## **1NC**

#### **1 – Reject new extrapolations for impacts. Their card causally asserts claims without delineated warrants – don’t fill in gaps for them. Reject new ‘yes extinction’ evidence – we strategized based on their horrible piece of impact evidence. Study indicts and answers to our arguments solve – anything else encourages sandbagging good evidence until after we read our case which triggers a race to the bottom where debaters test judges’ thresholds for warrants into huge impacts, which infinitely regresses the education of debate.**

#### **2 – Islands repopulate**

**Turchin 18** [Alexey Turchin – Scientist for the Foundation Science for Life Extension in Moscow, Russia, Founder of Digital Immortality Now, author of several books and articles on the topics of existential risks and life extension. Brian Patrick Green – Director of technology ethics at the Markkula Center for Applied Ethics, teaches AI ethics in the Graduate School of Engineering at Santa Clara University. <MKIM> “Islands as refuges for surviving global catastrophes”. September 2018. DOA: 7/20/19. https://www.emerald.com/insight/content/doi/10.1108/FS-04-2018-0031/full/html?fullSc=1&mbSc=1&fullSc=1] Recut Justin

**Different types of** possible **catastrophes suggest different scenarios for how survival could happen on an island. What is important is** that **the island should have properties which protect against the specific dangers of** particular global **catastrophic risks. Specifically, different islands will provide protection against different risks, and their natural diversity will contribute to a higher total level of protection:**

**Quarantined island survives pandemic**. An island could impose effective quarantine **if it is sufficiently remote and simultaneously able to protect itself**, possibly using military ships and air defense.

**Far northern aboriginal people survive an ice age. Many far northern people have adapted to survive in extremely cold** and dangerous **environments, and** under the right circumstances **could** potentially **survive the return of an ice age**. However, their cultures are endangered by globalization. If these people become dependent on the products of modern civilization, such as rifles and motor boats, and lose their native survival skills, then their likelihood of surviving the collapse of the outside world would decrease. Therefore, **preservation of** their **survival skills may be important as a defense against the risks connected with extreme cooling.**

**Remote polar island with high mountains survives** brief **global warming of** median surface temperatures, **up to 50˚C.** There is a theory that the climates of planets similar to the Earth could have several semi-stable temperature levels (Popp et al., 2016). If so, **because of climate change**, the **Earth could transition to a** second semi-stable **state with a median global temperature of around** 330 K, about **60˚C**, or about **45˚C above current global mean temperatures**. But **even in this climate, some regions of Earth could still be survivable for humans, such as the Himalayan plateau at elevations above 4,000 m, but below 6,000** (where oxygen deficiency becomes a problem), **or on polar islands with mountains** (however, global warming affects polar regions more than equatorial regions, and northern island will experience more effects of climate change, including thawing permafrost and possible landslides because of wetter weather). In the tropics, the combination of increased humidity and temperature may increase the wet bulb temperature above 36˚C, especially on islands, where sea moisture is readily available. In such conditions, proper human perspiration becomes impossible (Sherwood and Huber, 2010), and there will likely be increased mortality and morbidity because of tropical diseases. **If temperatures later returned to normal** – either naturally or through climate engineering – **the rest of the Earth could be repopulated.**

‘‘Swiss Family Robinsons’’ survive on a tropical island, unnoticed by a military robot ‘‘mutiny’’. Most AI researchers ignore medium-term AI risks, which are neither near-term risks, like unemployment, nor remote risks, like AI superintelligence. But a large drone army – if one were produced – could receive a wrong command or be infected by a computer virus, leading it to attack people indiscriminately. Remote islands without robots could provide protection in this case, allowing survival until such a drone army ran out of batteries, fuel, ammunition or other supplies:

**Primitive tribe survives civilizational collapse**. The **inhabitants of North Sentinel Island**, near the Andaman Islands in the Indian Ocean, **are hostile and uncontacted. The Sentinelese survived the 2004 Indian Ocean tsunami apparently unaffected** (Voanews, 2009), and **if the rest of humanity disappear, they might well continue their existence without change.**

**Tropical Island survives extreme global nuclear winter and glaciation** event. **Were a nuclear**, bolide impactor or volcanic **“winter”** scenario **to unfold**, these **islands would remain surrounded by Warm Ocean, and local volcanism or other energy sources might provide heat, energy and food. Such island refuges** may have **helped life on Earth survive during the “Snowball Earth” event in Earth’s distant past** (Hoffman et al., 1998).

Remote island base for project “Yellow submarine”. Some **catastrophic risks such as a** gamma ray burst, a **global nuclear war with high radiological contamination** or multiple pandemics **might be best survived underwater in nuclear submarines** (Turchin and Green, 2017). However, after a catastrophe, the submarine with survivors would eventually need a place to dock, and an island with some prepared amenities would be a reasonable starting point for rebuilding civilization.

Bunker on remote island. **For risks** which include multiple or complex catastrophes, **such as** a bolide impact, extreme volcanism, tsunamis, multiple pandemics and **nuclear war with radiological contamination, island refuges could be strengthened with bunkers. Richard Branson survived hurricane Irma on his own island in 2017 by seeking refuge in his concrete wine cellar** (Clifford, 2017). **Bunkers on islands would have higher survivability compared to those close to population centers**, as they will be neither a military target nor as accessible to looters or unintentionally dangerous (e.g. infected) refugees. **These bunkers could potentially be connected to water sources by underwater pipes, and passages could provide cooling, access and even oxygen and food sources.**

#### **3 – Alternative foods solve**

John **Denkenberger**, 1/05/20**17**, Feeding Everyone if the Sun is Obscured and Industry is Disabled, https://hal.archives-ouvertes.fr/hal-02113471/file/Feeding\_Everyone\_if\_the\_Sun\_is\_Obscured.pdf)// recut JZ

**For combined sun blocking and industrial failure scenarios**, the reduced output of conventional agriculture would present a threat of causing mass starvation. This study showed that one **solution** in the short term **is extract**ing edible **calories from** killed **leaves** using distributed mechanical processes. Then a constrained food web could be formed where part of the remainder from this could be **fed to chickens**, and the rest coupled with leaf litter could have mushrooms grown on it. A second group of solutions is **grow**ing **mushrooms on** dead trees and the **residue** going **to** cellulose digesting animals such as **cattle** and rabbits. Typically, in these catastrophes the sun is not blocked completely, so some **agriculture would be possible based off of existing farming in extreme environments (e.g. growing UV and cold tolerant crops in the tropics). Furthermore, the cooling climate would cool the upper layer of the ocean, causing upwelling of nutrient-rich deep ocean water. This would facilitate algae growth in the ocean, feeding fish; retrofitting of ships to be sail powered could enable significant fishing. The results of this study show these solutions could enable the feeding of everyone given minimal preparation**, and this preparation should be a high priority now.

