within a potentially vast perimeter that demarcates all possible minds/optimization processes. Rather than a *linear hierarchy* metaphor, AI safety experts thus advocate a *cartographic* metaphor in which different "cognitive spaces" (as I prefer to call them) can subsume, overlap with, or not overlap with other cognitive spaces. This is a quite different model than what Pinker identifies—for the purpose of knocking it down.

In this conception, intelligence is an all-powerful, wish-granting potion that agents possess in different amounts. Humans have more of it than animals, and an artificially intelligent computer or robot of the future ("an AI ," in the new count-noun usage) will have more of it than humans. Since we humans have used our moderate

---

[44] Indeed, the article that Pinker cites here is "Will They Make Us Better People?," published on *The Edge.org*. In it, Russell is quite explicit about the *existential importance* of aligning the values of AI with our human values. "As Steve Omohundro, Nick Bostrom, and others have explained," Russell writes, referring to the instrumental convergence thesis, "the combination of value misalignment with increasingly capable decision-making systems can lead to problems—perhaps even species-ending problems if the machines are more capable than humans." He concludes the article with this rumination: "Instead of pure intelligence, we need to build intelligence that is *provably* aligned with human values. This turns moral philosophy into a key industry sector. The output could be quite instructive for the human race as well as for the robots."

[45] Such as a directly programmed or neuromorphic HLMI or super-HLMI. In contrast, a quantitative intelligence would involve the *same cognitive architecture* that's merely able to process information faster.

endowment to domesticate or exterminate less well-endowed animals (and since technologically advanced societies have enslaved or annihilated technologically primitive ones), it follows that a supersmart Al would do the same to us.

This line of argumentation is Pinker's —and *Pinker's only*.[46] That is to say, no one in the AI safety community fears that because (P1) we have done X in the past, and (P2) doing X has depended upon our relatively high levels of intelligence; therefore (C) a superintelligence will also do X to us, where "X" stands for "(domesticating or) exterminating organisms with lower levels of intelligence." This simply does not track any of the serious arguments outlined by scholars in the technical literature.

Since an AI will think millions of times faster than we do, and use its superintelligence to recursively improve its superintelligence (a scenario sometimes called "foom," after the comic-book sound effect), from the instant it is turned on we will be powerless to stop it.[22] *But the scenario makes about as much sense as the worry that since jet planes have surpassed the flying ability of eagles, someday they will swoop out of the sky and seize our cattle.*

It's hard to see any actual parallels between these two scenarios. Pinker seems to be suggesting that if planes fly faster than eagles, they might swoop down and seize larger prey than mice or rabbits, and that this is essentially what those worried about superintelligence are claiming will happen to humanity once machines surpass our intellectual abilities. If so, the confusion is denser than a neutron star. The first thing to recognize is that intelligence, especially *general intelligence*, enables *self-improving positive feedback loops*. This is why IJ Good famously describes the first superintelligence as the last invention that humanity will ever need to make:

> let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any [human] however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an "intelligence explosion," and the intelligence of [humans] would be left far behind.

This sort of self-improving positive feedback mechanism simply isn't applicable to the case of fast-moving machines that can fly. Or, as Yudkowsky writes in this general context, "artificial Intelligence is not something you can casually mix into a *lumpenfuturistic* scenario of skyscrapers and flying cars and nanotechnological red blood cells that let you hold your breath for eight hours. *Sufficiently tall skyscrapers don't potentially start doing their own engineering*."[47]

The first fallacy is a confusion of intelligence with motivation—of beliefs with desires, inferences with goals, thinking with wanting.

This is a good candidate, in my view, for being among the most perplexing statements of the whole chapter. Consider that one of the *central ideas of AI safety research* is the "orthogonality thesis," which states that "intelligence and final goals are orthogonal: more or less any level of intelligence could in principle be combined with more or less any final goal." The reason is that "intelligence," as Pinker would concur, is just a measure of one's ability to secure effective means to achieve one's ends—this is the standard definition in the cognitive sciences and philosophy of mind, and it says nothing whatsoever

[46] If one were feeling rambunctious, perhaps one might call this "the Pinkerian fallacy."
[47] Italics added.