#### **4 – Adaptions solve**

Jeffrey **Ladish**, 11/06/20**20**, Nuclear war is unlikely to cause human extinction, LessWrong, https://www.lesswrong.com/posts/sT6NxFxso6Z9xjS7o/nuclear-war-is-unlikely-to-cause-human-extinction)// JZ

1) Kinetic destruction

**There simply aren't enough nuclear warheads to kill everyone directly with kinetic force, and there likely never will be.** There are ~14,000 nuclear weapons in the world, and let’s suppose they have an average yield of something like 1 megaton. This is a conservative guess, the actual average is probably closer to 100 kilotons. With a 1 megaton warhead, you can create a fireball covering 3 km², and a moderate pressure wave that knocks down most residential houses covering 155 km². The former kills nearly everyone and the latter kills a decent percentage of people but not everyone. Let's be conservative and assume the pressure wave kills everyone in its radius. 14,000 \* 155 = 2.17 million km². The New York Metro area is 8,683 km². So all the nuclear weapons in the world could destroy about 250 New York Metro areas. **This is a lot! But not near enough, even if someone intentionally tried to hit all the populations at once.** Total land surface of earth is: 510.1 million km². Urban area, by one estimate, is about 2%, or 10.2 million km.² Since the total possible area destroyed from nuclear weapons is ~2.17 million km² is considerably less than a lower bound on the area of human habitation, 10.2 million km², there should be basically no risk of human extinction from kinetic destruction. The circle with the white border indicates the zone of moderate blast damage radius (5 psi): 7.03 km (155 km²) from a 1,000 kiloton warhead, [link to nukemap](https://nuclearsecrecy.com/nukemap/?&kt=1000&lat=40.7648&lng=-73.9808&hob_psi=5&hob_ft=10245&psi=20,5,1&zm=11)

If you want to check my work there, I was using [nuke map](https://nuclearsecrecy.com/nukemap/?&kt=1000&lat=40.7648&lng=-73.9808&hob_psi=5&hob_ft=10245&psi=20,5,1&zm=11).

**The even more obvious reason why kinetic damage wouldn't lead to human extinction is that nuclear states only threaten one or several countries at a time, and never the population centers of the entire world.** Even if NATO countries and Russia and China all went to war at the same time, Africa, South America, and other **neutral regions would be spared any kinetic damage.**

**2) Radiation**

**Radiation won't kill everyone because there aren't enough weapons, and radiation from them would be concentrated in some areas and wholly absent from other areas. Even in the worst affected areas, lethal radiation from fallout would drop to survivable levels within weeks.**

Here it's worth noting that there is an inherent tradeoff between length of halflife and energy released by radionuclides. The shorter the half life the more energy will be released, and the longer the half life the less energy. The fallout products from modern nuclear weapons are very lethal, but only for days to several weeks. From Nuclear War Survival Skills, 1987 edition

Let's try the same calculation we used with kinetic damage, and see if an attack aimed at optimizing fallout for killing everyone could succeed. Using [Nukemap](https://nuclearsecrecy.com/nukemap/?&kt=1000&lat=40.7648&lng=-73.9808&airburst=0&hob_ft=0&fallout=1&psi=20,5,1&zm=8) again, I'll go with the fallout contour for 100 rads per hour. 400 rads is thought too be enough to kill 50% of people, so 100 rads per hour is likely to kill most all people not in some kind of shelter. We need to switch to using a groundburst detonation rather than an airburst detonation, because groundbursts create far more fallout. A 1mt ground burst would create an area of about 8,000 km² of >100 rads per hour. Okay, multiple that by 14,000 warheads, and we get 112 million km². That's a lot! It's still less than the 510.1 million km² of earth's land mass, but it's a lot more than the ~10.2 million km² of urban space. Presumably this is enough to cover every human habitation, so in principle, it might be possible to kill everyone with radiation from existing nuclear weapons.

The bright red and slightly less bright red indicate fallout contour for 1,000 rads and 100 rads per hour, covering 1,140 km² and 7,080 km² respectively, from a 1,000 kiloton ground burst. [Nukemap settings](https://nuclearsecrecy.com/nukemap/?&kt=1000&lat=40.7648&lng=-73.9808&airburst=0&hob_ft=0&fallout=1&psi=20,5,1&zm=8) In practice, it would be almost impossible to kill every human via radiation with the existing nuclear arsenals, even if they were targeted explicitly for this purpose. **The first reason is that fallout patterns are very uneven. After a ground burst, fallout is carried by the wind. Some areas will be hit bad and some areas will be hardly affected by fallout. Even if most human population centers were covered, a few areas would almost certainly escape.**

Two other things make extinction by radiation unlikely. Many countries, especially in the southern hemisphere, are unlikely to be affected by fallout much at all. **Since most of these countries are likely to be neutral in a conflict, and not near combatant countries, they should be relatively safe from fallout. While fallout might travel hundreds of kms, it still won't reach places separated by greater distances. Fallout that reaches the upper atmosphere will eventually fall back down, but usually after the period of lethal radioactivity.** The other mitigating factor is that in typical nuclear war plans, ground bursts are usually restricted to hardened targets, and air bursts are favored for population and industry centers. This is because air bursts maximize the size of the destructive pressure wave. Air burst detonations result in little lethal fallout reaching the ground, so populations not downwind of military targets would likely be safe from the worst of the radiological effects in a war scenario.

The final protection from extinction by radiation is simply large amounts of mass between people and the radiation source, in other words, fallout shelters. After several weeks, the radionuclides in fallout from ground burst detonations will have decayed to the point where humans can survive outside of shelters. **Many fallout shelters exist in the world, and many more could be made easily in a day or two with a shovel, some ground, and some boards. Even if lethally radioactive fallout from ground bursts covered all population centers, many humans would still survive in shelters.**

The risks of extinction from nuclear-weapon-induced-radiation wouldn't be complete without discussing two factors: nuclear power plants and radiological weapons. I'm only going to cover these briefly, but they both don't change the conclusions much.

Nuclear power plants could be targeted by nuclear weapons to create large amounts of fallout with a longer half-life but less energy per unit time. The main concern here is that nuclear power plants and spent fuel sites contain a much greater \*mass\* of radioactive material than nuclear missiles can carry. The danger comes primarily from spreading the already very radiative spent or unspent nuclear fuel. The risk this poses requires a longer analysis, but the short version is that while nuking a nuclear power plant or stored fuel site would indeed create some pretty long-lived fallout it would still be concentrated in a relatively small area. Fortunately, even a nuclear detonation wouldn't spread the nuclear fuel more than several hundred km at most. Having regions of countries covered in spent nuclear fuel would be awful, but it doesn't much raise the risk of extinction.

Radiological weapons are nuclear weapons designed to maximize the spread of lethal fallout rather than destructive yield. The particular concern from the extinction perspective is that they can be designed to create fallout that continues to emit levels of radiation that can make an area uninhabitable for months to years. These kind of radiological weapons kill more slowly, but they still kill. In principle, radiological weapons could be used to kill everyone on earth. However, in practice, the same constraints that apply to standard nuclear weapons apply to weapons optimized for long-lasting fallout, as well as some additional constraints.

Radiological weapons wouldn't produce more fallout than standard warheads, they would just produce fallout with different characteristics. As a result the amount of radiological weapons required to cover every part of earth's surface would be massively expensive (likely as expensive as the largest existing nuclear arsenals), and serve no military purpose. Their inefficiency in destruction and death compared to standard nuclear weapons is probably why radiological weapons have never been developed or deployed in large numbers. This makes them an ongoing theoretical concern, but not an existential risk in the immediate future. A [concerning development](https://en.wikipedia.org/wiki/Status-6_Oceanic_Multipurpose_System) is Russia's claim to have developed a large-yield (100mt) submersible nuclear weapon with the suggestion that it could be used as a radiological weapon, but even if this is true, it's unlikely to be deployed in large numbers.

3) Climate alteration

The bulk of the risk of human extinction from nuclear weapons come from risks of catastrophic climate change, nuclear winter, due to secondary effects from nuclear detonations. **However, even in most full-scale nuclear exchange scenarios, the resulting climate effects are unlikely to cause human extinction.**

**Reasons for this:**

**a) Under scenarios where a severe nuclear winter occurs as described by Robock et al, some human populations would likely survive.**

**b) The Robock group’s models are probably overestimating the risk**

**c) Nuclear war planners are aware of nuclear winter risks and can incorporate these risks into their targeting plans**

Before diving into each subject, it’s worth understanding the background of nuclear winter research. In the 1980s a group of atmospheric scientists proposed the hypothesis that a nuclear war would result in massive firestorms in burning cities, which would loft particles high into the atmosphere and cause catastrophic cooling that would last for years. Many found it alarming that such an effect could be possible and go unnoticed for decades while the risk existed. Some scientists also thought the proposed effect was too strong, or unlikely to occur at all. Until a few years ago, if you looked only at peer reviewed literature you would only find papers forecasting severe nuclear winter effects in the event of a nuclear war. Understandably, many people assumed that this was the scientific consensus. Unfortunately, this misrepresented the scientific community’s state of uncertainty about the risks of nuclear war. There have only ever been a small numbers of papers published about this topic (<15 probably), mostly from one group of researchers, despite the topic being one of existential importance.

I’m very glad Robock, Toon, and others have spent much of their careers studying nuclear winter effects, and their models are useful in estimating potential climate change caused by nuclear war. However, I’ve become less convinced over time the Robock model is largely correct. See section B below for why I’ve changed my mind. However, I’m quite uncertain about the probability of strong cooling effects from nuclear war, and am still quite concerned about the potential for severe cooling, even if the risk of extinction from such events is small.

A: Under scenarios where a severe nuclear winter occurs as described by Robock et al, some human populations would likely survive.

The latest and most detailed model of potential cooling effects from a fullscale nuclear exchange comes from, Robock et al., “Nuclear winter revisited with a modern climate model and current nuclear arsenals: Still catastrophic consequences” found [here](http://climate.envsci.rutgers.edu/pdf/RobockNW2006JD008235.pdf).

The effects from this model are severe. In the 150Tg case, after a year, summer temperatures in the Northern hemisphere are 10-30 degrees C cooler. The effects are less severe at the equator (5 degrees C), but basically all places in the world are affected. The most likely outcome is that most people starve to death. **Many would freeze too, but starvation is likely the greatest risk. Even in this model, it appears that in equatorial regions, some farming would still be possible, enough for some populations to survive. After a 10-15 years, agriculture in most of the world would be possible** at reduced capacity. Surface air temperature changes for the 150 Tg case averaged for June, July, and August of the year of smoke injection and the next year. [Robock et al., 2007](http://climate.envsci.rutgers.edu/pdf/RobockNW2006JD008235.pdf)

Carl Shulman asked one of the authors of this paper, Luke Oman, his probability that the 150Tg nuclear winter scenario discussed in the paper would result in human extinction, [the answer he gave](https://www.overcomingbias.com/2012/11/nuclear-winter-and-human-extinction-qa-with-luke-oman.html) was “in the range of 1 in 10,000 to 1 in 100,000.” This strikes me as quite plausible, though one expert opinion is no substitute for a deep analysis. The [Q&A with Oman](https://www.overcomingbias.com/2012/11/nuclear-winter-and-human-extinction-qa-with-luke-oman.html) contains his reasoning for this assessment.

Two different analyses are required to calculate the chances of human extinction from nuclear winter. The first is the analysis of the climate change that could result from a nuclear war, and the second is the adaptive capacity of human groups to these climate changes. I have not seen an in depth analysis of the latter, but I believe such an assessment would be worthwhile.

My own guess is that **humans are capable of surviving far more severe climate shifts than those projected in nuclear winter scenarios.** Humans are more robust than most any other mammal to drastic changes in temperature, as evidenced by our global range, even in pre-historic times. While a loss of most agriculture would likely kill most people on earth, modern technology would enable some populations to survive. **Great stores of food currently exist in the world, and it is l likely that some of these would be seized and protected by small groups, providing enough food to last for years.** While even such populations with food stores wouldn’t have enough to survive for 10-15 years, such food stores would give groups time to adapt to new food sources. The organization [ALLFED](https://allfed.info/) has explored a number of alternative food sources that could keep populations alive in the event of a nuclear war or other large solar disruption, and I expect great necessity to drive the discovery of even more in the event of such a disaster.

B: The Robock group’s models are probably overestimating the risk

The nuclear winter model at its simplest: Nuclear detonations → Fires in cities → Firestorms in cities → Lofted black carbon into the upper atmosphere → black carbon persists in upper atmosphere, reflecting sunlight and causes massive cooling Each step is required in order for the effect to occur. If nuclear war causes massive fires in cities but does not lead to firestorms that loft particles, then no long term cooling is going to occur.

#### **Society won't redevelop hazardous advanced tech post-fallout.**

**Dartnell 15** [Lewis Dartnell, professor of science communication at the University of Westminster; April 13, 2015; “Out of the ashes,” https://aeon.co/essays/could-we-reboot-a-modern-civilisation-without-fossil-fuels]

**Imagine that** the world as we know it ends tomorrow. **There’s a** global catastrophe: a pandemic virus, an asteroid strike, or perhaps a **nuclear holocaust. The vast majority of the human race perishes.** Our civilisation collapses. The post-apocalyptic survivors find themselves in a devastated world of decaying, deserted cities and roving gangs of bandits looting and taking by force.

Bad as things sound, **that’s not the end for humanity. We bounce back**. Sooner or later, peace and order emerge again, just as they have time and again through history. **Stable communities take shape. They begin** the agonising process of **rebuilding their technological base from scratch.** But here’s the question: how far could such a society rebuild? **Is there any chance**, for instance, **that a post-apocalyptic society could reboot a technological civilisation**?

Let’s make the basis of this thought experiment a little more specific. Today, **we have already consumed the most easily drainable crude oil and**, particularly in Britain, much of the shallowest, most readily mined **deposits of coal. Fossil fuels are central to** the organisation of **modern industrial society**, just as they were central to its development. Those, by the way, are distinct roles: even if we could somehow do without fossil fuels now (which we can’t, quite), it’s a different question whether we could have got to where we are without ever having had them.

So, would a society starting over on a planet stripped of its fossil fuel deposits have the chance to progress through its own Industrial Revolution? Or to phrase it another way, what might have happened if, for whatever reason, the Earth had never acquired its extensive underground deposits of coal and oil in the first place? Would our progress necessarily have halted in the 18th century, in a pre-industrial state?

**It’s easy to underestimate our current dependence on fossil fuels**. In everyday life, their most visible use is the petrol or diesel pumped into the vehicles that fill our roads, and the coal and natural gas which fire the power stations that electrify our modern lives. But we also rely on a range of different industrial materials, and in most cases, high temperatures are required to transform the stuff we dig out of the ground or harvest from the landscape into something useful. **You can’t smelt metal, make glass, roast** the ingredients of **concrete, or synthesise artificial fertiliser** without a lot of heat. It is fossil fuels – coal, gas and oil – that provide most of this thermal energy.

In fact, **the problem is even worse than that. Many of the chemicals required** in bulk **to run the modern world**, from pesticides to plastics, **derive from** the diverse organic compounds in **crude oil**. Given the dwindling reserves of crude oil left in the world, it could be argued that the most wasteful use for this limited resource is to simply burn it. We should be carefully preserving what’s left for the vital repertoire of valuable organic compounds it offers.