about *what* one's ends are. Thus, a mass murderer who kills his victims with great efficiency could be described as "intelligent" no less than a philanthropic altruist who discovers a highly effective way to save the lives of millions of people.[48] It follows that there is no contradiction between a machine whose (a) problem-solving capacities exceed that of all possible humans by orders of magnitude, and (b) ultimate goals are no less mundane or bizarre, from our human perspective, than calculating as many digits of pi as possible, rewriting Hamlet over and over again until the heat death of the universe, collecting as many stamps as possible for one year,[49] or "worshipping" some AI god that its world-model identifies as having created the universe. As alluded to above, this is a crucial premise in the argument *for* taking the superintelligence threat seriously—yet Pinker seems to believe that it constitutes a problem for the proposition, "superintelligence poses an existential risk."

Even if we did invent superhumanly intelligent robots, why would they want to enslave their masters or take over the world? Intelligence is the ability to deploy novel means to attain a goal. But the goals are extraneous to the intelligence: being smart is not the same as wanting something. It just so happens that the intelligence in one system, Homo sapiens*, is a product of Darwinian natural selection, an inherently competitive process. In the brains of that species, reasoning comes bundled (to varying degrees in different specimens) with goals such as dominating rivals and amassing resources. But it's a mistake to confuse a circuit in the limbic brain of a certain species of primate with the very nature of intelligence. An artificially intelligent system that was designed rather than evolved could just as easily think like shmoos, the blobby altruists in Al Capp's comic strip Li'l Abner, who deploy their considerable ingenuity to barbecue themselves for the benefit of human eaters. There is no law of complex systems that says that intelligent agents must turn into ruthless conquistadors. Indeed, we know of one highly advanced form of intelligence that evolved without this defect. They're called women.*

It is difficult to know where to begin with this. Once again, the concern isn't that a superintelligence will possess some alpha-male or machismo "will to power," or that it will desire to "enslave their masters or take over the world."[50] The key idea here concerns what's called the "instrumental convergence thesis." This states that there exists a small catalogue of *instrumental goals* that all intelligent agents will predictably pursue to achieve their final goals independent of what those final goals happen to be. For example, if I have a life-goal of manufacturing 1 million paperclips—say, this is my ultimate passion in life—then it is in my interest to ensure the following things: (i) make sure I don't die, because if I die, I won't be able to achieve my life-goal; (ii) make sure that no one alters my life-goal, because if this occurs and I abandon my passion, then I won't succeed at creating 1 million paperclips; (iii) improve my intelligence, because the smarter I am, the better strategies I'll be able to devise for manufacturing paperclips; (iv) learn as much as I can about physics, chemistry, and so on,

[48] Here we might here distinguish between *instrumental rationality* and *value rationality*. The first term roughly corresponds to the definition of "intelligence" above, whereas the latter measures the extent to which one's final goals are consciously and deliberately chosen for moral, aesthetic, epistemic, and so on, reasons—where the first two are entirely human-relative.

[49] The hyperlink here connects to an excellent Computerphile video featuring Rob Miles.

[50] It is somewhat ironic that Pinker has repeatedly put forth a gender-based argument against superintelligence worries given that he's no fan of gender studies, which is precisely the field that would reject superintelligence worries as stemming from alpha-male tendencies (that is, if those scholars were sufficiently unfamiliar with the actual arguments of the AI safety literature).

because a better mental model of the universe will help me devise better ways to manufacture paperclips; (v) invent better technologies—not just to increase the efficiency of paperclip manufacturing, but to augment my cognitive abilities and knowledge about the universe; and (vi) acquire as many resources as possible, since doing so would enhance my ability to accomplish all of the above instrumental goals.[51]