But my topic here is not what we should do now. Presumably everybody knows that we must transition to a low-carbon economy one way or another. No, I want to answer a question whose interest is (let’s hope) more theoretical. Is the emergence of a technologically advanced civilisation necessarily contingent on the easy availability of ancient energy? **Is it possible to build an industrialised civilisation without fossil fuels?** And the answer to that question is: **maybe – but it would be extremely difficult.** Let’s see how.

We’ll start with a natural thought. Many of our alternative energy technologies are already highly developed. Solar panels, for example, represent a good option today, and are appearing more and more on the roofs of houses and businesses. It’s tempting to think that a rebooted society could simply pick up where we leave off. **Why couldn’t our civilisation 2.0 just start with renewables**?

Well, **it could, in a very limited way**. If you find yourself among the survivors in a post-apocalyptic world, you could scavenge enough working solar panels to keep your lifestyle electrified for a good long while. Without moving parts, photovoltaic cells require little maintenance and are remarkably resilient. **They do deteriorate over time, though**, from moisture penetrating the casing and from sunlight itself degrading the high-purity silicon layers. The electricity generated by a solar panel declines by about 1 per cent every year so, **after a few generations, all our hand-me-down solar panels will have degraded** to the point of uselessness. Then what?

**New ones would be fiendishly difficult to create from scratch.** Solar panels are made from thin slices of extremely pure silicon, and although the raw material is common sand, it must be processed and refined using complex and precise techniques – the same technological capabilities, more or less, that we need for modern semiconductor electronics components. These techniques took a long time to develop, and would presumably take a long time to recover. So photovoltaic solar power would not be within the capability of a society early in the industrialisation process.

#### **Suffering risks outweighs**

Di **Minardi**, 10/15/20**20**, The grim fate that could be ‘worse than extinction,’ BBC,

https://www.bbc.com/future/article/20201014-totalitarian-world-in-chains-artificial-intelligence)// JZ

Toby Ord, a senior research fellow at the Future of Humanity Institute (FHI) at Oxford University, believes that the odds of an existential catastrophe happening this century from natural causes are less than one in 2,000, because humans have survived for 2,000 centuries without one. However, when he adds the probability of human-made disasters, Ord believes the chances increase to a startling one in six. He refers to this century as “the precipice” because the risk of losing our future has never been so high.

Researchers at the Center on Long-Term Risk, a non-profit research institute in London, have expanded upon x-risks with the even-more-chilling prospect of **suffering risks.** These “s-risks” **are** defined as “**suffering on an astronomical scale, vastly exceeding all suffering that has existed on Earth so far.” In these scenarios, life continues for billions of people, but the quality is so low and the outlook so bleak that dying out would be preferable. In short: a future with negative value** is worse than one with no value at all.

This is where the “world in chains” scenario comes in. **If a malevolent group or government suddenly gained world-dominating power through technology, and there was nothing to stand in its way, it could lead to an extended period of abject suffering and subjugation.** A 2017 report on existential risks from the Global Priorities Project, in conjunction with FHI and the Ministry for Foreign Affairs of Finland, warned that “a long future under a particularly brutal global totalitarian state could **arguably be worse than complete extinction”.**

Singleton hypothesis

Though global totalitarianism is still a niche topic of study, researchers in the field of existential risk are increasingly turning their attention to its most likely cause: artificial intelligence.

#### **The risk is extremely high.**

#### **1 – Incidental risks.**

**Baumann 22** [Tobias Baumann, former researcher at University College London focusing on Cooperative Artificial Intelligence, co-founder of the Center for Reducing Suffering; April 22; Avoiding the Worst: How to Prevent a Moral Catastrophe]

Incidental s-risks

Incidental s-risks arise when efficient ways to achieve a certain goal creates a lot of suffering in the process, without anyone actively trying to cause suffering per se. The agent or agents that cause the s-risk are either indifferent to that suffering, or they would prefer a suffering-free alternative in theory, but are not willing to bear the necessary costs in practice.

We can further divide incidental s-risks into subcategories based on the underlying motivation. In one class of scenarios, **economic incentives** and market forces **cause large amounts of suffering because that suffering is a byproduct of high economic productivity.** Animal suffering in **factory farms is a case in point.** It just so happens that the most economically efficient way to satisfy the demand for cheap animal products **entails miserable conditions for farmed animals.**4 **Future technology might enable similar dynamics but on a much larger scale.** For instance, **the fact that evolution uses pain suggests that learning might be more efficient if negative reward signals are also used**, and we might consider sufficiently advanced and complex reinforcement learners to be capable of suffering.

Another possibility involves suffering that is instrumental for information gain. **Experiments on sentient creatures can be useful for scientific purposes**, while causing serious harm to those experimented on. Again, **future tech**nology **may enable such practices** on a much larger scale. As discussed in the previous chapter, **it may become possible to run a large number of simulations of artificial minds that are capable of suffering**. And if there are instrumental reasons to run many such simulations, this could lead to vast amounts of suffering. For example, an **advanced AI** system might **run many simulations to improve its knowledge of human psychology or** in an attempt **to predict what other agents will do in a certain situation.**

**It is also conceivable that complex simulations will be used for entertainment purposes in the future, which could cause serious suffering** if these simulations contain artificially sentient beings. **Many people enjoy violent forms of entertainment, as evidenced by** countless historical examples, from **gladiator fights** to **public executions.** Another case in point is the content of today’s **video games or movies. Such forms** of entertainment are victimless as long as they are fictional — but in combination with sentient artificial minds, they **could be a potential s-risk.**

#### **2 – Agential risks.**

**Baumann 22** [Tobias Baumann, former researcher at University College London focusing on Cooperative Artificial Intelligence, co-founder of the Center for Reducing Suffering; April 22; Avoiding the Worst: How to Prevent a Moral Catastrophe]

**Agential s-risks might also arise when people harbor strong** feelings of **hatred towards others. One relevant factor is the human tendency to** form tribal identities and **divide the world into an ingroup and an outgroup**. In extreme cases, **such tribalism spirals into a desire to harm the other side as much as possible. History features many well-known examples of atrocities committed against** those that belong to **the “wrong” religion, ethnic group, or political ideology. Similar dynamics could unfold on an astronomical scale in the future.**

**Another theme is retributivism: seeking vengeance for** actual or perceived **wrongdoing by others**. **A concrete example is excessive criminal punishment, as evidenced by historical** and contemporary penal systems that have inflicted **extraordinarily cruel forms of punishment.**

These different themes can overlap or take place as part of an escalating conflict. For instance, **large-scale warfare** or terrorism **involving advanced tech**nology **might amount to an s-risk.** This is both because of the potential suffering of the combatants themselves and because extreme conflict tends to reinforce negative dynamics such as sadism, tribalism, and retributivism. War often brings out the worst in people. It is also conceivable that agents would, as part of an escalating conflict or war, make threats to deliberately bring about worst-case outcomes in an attempt to force the other side to yield.

#### **3 – Selective alignment.**

**Sauer 22** [AI expert and researcher, December 24, 2022; “The case against AI alignment,” https://www.lesswrong.com/posts/CtXaFo3hikGMWW4C9/the-case-against-ai-alignment]

I am not here to level criticisms of this type at the AI alignment community. I accept most of the descriptive positions endorsed by this community: I believe that **AGI is possible and will inevitably be achieved within the next few decades, I believe that** the alignment problem is not trivial and that **unaligned AGI will likely** act against human interests to such an extent as to **lead to the extinction of** the human race and probably **all life** as well. My criticism is rather on a moral level: do these facts mean that we should attempt to develop AI alignment techniques?

I say we should not, because **although the** risks and **downsides of unaligned strong AI are great, I do not believe that they even remotely compare** in scope **to the risks from strong AI alignment** techniques **in the wrong hands. And** I believe that **the vast majority of hands this technology could end up in are** the **wrong** hands.

You may reasonably ask: How can I say this, when I have already said that unaligned strong AI will lead to the extinction of humanity? What can be worse than the extinction of humanity? The answer to that question can be found very quickly by examining many possible nightmare scenarios that AI could bring about. And the common thread running through all of these nightmare scenarios is that the AI in question is almost certainly aligned, or partially aligned, to some interest of human origin.

**Unaligned AI will kill you**, because you are made of atoms which can be used for paper clips instead. It will kill you **because it is completely uninterested in you. Aligned**, or partially aligned **AI**, by contrast, **may well take a considerable interest in you** and your well-being or lack thereof. It does not take a very creative mind to imagine how **this can be significantly worse**, and a superintelligent AI is more creative than even the most deranged of us.

I will stop with the euphemisms, because this point really needs to be driven home for people to understand exactly why I am so insistent on it. **The world** as it exists **today**, at least sometimes, **is unimaginably horrible**. People have endured things that would make any one of us go insane, more times than one can count. Anything you can think of which is at all realistic has happened to somebody at some point in history. **People have been skinned alive, burned and boiled alive, wasted away from agonizing disease, crushed to death, impaled, eaten alive**, succumbed to thousands of minor cuts, been raped, been forced to rape others, drowned in shit, trampled by desperate crowds fleeing a fire, and really anything else you can think of. People like Junko Furuta have suffered torture and death so bad you will feel physical pain just from reading the Wikipedia article. Of course, **if you care about animals, this gets many orders of magnitude worse**. I will not continue to belabor the point, since others have written about this far better than I ever can. On the Seriousness of Suffering (reducing-suffering.org) The Seriousness of Suffering: Supplement – Simon Knutsson

I must also stress that all of **this has happened in a world significantly smaller than one an AGI could create, and with a limited capacity for suffering. There is only so much harm** that **your body** and mind **can physically take** before they give out. **Torturers have to restrain themselves** in order to be effective, since if they do too much, their victim will die and their suffering will end. **None of these things are guaranteed to be true in a world augmented with** the technology of **mind uploading. You can** potentially **try every torture you can think of**, physically possible or no, on someone in sequence, complete with modifying their mind so they never get used to it. You can **create** new digital beings by the **trillions just for this purpose** if you really want to.

I ask you, **do you really think that an AI aligned** to human values **would refrain** from doing something like this to anyone? **One of the most fundamental** aspects of **human values is the hated outgroup. Almost everyone has somebody they’d love to see suffer**. How many times has one human told another “burn in hell” and been entirely serious, believing that this was a real thing, and 100% deserved? Do you really want technology under human control to advance to a point where this threat can actually be made good upon, with the consent of society? **Has there ever been any technology** invented in history which has **not** been terribly and **systematically misused at some point?**

Mind uploading will be abused in this way if it comes under the control of humans, and it almost certainly will not stop being abused in this way when some powerful group of humans manages to align an AI to their CEV. **Whoever controls the AI will** most likely **have somebody whose suffering they don’t care about, or that they want to enact**, or that they have some excuse for, because that describes the values of the vast majority of people. **The AI will perpetuate it** because that is what the CEV of the controller will want it to do, **and with value lock-in, this will never stop happening until the stars burn themselves out** and there is no more energy to work with.

**Do you really think** extrapolated **human values don’t have this potential? How many** ordinary, regular people **throughout history have become the worst kind of sadist under the slightest excuse or social pressure** to do so to their hated outgroup? What society hasn’t had some underclass it wanted to put down in the dirt just to lord power over them? How many people have you personally seen who insist on justifying some form of suffering for those they consider undesirable, calling it “justice” or “the natural order”?

#### **It’s coming now, and causes extinction.**

**McAleese 22** [Stephen; August; software engineer, Amazon Web Services, CloudWatch monitoring, Machine Learning Safety Scholar, Center for AI Safety; Arxiv, “How Do AI Timelines Affect Existential Risk?” https://arxiv.org/pdf/2209.05459.pdf]

1 Introduction

Recent **progress in AI suggests** that artificial general intelligence (**AGI**) that is as capable as humans on a wide variety of tasks **is likely to be created this century. Once AGI exists**, further **improvement of** its **abilities** would **enable it to surpass human intelligence resulting in** the creation of a **s**uperintelligent **a**rtificial general **i**ntelligence **that is vastly more capable than humans** at many important cognitive tasks.

A similar concept is “transformative AI” which is defined as “AI that precip- itates a transition comparable to (or more significant than) the agricultural or industrial revolution” [11]. For the purpose of this report I’m going to use the acronym “ASI” (artificial superintelligence) as a short-hand for superintelligent general AI that is “much smarter than the best human brains in practically every field, including scientific creativity, general wisdom and social skills” [2].

**After** an **ASI** is created, **it would** probably **have a significant** and long-lasting **effect on human civilization and its trajectory.** An **ASI could also be a significant source of existential risk** and could result in a highly undesirable outcome such as human extinction.

Although recently created AI systems such as GPT-3 can accomplish a wide variety of tasks, their understanding of the world, generality and performance is not high enough for them to be a significant existential risk to humanity.

**To be** a major source of **existential** risk, an **AI** system **might need to have** the kind of **deep**, cross-domain **understanding humans have that enables them to significantly change the world by** taking actions such as **creating** and implementing **complex long-term plans or inventing powerful new tech**nologies.

For example, an AI that wanted to invent and deploy advanced nanotechnology might need to have the ability to read and understand scientific papers, plan, carry out and interpret the results of experiments, and model the behavior of other actors such as humans.

Since current AI systems are not intelligent enough to have one or more of these general abilities or significantly transform the world, they are not a major source of existential risk. But as AI progress continues, humanity might someday create ASI systems that are intelligent and powerful enough to be an existential risk to humanity.

1.1 Motivation

Some philosophers believe that **reducing existential risk could have extremely high expected value** [1, 10]. For example, axiological strong longtermism states that “**impact on the far future is** the **most important feature of our actions today**” and that “every option that is near-best overall is near-best for the far future” [10]. Research on existential risk is also generally neglected [1].

**Since** the creation of **ASI may be** a **major** source of e**x**istential **risk, actions that reduce** the **level of** existential **risk posed by ASI would have extremely high** expected **value**.

**Researchers** such as Ajeya Contra have **analyzed** **when humanity might create** transformative **AI** (TAI) in the future by, for example, **using** **biological anchors to estimate** the amount of **computational power necessary** for TAI and extrapolating past progress in AI to estimate when TAI will be created [5]. **This** kind of **research is important because it informs humanity on how it should** act to **minimize AI risk.** For example, AI safety research involving current machine learning methods is more likely to be relevant and valuable for TAI safety if shorter rather than longer timelines are expected. AI timelines could also affect funding decisions or other priorities.

Although a significant amount of research effort has been put into estimating AI timelines, apparently much less research has been directed at the question of how AI timelines affect existential risk and similarly which ASI arrival date would be most desirable from the standpoint of existential risk reduction. **Most** previous **research seems to have been descriptive rather than normative** with a focus on predicting the arrival date of ASI as if it were a fixed and inevitable moment in the future **with less attention being directed at how AI timelines affect existential risk** and which ASI arrival date would be most desirable given the goal of minimizing existential risk.