The example of making paperclips is obviously silly, but here's the point: If my one and only final goal were to cure all human diseases, end world hunger, convince humanity to adopt a world government, prove the Riemann hypothesis, build a Dyson swarm around the sun, colonize the Milky Way galaxy, *and so on*, it would *also* be in my interest to pursue (i) through (vi). And here's the catch: if we aren't really careful about which final goal(s) we give a machine that ends up being more intelligent than us in every domain, the instrumental goal of "open-ended resource acquisition" specified by (vi) would almost certainly result in our annihilation; that is, *not* because the superintelligence is "evil" or "malevolent" or "malicious" or "misanthropic" but simply because our bodies are made of conveniently located atoms that it could use for something else. The now-hackneyed analogy that scholars use to describe this goes as follows: humans and ants have different value systems. Humans want to (in this example) build new suburban neighborhoods, whereas ants want to construct underground colonies. Yet humans happen to be quite a bit more intelligent than ants. As a result, humans cut down trees, dig into the ground, and construct houses without much resistance from the poor ants who are killed as a result, in a mass genocide that we humans hardly notice, much less care about. The result is the destruction of entire ant colonies not because humans hate ants, are evil, and so on, but simply because (a) we share different value systems, and (b) our intelligence enables us to overcome any rebellion the ants might have organized. Now change some details and replace "ants" with "humans" and "humans" with "superintelligence." *This* is the worry that Pinker mischaracterizes and, consequently, fails to address.

But there's a double catch here: to prevent a bad outcome from occurring with superintelligence, humanity will need to overcome two challenges. First, there's the *philosophical* problem of figuring out which set of values we would like the superintelligence to pursue, and second, there's the *technical* problem of figuring out how to load these values, once we agree on what they should be, into the AI. Both turn out to be exceptionally difficult—and this is why a growing number of scientists and philosophers are concerned that superintelligence could constitute the greatest (known) threat to the long-term survival of our species.

The second fallacy is to think of intelligence as a boundless continuum of potency, a miraculous elixir with the power to solve any problem, attain any goal.[23]

Another straw man clubbed to death by the bludgeon of mischaracterization. Let's be precise here: in his canonical book on the topic, Bostrom "tentatively" defines the term

[51] Note: the usual example is of a "paperclip maximizer" whose aim is to produce *as many paperclips as possible*. But the example works just as well if one's goal is to manufacture a *finite* quantity: for example, insofar as one is a good Bayesian reasoner, one will never assign a probability of 1 to having accomplished one's goal, because there will always remain uncertainty about whether one might have miscounted and thus fallen short of or overshot the goal, and so on. Thus, a superintelligence with this *limited* final goal would also pursue open-ended resource acquisition to count and recount the total number of paperclips that it's manufactured, with each recount increasing the likelihood that it has achieved its goal without ever reaching a state of certitude.

"superintelligence" as "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest." Note that *exceeding* the cognitive performance of humans does not imply, at all, that intelligence constitutes a "boundless continuum of potency" or the capacity "to solve any problem, attain any goal." The AI safety community isn't worried about a machine that can solve *any* problem or attain *any* goal; rather, it's worried about an algorithm that can make, to paraphrase Stuart Russell, *higher-quality decisions* than any possible organism with a human-type genome.[52] Put differently, the existential dangers associated with the instrumental convergence thesis don't require an omniscient machine; they only require a machine that is *relatively superior* to humans in "all domains of interest."

The fallacy leads to nonsensical questions like when an AI will "exceed human-level intelligence," and to the image of an ultimate "Artificial General Intelligence" (AGI) with God like omniscience and omnipotence. Intelligence is a contraption of gadgets: software modules that acquire, or are programmed with, knowledge of how to pursue various goals in various domains.[24] People are equipped to find food, win friends and influence people, charm prospective mates, bring up children, move around in the world, and pursue other human obsessions and pastimes. Computers may be programmed to take on some of these problems (like recognizing faces), not to bother with others (like charming mates), and to take on still other problems that humans can't solve (like simulating the climate or sorting millions of accounting records). The problems are different, and the kinds of knowledge needed to solve them are different. Unlike Laplace's demon, the mythical being that knows the location and momentum of every particle in the universe and feeds them into equations for physical laws to calculate the state of everything at any time in the future, a real-life knower has to acquire information about the messy world of objects and people by engaging with it one domain at a time. Understanding does not obey Moore's Law: knowledge is acquired by formulating explanations and testing them against reality, not by running an algorithm faster and faster.[25] Devouring the information on the Internet will not confer omniscience either: big data is still finite data, and the universe of knowledge is infinite.

To be clear, none of the serious arguments for worrying about superintelligence rely, in any way, on any "laws" of exponential development, such as Moore's law.