This report is focused on answering the latter question: how do AI timelines affect total existential risk? And given the goal of minimizing total existential risk, should we prefer a world where ASI is created in the near or far future?

Existential catastrophe occurring over time. Since the creation of ASI is likely one of the main sources of existential risk facing humanity this century [15], it would be valuable to study how various AI development trajectories affect total existential risk.

2 What is the magnitude of existential risk from ASI this century?

Before we compare existential risks, it will be useful to estimate the existential risk contribution of ASI to total existential risk.

The amount of existential risk from ASI this century depends on how likely ASI is to be created this century and how likely an existential catastrophe is to occur afterwards. I’ll describe several sources of information we can use to estimate these two variables.

2.1 Expert surveys

A 2014 **survey asked experts when** they thought high-level machine intelligence (**HLMI) would be created where HLMI was defined as a machine that can “carry out most human professions at least as well as a typical human.**” The survey found a median estimate that **HLMI had a 50% chance of** being **developed by 2040.** Respondents were also asked when they thought superintelligence would be developed which was defined as “machine intelligence that greatly surpasses the performance of every human in most professions.” The median **estimate was** a 10% **probability of superintelligence** within 2 years after the creation of HLMI and a **75% probability within 30 years.** The same survey found that, on average, experts believed that HLMI had an 18% chance of causing an existential catastrophe [14].

Another survey [9] estimated a 50% probability of HLMI 50 years after 2016 and that the median expert believed the probability of an extremely bad outcome was 5%.

2.2 The Precipice

In The Precipice [15], Toby Ord estimates that the probability of unaligned AI causing an existential catastrophe is about 10% in the 21st century.

2.3 Metaculus

The Metaculus prediction market currently predicts that there is a 50% proba- bility of artificial general intelligence being created by 2041 [8].

2.4 Ajeya Contra

Ajeya Contra is a researcher at Open Philanthropy who has spent a signifi- cant amount of time predicting when transformative AI (TAI) will be created. Recently, she revised her median prediction from 2050 to 2040 [6].

2.5 A priori arguments for ASI being a significant source of existential risk

Humanity does not have any previous experience dealing with ASI systems. Therefore, we have little evidence we can draw on to estimate how much existential risk would be associated with the creation of an ASI. However, **there are** still **a priori reasons to believe** that **ASI would be a significant** source of e**x**istential **risk** [7].

What if humanity created an ASI and programmed it with a random goal? What would we expect to happen? Should we expect a positive or negative outcome by default? This section explains why ASI systems are likely to be harmful without careful countermeasures to make them aligned and beneficial.

In a chapter named “Is the default outcome doom?” (p. 140) in Superin- telligence, Nick Bostrom explains why programming an ASI to be beneficial wouldn’t be easy [3]. He gives three reasons:

1. The **orthogonality thesis says that there is no correlation between intelligence and having beneficial goals.** Therefore, **we cannot expect an AI to acquire beneficial goals** simply by increasing its intelligence.

2. **It is much easier to program an ASI to have a meaningless goal** such as “count digits of Pi” **rather than a complex** and valuable **goal** like “achieve human flourishing”.

3. The instrumental **convergence thesis says** that **AIs will have** certain **subgoals such as resource acquisition for** a wide variety of **final goals.** For example, **if we gave an AI a** random **goal, it is likely that it would acquire resources to** help it **achieve** its goal. An ASI programmed to maximize the probability of some goal being achieved would be **incentivized to pursue extreme** and extensive **resource acquisition** efforts such as building huge numbers of mines and power plants **that would make Earth uninhabitable for life.** Bostrom calls this particular risk infrastructure profusion.

#### **Exponential growth rates and quantum computing make it a structural inevitability.**

**Dilmegani 24** [Cem Dilmegani, Computer engineer from Bogazici University, M.B.A. from Columbia Business School; January 1, 2024; “When will singularity happen? 1700 expert opinions of AGI [2024],” https://research.aimultiple.com/artificial-general-intelligence-singularity-timing/]

**Will AGI** / singularity **ever happen? According to most AI experts, yes.**

When will the singularity / AGI happen? Before the end of the century. The consensus view was that it would take around 50 years in 2010s. After the advancements in Large Language Models (LLMs), some leading AI researchers updated their views. For example, Hinton believed in 2023 that it could take 5-20 years.1

What is our current status? While there are narrow AI solutions that exceed humans in many tasks, a generally intelligent machine doesn’t exist even though some researchers believe that large language models exhibit emerging, more generalist capabilities than other existing AI models.2

The more nuanced answers are below. There have been several surveys and research of AI scientists asking about when such developments will take place.

Understand the results of major surveys of AI researchers in 2 minutes

**We looked at the results of 5 surveys with** around **1700 participants** where researchers estimated when singularity would happen. **In all cases, the majority of participants expected AI singularity before 2060.**

In the 2022 Expert Survey on Progress in AI, conducted with 738 experts who published at the 2021 NIPS and ICML conferences, AI experts estimate that there’s a 50% chance that high-level machine intelligence will occur until 2059.

Older surveys had similar conclusions. In 2009, 21 AI experts participating the in AGI-09 conference were surveyed. Experts believed AGI will occur around 2050, and plausibly sooner. You can see above their estimates regarding specific AI achievements: passing the Turing test, passing third grade, accomplishing Nobel worthy scientific breakthroughs and achieving superhuman intelligence.

In 2012/2013, Vincent C. Muller, the president of the European Association for Cognitive Systems, and Nick Bostrom from the University of Oxford, who published over 200 articles on superintelligence and artificial general intelligence (AGI), conducted a survey of AI researchers. 550 participants answered the question: “When is AGI likely to happen?” The answers are distributed as

10% of participants think that AGI is likely to happen by 2022

For 2040, the share is 50%

90% of participants think that AGI is likely to happen by 2075.

In 2017 May, 352 AI experts who published at the 2015 NIPS and ICML conferences were surveyed. Based on survey results, experts estimate that there’s a 50% chance that AGI will occur until 2060. However, there’s a significant difference of opinion based on geography: Asian respondents expect AGI in 30 years, whereas North Americans expect it in 74 years. Some significant job functions that are expected to be automated until 2030 are: Call center reps, truck driving, and retail sales.

In 2019, 32 AI experts participated in a survey on AGI timing:

45% of respondents predict a date before 2060

34% of all participants predicted a date after 2060

21% of participants predicted that singularity will never occur.

AI entrepreneurs are also making estimates on when we will reach singularity and they are a bit more optimistic than researchers:

Louis Rosenberg, computer scientist, entrepreneur, and writer: 2030

Patrick Winston, MIT professor and director of the MIT Artificial Intelligence Laboratory from 1972 to 1997: He mentioned 2040 while stressing that while it would take place, it is a very hard-to-estimate date.

Ray Kurzweil, computer scientist, entrepreneur, and writer of 5 national best sellers including The Singularity Is Near: 2045

Jürgen Schmidhuber, co-founder at AI company NNAISENSE and director of the Swiss AI lab IDSIA: ~2050

Keep in mind that AI researchers were over-optimistic before

Examples include:

AI pioneer Herbert A. Simon in 1965: “machines will be capable, within twenty years, of doing any work a man can do.”

Japan’s Fifth Generation Computer in 1980 had a ten-year timeline with goals like “carrying on casual conversations”

This historical experience contributes to most current scientists shying away from predicting AGI in bold time frames like 10-20 years. However, just because they are more conservative now doesn’t mean that they are right this time around.