For these reasons, many AI researchers are annoyed by the latest round of hype (the perennial bane of AI) which has misled observers into thinking that Artificial General Intelligence is just around the corner.[26]

Among the articles that Pinker cites in footnote 26 is a long thread on *The Edge.org* that includes a plethora of responses to a video conversation with Jaron Lanier (an AI skeptic). Among the respondents are Kevin Kelly (mentioned above), George Dyson, and Lee Smolin. This thread was posted roughly five months after Bostrom's book *Superintelligence* was published, and it contains numerous statements of skepticism about the plausibility of and/or hazards posed by superintelligent machines. Yet many of these skeptics don't appear—*at all*, in some cases—to understand the actual worries articulated by AI safety experts and delineated with nuance and sophistication in Bostrom's book. In fact, toward the end of the thread, Russell notes that

[52] Nor are any of the worries outlined by the AI safety community in any way *predicated* upon intelligence having the properties that Pinker species.

> there have been many unconvincing arguments [that AI poses a threat]—especially those involving blunt applications of Moore's law or the spontaneous emergence of consciousness and evil intent. Many of the contributors to this conversation seem to be responding to those arguments and ignoring the more substantial arguments proposed by Omohundro, Bostrom, and others.

For example, the roboticist Rodney Brooks, who Pinker cites as an authority on the issue, writes the following:

> (i) "Just how open the question of time scale for when we will have human level AI is highlighted by a recent report by Stuart Armstrong and Kaj Sotala, of the Machine Intelligence Research Institute [MIRI], an organization that itself has researchers worrying about evil AI."
> (ii) "I think it is a mistake to be worrying about us developing malevolent AI anytime in the next few hundred years."
> (iii) "Worrying about AI that will be intentionally evil to us is pure fear mongering."

There are myriad problems with these claims. First, MIRI is worried about value-misaligned superintelligence—because of the orthogonality, instrumental convergence, and other theses discussed in section 2.3 of this paper—*not* about "evil," "malevolent," or "intentionally evil" AI. This is, indeed, one of the top "Myths about AI Safety" that Max Tegmark identifies in a Future of Life Institute article, and it indicates, quite unambiguously, that Brooks is deeply unfamiliar with the relevant body of research. Along these lines, Michael Shermer (also disgraced, for the same ignominious reason that Krauss is) shoots but misses the target when he declares that "the latest round of handwringing over the potential for computers, machines, or robots to turn evil overlooks the fundamental difference between artificial intelligence (AI) and natural intelligence (NI)." I have written a detailed response to Shermer's view of superintelligence in an article for the *Bulletin of the Atomic Scientists*, so readers should go there for more.

Finally, Pinker claims that the latest hype-round "has misled observers into thinking that Artificial General Intelligence is just around the corner." Once more, I have no idea who Pinker is referring to by "observers"—and this is not from ignorance of the field. Bostrom, Yudkowsky, Russell, and every other respectable scholar are, indeed, *explicit* that the arguments for worrying about a superintelligence takeover have nothing whatsoever to do with the timeline of its development.[53] As Yudkowsky points out, just as it took millennia for physicists to reach the point of initiating an explosive nuclear chain reaction, so too could the emergence of a superintelligent AI occur on a completely different timescale than the incremental research that led to its "nucleation." In his words: "The first moral is that confusing the speed of *AI research* with the speed of *a real AI* once built is like confusing the speed of physics research with the speed of nuclear reactions." So, no one dogmatically maintains that AGI or HLMI is "just around the corner," although this *may* be the case; judicious minds are simply agnostic about the issue. This being said, we should recall the previous survey that Pinker himself cites, and which suggests a nearly 100 percent chance of machine general intel-

[53] To be clear, the timeline does matter insofar as it will determine the amount of time we have to solve the value-alignment problem. If superintelligence will arrive circa 2093 but humanity needs another 100 years to solve it, then we are in big trouble. This being said, the phenomena of orthogonality, instrumental convergence, etc.—the philosophical heart of the alleged superintelligence threat—do not depend on any near-term breakthroughs in AI development.

ligence by the end of this century. Depending on how much one "zooms-out" to geological time, perhaps there is a *sense* in which superintelligence is right over the horizon. But this qualification is important.

* * *

It's here that, somewhat arbitrarily, I would like to interrupt this critique of Pinker's chapter on existential threats. There is *a lot of chapter left*—indeed, we have only made it perhaps one-fourth or one-third through the entire document. Yet this may be excusable insofar as I believe that the criticisms above are *sufficient* for readers to grasp the ubiquity and severity of problems with Pinker's general approach and more specific theses. Indeed, the remaining paragraphs are, I would argue, no less permeated by cherry-picked data, questionable citations, and other manifestations of authorial malpractice. (After a cursory glance over Pinker's footnotes for the rest of the chapter, I have found several other quotes and citations that appear to be, I say quite seriously, taken out of context.) To see these additional problems, readers have three options: (i) peruse the scholarly literature on existential risks to gain some degree of expertise and then use this to read Pinker through an informed exegetical lens, (ii) consult individuals who do have some expertise on existential risks about particular questions and curiosities; or (iii) contact me and ask me to produce a longer document—as mentioned at the beginning of this paper, I would consider doing so if readers found this sort of paragraph-by-paragraph dissection useful.

Before concluding this paper, though, I would like to focus on one last passage that is, in my view, among the most patently false in the entire chapter. Let's begin by quoting Pinker at length:

> *… Has technological progress ironically left the world newly fragile?*
>
> *No one can know with certainty, but when we replace worst-case dread with calmer consideration, the gloom starts to lift. Let's start with the historical sweep: whether mass destruction by an individual is the natural outcome of the process set in motion by the Scientific Revolution and the Enlightenment. According to this narrative, technology allows people to accomplish more and more with less and less, so given enough time, it will allow one individual to do anything—and given human nature, that means destroy everything.*
>
> *But Kevin Kelly, the founding editor of* Wired *magazine and author of* What Technology Wants*, argues that this is in fact not the way technology progresses.[40] Kelly was the co-organizer (with Stewart Brand) of the first Hackers' Conference in 1984, and since that time he has repeatedly been told that any day now technology will outrun humans' ability to domesticate it. Yet despite the massive expansion of technology in those decades (including the invention of the Internet), that has not happened. Kelly suggests that there is a reason: "The more powerful technologies become, the more socially embedded they become." Cutting-edge technology requires a network of cooperators who are connected to still wider social net works, many of them committed to keeping people safe from technology and from each other. (As we saw in chapter 12, technologies get safer over time.) This undermines the Hollywood cliche of the solitary evil genius who commands a high-tech lair in which the technology miraculously works by itself. Kelly suggests that because of the social embeddedness of technology, the destructive power of a solitary individual has in fact not increased over time:*
>
> > *The more sophisticated and powerful a technology, the more people are needed to weaponize it. And the more people needed to weaponize it, the more societal controls work to defuse, or soften, or prevent harm from happening. I add one additional thought.*

> *Even if you had a budget to hire a team of scientists whose job it was to develop a species extinguishing bioweapon, or to take down the internet to zero, you probably still couldn't do it. That's because hundreds of thousands of man-years of effort have gone into preventing this from happening, in the case of the internet, and millions of years of evolutionary effort to prevent species death, in the case of biology. It is extremely hard to do, and the smaller the rogue team, the harder. The larger the team, the more societal influences.*[41]

Pinker then proceeds to argue, among other things, that the number of sufficiently competent malicious agents willing to cause global-scale harm "may not be zero, but it surely isn't high." First, I have published several papers on this exact topic, and while I would agree that the *percentage* of "omnicidal agents" is small, my research leads me to believe that the *absolute number* is not negligible. Readers are encouraged to see this paper for a theoretical overview of the topic and this paper for a concrete list of actual "agents of doom" who would almost certainly have brought about an existential catastrophe *if only* the technological means had been available. These are among the first papers ever published on the topic of "agential risks"—and they were published after Pinker's book—so it's unfair to complain that Pinker doesn't cite them (although, then again, he generally doesn't cite *anyone* working on existential risk issues).