Understand why reaching AGI seems inevitable to most experts

**These may seem like wild predictions, but they seem quite reasonable when you consider these facts**:

Human intelligence is fixed unless we somehow merge our cognitive capabilities with machines. Elon Musk’s neural lace startup aims to do this but research on brain-computer interfaces is in the early stages.

**Machine intelligence depends on algorithms, processing power, and memory. Processing power and memory have been growing at an exponential rate.** As for algorithms, until now we have been good at supplying machines with the necessary algorithms to use their processing power and memory effectively.

Considering that our intelligence is fixed and machine intelligence is growing, **it is only a matter of time before machines surpass us unless there’s some hard limit to their intelligence. We haven’t encountered such a limit yet.**

This is a good analogy for understanding exponential growth. While machines can seem dumb right now, they can grow quite smart, quite soon.

If classic computing slows its growth, quantum computing could complement it

Classic computing has taken us quite far. AI algorithms on classical computers can exceed human performance in specific tasks like playing chess or Go. For example, AlphaGo Zero beat AlphaGo by 100-0. AlphaGo had beaten the best players on earth. However, we are approaching the limits of how fast classical computers can be.

Moore’s law, which is based on the observation that the number of transistors in a dense integrated circuit double about every two years, implies that the cost of computing halves approximately every 2 years. However, most experts believe that Moore’s law is coming to an end during this decade. Though there are efforts to keep improving application performance, it will be challenging to keep the same rates of growth.

**Quantum Computing**, which is still an emerging technology, **can contribute to reducing computing costs** after Moore’s law comes to an end. Quantum Computing is based on the evaluation of different states at the same time whereas classical computers can calculate one state at one time. The unique nature of quantum computing can be used to efficiently train neural networks, currently the most popular AI architecture in commercial applications. **AI algorithms running on stable quantum computers have a chance to unlock singularity.**

#### **AI kills everyone**

Emile P. **Torres**, 8/06/20**22**, Would "artificial superintelligence" lead to the end of life on Earth? It's not a stupid question, Salon, A philosopher and historian whose work focuses on existential threats.

https://www.salon.com/2022/08/06/would-artificial-superintelligence-lead-to-the-end-of-life-on-earth-its-not-a-stupid-question/)// JZ

The activist group Extinction Rebellion has been remarkably successful at raising public awareness of the ecological and climate crises, especially given that it was established only in 2018.

The dreadful truth, however, is that climate change isn't the only global catastrophe that humanity confronts this century. Synthetic biology could make it possible to create designer pathogens far more lethal than COVID-19, nuclear weapons continue to cast a dark shadow on global civilization and advanced nanotechnology could trigger arms races, destabilize societies and "[enable powerful](https://www.sciencedirect.com/science/article/pii/S0016328717301908) new types of weaponry."

**Yet another serious threat comes from artificial intelligence, or AI.** In the near-term, AI systems like those sold by IBM, Microsoft, Amazon and other tech giants could [exacerbate inequality](https://www.aclu.org/news/privacy-technology/how-artificial-intelligence-can-deepen-racial-and-economic-inequities) due to gender and racial biases. According to a paper co-authored by [Timnit Gebru](https://en.wikipedia.org/wiki/Timnit_Gebru), the former Google employee who [was fired](https://www.nytimes.com/2020/12/03/technology/google-researcher-timnit-gebru.html) "after criticizing its approach to minority hiring and the biases built into today's artificial intelligence systems," facial recognition software [is](https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/) "less accurate at identifying women and people of color, which means its use can end up discriminating against them." These are very real problems affecting large groups of people that require urgent attention.

But there are also longer-term risks, as well, arising from the possibility of algorithms that exceed human levels of general intelligence. **An artificial superintelligence, or ASI, would by definition be smarter than any possible human being in every cognitive domain of interest, such as abstract reasoning, working memory, processing speed and so on.** Although there is no obvious leap from current "deep-learning" algorithms to ASI, there is a good case to make that the creation of an ASI is not a matter of if but when: Sooner or later, scientists will figure out how to build an ASI, or figure out how to build an AI system that can build an ASI, perhaps by modifying its own code.

When we do this, it will be the most significant event in human history: Suddenly, for the first time, humanity will be joined by a problem-solving agent more clever than itself. What would happen? Would paradise ensue? Or would the **ASI promptly destroy us?**

**Even a low probability that machine superintelligence leads to "existential catastrophe" presents an unacceptable risk — not just for humans but for our entire planet.**

I believe we should take the arguments [for why](https://www.google.de/books/edition/Superintelligence/C-_8AwAAQBAJ?hl=en&gbpv=1&dq=%22a+plausible+default+outcome+of+the+creation+of+machine+superintelligence+is+existential+catastrophe%22&pg=PA115&printsec=frontcover) "a plausible default outcome of the creation of machine superintelligence is existential catastrophe" very seriously. Even if the probability of such arguments being correct is low, a risk is standardly [defined](https://plato.stanford.edu/entries/risk/#DefRis) as the probability of an event multiplied by its consequences. And since the consequences of total annihilation would be enormous, even a low probability (multiplied by this consequence) would yield a sky-high risk.

**Even more, the very same arguments for why an ASI could cause the extinction of our species also lead to the conclusion that it could obliterate the entire biosphere.** Fundamentally, the risk posed by artificial superintelligence is an environmental risk. It is not just an issue of whether humanity survives or not, but an environmental issue that concerns all earthly life, which is why I have been calling for an Extinction Rebellion-like movement to form around the dangers of ASI — a threat that, like climate change, could potentially harm every creature on the planet.

Although no one knows for sure when we will succeed in building an ASI, one [survey of experts](https://nickbostrom.com/papers/survey.pdf) found a 50 percent likelihood of "human-level machine intelligence" by 2040 and a 90 percent likelihood by 2075. A human-level machine intelligence, or artificial general intelligence, abbreviated AGI, is the stepping-stone to **ASI, and the step from one to the other might be very small, since any sufficiently intelligent system will quickly realize that improving its own problem-solving abilities will help it achieve a wide range of "final goals," or the goals that it ultimately "wants" to achieve** (in the same sense that spellcheck "wants" to correct misspelled words).

Furthermore, one [study from 2020](https://gcrinstitute.org/papers/055_agi-2020.pdf) reports that at least 72 research projects around the world are currently, and explicitly, working to create an AGI. Some of these **projects** are just as explicit that they **do not take seriously** the potential **threats posed** by ASI. For example, a company called 2AI, which runs the Victor project, writes on [its website](http://www.2ai.org/killerai/):

There is a lot of talk lately about how dangerous it would be to unleash real AI on the world. **A program that thinks for itself might become hell-bent on self preservation, and in its wisdom may conclude that the best way to save itself is to destroy civilization as we know it.** Will it flood the internet with viruses and erase our data? Will it crash global financial markets and empty our bank accounts? Will it create robots that enslave all of humanity? Will it trigger global thermonuclear war?… We think this is all crazy talk.

But is it crazy talk? In my view, the answer is no. The arguments for why **ASI** could **devastate the biosphere and destroy humanity**, which are primarily philosophical, are complicated, with many moving parts. But the central conclusion is that by far the greatest concern is the **unintended consequences of the ASI striving to achieve its final goals.** Many technologies have unintended consequences, and indeed **anthropogenic climate change is** an **unintended consequence of large numbers of people burning fossil fuels.** (Initially, the transition from using horses to automobiles powered by internal combustion engines was hailed as a solution to the problem of urban pollution.)