Second, the citation provided in footnote 41 is "personal communication." In my opinion, as someone who studies this issue, the first two sentences of Kelly's response are perhaps the most flagrantly and outrageously wrong that I've ever stumbled upon. It is, indeed, *simply not true* that more sophisticated and powerful technologies require more people to weaponize it. I am not here regurgitating some doom-mongering "narrative"—I'm stating an empirical fact. Consider just a few recent phenomena (some of this is excerpted from *The End* and *Morality*):

> (1) People can now print guns that fire real bullets from 3D printers. The first of its kind was dubbed the "Liberator" by its libertarian creator, and while the US government removed downloadable designs for the gun within two days of it being released online (for free), not only had the designs been downloaded more than 100,000 times, but they remain available on The Pirate Bay.
> (2) A single group of Australian scientists who were trying to create a mouse contraceptive by modifying the mousepox virus inadvertently made it 100 percent lethal in all mice, including those that had previously been vaccinated against it and those with a natural immunity. This confirms that making an already-virulent virus *even more* virulent is possible with genetic engineering techniques, and it has bioterrorism implications because the smallpox and mousepox viruses are quite similar.
> (3) A year later, in 2002, scientists at Stony Brook University synthesized "a live polio virus from chemicals and publicly available genetic information." Specifically, they created "the virus using its genome sequence, which is available on the Internet, as their blueprint and genetic material from one of the many companies that sell made-to-order DNA." This project was, in fact, funded by the Pentagon, and the point was "to send a warning that terrorists might be able to make bio- logical weapons without obtaining a natural virus." As the study's lead scientist chillingly put it, "you no longer need

the real thing in order to make the virus and propagate it."[54]

(4) The biohacking community continues to make setting up a laboratory in one's basement or garage increasingly affordable; indeed, readers can buy a "DIY Bacterial Gene Engineering CRISPR Kit" online for a mere $159.00.

(5) The *explicit aim* of synthetic biology is to create standardized, modular genetic elements that users with little expertise can effectively manipulate. The result is a "de-skilling" trend that has "the potential to … decrease the skill gradient separating elite practitioners from non-experts." Consequently, this domain presents special difficulties for regulators. In the words of Ali Nouri and Christopher Chyba, "biological weapons proliferation poses challenges more similar to those presented by cyber attacks or cyber-terrorism than to those due to nuclear or chemical weapons. … Internet technology is so widely available that only a remarkably invasive inspection regime could possibly monitor it."[55]

(6) The 9/11 terrorist attack—which Pinker mentions—was perpetrated by only 19 radical Islamists associated with al-Qaeda, a terrorist organization whose "core membership" consisted of a mere 500 to 1,000 people at the time. Yet this attack, which directly killed roughly 3,000 people, resulted in two major wars that left 4,486 soldiers and almost half-a-million civilians dead, and cost the US taxpayer some $2.4 trillion.

(7) The 2016 Dyn cyberattack, which may have been perpetrated by a *lone* "angry gamer," single-handedly disrupted the websites of Airbnb, Amazon, BBC, *The Boston Globe*, CNN, Comcast, *FiveThirtyEight*, Fox News, *The Guardian*, iHeartRadio, Imgur, National Hockey League, Netflix, *The New York Times*, PayPal, Pinterest, Pixlr, Reddit, SoundCloud, Squarespace, Spotify, Starbucks, Storify, the Swedish Government, Tumblr, Twitter, Verizon Communications, Visa, Vox Media, Walgreens, *The Wall Street Journal*, *Wired*, Yelp, and Zillow—*among many others*.[56]

(8) New uranium enrichment technologies, such as SILEX (mentioned above), "could make nuclear power slightly cheaper, but could also be used to covertly manufacture nuclear weapons." Consequently, it has stoked new fears of nuclear proliferation among state and non-state actors.

---

[54] Indeed, need I mention that anyone can access the genomes of some of the most horrific pathogens, including Ebola and smallpox, with just a smartphone?

[55] Furthermore, whereas nuclear weapons require rare materials like enriched uranium and plutonium, biological microorganisms are self-replicating, some on timescales of "just twenty minutes, allowing microscopic amounts of organisms to be mass- produced in a brief period of time." And certain pathogens, such as anthrax, can be found in one's backyard; a 2015 study even discovered traces of anthrax and the bubonic plague in New York City subways (excerpted from *Morality*).

[56] Somewhat ironically, elsewhere in this chapter, Pinker cites an article by Bruce Schneier at the end of this sentence: "Military, financial, energy, and Internet infrastructure should be made more secure and resilient." The article is about the DDoS ("distributed denial-of-service") attack mentioned above, and it repeatedly emphasizes the increasing ease by which groups or individuals can wreak ever-greater damage! For example, he writes that "in December 2014, there was a legitimate debate in the security community as to whether [a previous] massive attack against Sony had been perpetrated by a nation-state with a $20 billion military budget or a couple of guys in a basement somewhere. … These attacks are getting larger. … A year ago, it was unheard of. Now it occurs regularly. … Expect these attacks to similarly increase." It's unclear—or perhaps all too clear—why Pinker ignores these passages in evaluating Kelly's claims. Indeed, elsewhere Schneier writes that "sooner or later, the technology will exist for a hobbyist to explode a nuclear weapon, print a lethal virus from a bio-printer, or turn our electronic infrastructure into a vehicle for large-scale murder. We'll have the technology eventually to annihilate ourselves in great numbers, and sometime after, that technology will become cheap enough to be easy."

> (9) A short video produced by the Campaign to Stop Killer Robots, called "Slaughterbots," depicts a realistic near-future scenario in which lethal autonomous weapons (LAWs), released by a potentially small group of anonymous agents, assassinates large groups of people. This story is similar to a scenario outlined by Stuart Russell: "A very, very small quadcopter, one inch in diameter can carry a one- or two-gram shaped charge. You can order them from a drone manufacturer in China. You can program the code to say: 'Here are thousands of photographs of the kinds of things I want to target.' A one-gram shaped charge can punch a hole in nine millimeters of steel, so presumably you can also punch a hole in someone's head. You can fit about three million of those in a semi-tractor-trailer. You can drive up I-95 with three trucks and have 10 million weapons attacking New York City. They don't have to be very effective, only 5 or 10% of them have to find the target." The punch-line is that "there will be manufacturers producing millions of these weapons that people will be able to buy just like you can buy guns now, except millions of guns don't matter unless you have a million soldiers. You need only three guys to write the program and launch [these drones]."

The list could go on—interminably. It could include both recent cases of small groups or lone wolves unilaterally inflicting unprecedented damage on societies, corporations, or whole governments, as well as probable future scenarios involving advanced synthetic biology, molecular nanotechnology, artificial intelligence, and other emerging artifacts that humanity has never before created, and thus has no track record of surviving. This should give us pause and, ultimately, galvanize humanity to devise effective strategies for realizing the good aspects of technology while neutralizing the bad aspects. (As John Sotos demonstrates, even an extremely small probability of an omnicidal agent successfully bringing about an existential catastrophe could more or less guarantee doom over periods of decades or centuries; probabilities accumulate![57]) If we manage to do this—if we "play our cards right"—the future could be unimaginably wonderful, and perhaps even utopian.

* * *

Why did I take the time to compose this nearly 22,000-word document (only about 4,000 words of which are Pinker's)? Because I worry that *Enlightenment Now* will have a far greater influence on culture, including intellectual culture, than any of the aforementioned books in the first section of this paper, namely, *The Future of Violence*, *Our Final Hour*, *Global Catastrophic Risks*, and *Here Be Dragons* (or my own *The End* and *Morality*). Consequently, a number of inaccurate, incomplete, and misleading ideas could become more or less permanent fixtures of our culture's shared futurological worldview, so to speak, and that these will cause people *not* to take seriously the increasingly vociferous warnings of scientists and philosophers who (a) hope for a long and prosperous life for humanity, and (b) are worried that our own actions, or our failure to take actions, could cut this life short. To underline this point, I would like to close this paper by excerpting the entire final section of *The End*, facetiously titled "The End is Here," which gestures at *why* the field of existential risk studies matters, a lot:

> As you read this sentence, the earth sits at the center of a giant, expanding bubble of leaked

[57] I provide similar calculations in this paper.

electromagnetic radiation hurtling through the universe at the cosmic speed limit of light. The outermost edges of this bubble—now some 140 light-years in diameter—contain programs like *The Howdy Doody Show*, *Meet the Press*, and *I Love Lucy* that were broadcast by early television stations, a major source of leakage, along with military radar. These signals were picked up by the antenna atop people's homes, but some ricocheted into the vastness of space.

A civilization living on the far side of our galaxy could, with the right technology (namely radio telescopes), potentially detect these stray signals from Earth and infer the existence of another life-form—one clever enough to accidentally seep streams of very high frequency (VHF) light into the sky. (Unfortunately, the inverse square law entails a significant degradation of these signals as they travel further from Earth.) The reverse situation applies as well: if a civilization developed to our level of sophistication, we could potentially detect its bubble of radiation spillage expanding toward our lonely planet.

Yet in all directions, the universe looks like a barren wasteland of dead matter mindlessly acting out a cosmic screenplay written by nature's laws. This panoply can be beautiful, for sure, but it lacks any convincing signs that intelligent life is crying out for companionship in a universe bereft of intrinsic meaning. There are no ripples of leaked radiation splashing against the shores of Earth. The sky is quiet—not a whisper, much less a shriek. The conundrum is that this is exactly opposite of what we would expect, given what we know about the natural world. The universe should be teeming with life, according to some estimates using the Drake Equation; we should be able to point our telescopes at the midnight firmament and see, at least on occasion, a spaceship flying by.

What explains this Great Silence? Perhaps there is a period in every intelligent species' life at which the *archaic* (old brains filled with old ideas about the universe) catastrophically collides with the *neoteric* (new technologies capable of rearranging the universe in new ways). We live in a world, today, in which scientists are literally studying the first *billionth* of a second after the Big Bang. At the very same time, a sizable portion of evangelicals await being suddenly "caught up" into the clouds with Jesus, and many Muslims are eagerly anticipating the appearance of the Mahdi, followed by Jesus in the arms of two angels. We live in a world that contains both Noam Chomsky and John Hagee, [Lord] Martin Rees and Sarah Palin, and Steven Pinker and Abu Bakr al-Baghdadi. If the rationality of our *ends* fails to match the rationality of our *means*, there's a fairly good chance that this century, the forty-five millionth since Earth formed, could be our last.

The universe, as J. B. S. Haldane once quipped, is not only stranger than we imagined, but stranger than we *could have* imagined. Consider that matter is literally more than 99.9% empty space; one species can morph into another species through evolution; the loudest animal relative to its body size is a tiny insect that produces sound by rubbing its penis against its stomach; the fastest organism in the world is the white mulberry tree (which propels its pollen at half the speed of sound); the universe literally has no center and no boundaries; fruit flies have sperm that are 2.3 inches long when uncoiled; as part of its life cycle, a marine invertebrate called the tunicate eats its own brain; the atmosphere gets colder, then warmer, then colder as you move toward space; most humans have some Neanderthal genes; the cosmos isn't expanding *into* space, rather space *itself* is expanding; if you leave your refrigerator door open on a hot summer day, the room will warm up rather than cool down; if you stop suddenly with a helium balloon in your car, it will dart backward rather than forward; and if you add up every single number from

one to infinity, the resulting sum will be exactly *negative one-twelfth*.[58]

Who knows what other mysteries await our discovery. Who knows what wonders might dazzle our imaginations and tickle our intellects with eureka, "ah-ha" moments. And who knows what our frail-brained species could become if the *good* uses of technology are allowed to realize the evolutionary adventure of redesigning our bodies and brains for purposes that conduce to happiness and morality.

While science, philosophy, art, culture, music, literature, poetry, fashion, sports, and all the other objects of civilization make life *worth* living, avoiding an existential catastrophe makes it *possible*. This makes eschatology, with its two interacting branches [i.e., religious and scientific], the most important subject that one could study. Without an understanding of what the risks are before us, without an understanding of how the clash of eschatologies [here] has shaped the course of world history, we will be impotent to defend against the threat of (self-)annihilation. The fact is that we are the only remaining species of human on the planet—the last one, *Homo floresiensis*, having died out some 12,000 years ago in Indonesia. Our situation has always been precarious, but it's never been as precarious as it is today. If we want our children to have the opportunity of living the Good Life, or even existing at all, it's essential that we learn to favor evidence over faith, observation over revelation, and science over religion as we venture into a dangerously wonderful future.

---

[58] The original footnote here states: "While this is not inaccurate, the statement requires some qualifications. For a good discussion of how this infinite series could possibly equal a negative fraction, see Evelyn Lamb, "Does 1+2+3 Really Equal -1/12?" *Scientific American*, January 20, 2014, and Phil Plait, "Follow-Up: The Infinite Series and the Mind- Blowing Result," *Slate*, January 18, 2014."