Most **new technologies have unintended consequences**, and ASI would be the most powerful technology ever created, so we should **expect its potential unintended consequences to be massively disruptive.**

An ASI would be the most powerful technology ever created, and for this reason we should expect its potential unintended consequences to be even more disruptive than those of past technologies. Furthermore, unlike all past technologies, **the ASI would be a fully autonomous agent in its own right, whose actions are determined by a superhuman capacity** to secure effective means to its ends, along with an ability to process information many orders of magnitude faster than we can.

Consider that an ASI "thinking" one million times faster than us would see the world unfold in super-duper-slow motion. A single minute for us would correspond to roughly two years for it. To put this in perspective, it takes the average U.S. student 8.2 years to earn a PhD, which amounts to only 4.3 minutes in ASI-time. Over the period it takes a human to get a PhD, the ASI could have earned roughly 1,002,306 PhDs.

This is why the **idea that we could simply unplug a rogue ASI if it were to behave in unexpected ways is unconvincing**: The time it would take to reach for the plug would give the ASI, with its superior ability to problem-solve, ages to figure out how to prevent us from turning it off. Perhaps it quickly connects to the internet, or shuffles around some electrons in its hardware to influence technologies in the vicinity. Who knows? Perhaps **we aren't even smart enough to figure out all the ways it might stop us from shutting it down.**

But why would it want to stop us from doing this? The idea is simple: **If you give an algorithm some task — a final goal — and if that algorithm has general intelligence**, as we do, it will, **after a moment's reflection, realize that one way it could fail to achieve its goal is by being shut down. Self-preservation**, then, **is** a **predictable subgoal that sufficiently intelligent systems will automatically end up with**, simply by reasoning through the ways it could fail.

What, then, if we are unable to stop it? **Imagine that we give the ASI the single goal of establishing world peace.** What might it do? **Perhaps it would immediately launch all the nuclear weapons in the world to destroy the entire biosphere**, reasoning — logically, you'd have to say — **that if there is no more biosphere there will be no more humans, and if there are no more humans then there can be no more war** — and what we told it to do was precisely that, even though what we intended it to do was otherwise.

Fortunately, there's an easy fix: Simply add in a restriction to the ASI's goal system that says, "Don't establish world peace by obliterating all life on the planet." Now what would it do? Well, how else might a literal-minded agent bring about world peace? Maybe **it would place every human being in suspended animation, or lobotomize us all, or use invasive mind-control technologies to control our behaviors.**

Again, there's an easy fix: Simply add in more restrictions to the ASI's goal system. The point of this exercise, however, is that **by using our merely human-level capacities, many of us can poke holes in just about any proposed set of restrictions, each time resulting in more and more restrictions having to be added.** And we can keep this going indefinitely, with no end in sight.

Hence, given the seeming interminability of this exercise, the disheartening question arises: How can we ever be sure that we've come up with a complete, exhaustive list of goals and restrictions that guarantee the ASI won't inadvertently do something that destroys us and the environment? The **ASI thinks a million times faster than us. It could quickly gain access and control** over the economy, laboratory equipment and military technologies. And for any final goal that we give it, **the ASI will automatically come to value self-preservation as a crucial instrumental subgoal.**

**How can we come up with a list of goals and restrictions that guarantee the ASI won't do something that destroys us** and the environment? **We can't.**

Yet self-preservation isn't the only subgoal; so is **resource acquisition.** To do stuff, **to make things happen, one needs resources** — and usually, the more resources one has, the better. The problem is that **without giving the ASI all the right restrictions, there are a seemingly endless number of ways it might acquire resources that would cause us**, or our fellow creatures, **harm.** Program it to cure cancer: It immediately converts the entire planet into cancer research labs. Program it to solve the Riemann hypothesis: It immediately converts the entire planet into a giant computer. Program it to maximize the number of paperclips in the universe (an intentionally silly example): It immediately converts everything it can into paperclips, launches spaceships, builds factories on other planets — and perhaps, in the process, if there are other life forms in the universe, destroys those creatures, too.

It cannot be overemphasized: an ASI would be an extremely powerful technology. And power equals danger. Although Elon Musk is very often wrong, he was right when he tweeted that advanced artificial intelligence could be "more dangerous than nukes." **The dangers posed by this technology**, though, would not be limited to humanity; they **would imperil the whole environment.**

This is why we need, right now, in the streets, lobbying the government, sounding the alarm, an Extinction Rebellion-like movement focused on ASI. That's why I am in the process of launching the Campaign Against Advanced AI, which will strive to educate the public about the immense risks of ASI and convince our political leaders that they need to take this threat, alongside climate change, very seriously.

A movement of this sort could embrace one of two strategies. A "weak" strategy would be to convince governments — all governments around the world — to impose strict regulations on research projects working to create AGI. Companies like 2AI should not be permitted to take an insouciant attitude toward a potentially transformative technology like ASI.

A "strong" strategy would aim to **halt all ongoing research aimed at creating AGI.** In his 2000 article "Why the Future Doesn't Need Us," Bill Joy, cofounder of Sun Microsystems, argued that some domains of scientific knowledge are simply too dangerous for **us** to explore. Hence, he contended, we should impose moratoriums on these fields, doing everything we can to prevent the relevant knowledge from being obtained. Not all knowledge is good. Some knowledge poses "information hazards" — and once the knowledge genie is out of the lamp, it cannot be put back in.

Although I am most sympathetic to the strong strategy, I am not committed to it. More than anything, it should be underlined that almost no sustained, systematic research has been conducted on how best to prevent certain technologies from being developed. One goal of the Campaign Against Advanced AI would be to fund such research, to figure out responsible, ethical means of preventing an ASI catastrophe by putting the brakes on current research. We must make sure that superintelligent algorithms are environmentally safe.

If experts are correct, an ASI could make its debut in our lifetimes, or the lifetimes of our children. But even if ASI is far away — or even if it turns out to be impossible to create, which is a possibility — we don't know that for sure, and hence **the risk posed by ASI may still be enormous**, perhaps comparable to or exceeding the risks of climate change (which are huge). This is why we need to rebel — not later, but now.

#### **Extinction becomes mathematically inevitable as technology advances and society complexifies, catalyzing total system failure.**

**Manheim 20** [Dr. David Manheim, PhD in Public Policy, head of research and policy at the Association for Long Term Existence and Resilience (ALTER); September 2020; Futures, vol. 122, https://doi.org/10.1016/j.futures.2020.102570]

1. Introduction

**The risk of humanity's possible extinction could take various forms**, from diseases to runaway climate change to asteroid strikes. It is widely agreed that either extinction or global catastrophic risks leading to most humans dying would be bad, even if the exact relative importance of avoiding extinction is unclear. Schubert, Caviola, and Faber (2019) **It does seem to be the case**, however, **that technology** or other modern developments **are far more likely to cause human extinction than natural events**. Snyder-Beattie, Ord, and Bonsall (2019)

Bostrom has argued that the prevention of potential extinction is a current moral priority. Bostrom (2013) Because of this, Bostrom has researched potential sources of extinction risk extensively. Some past technologies are understood to potentially have such devastating consequences, such as nuclear war. Newer concerns include malicious **artificial superintelligence**, Pistono and Yampolskiy (2016) **malicious use of biological engineering**, Gronvall (2016) and technologies that could destabilize the international balance that makes the use of nuclear weapons less likely. Geist and Lohn (2018) In each case, the risks of these technologies are being investigated by an active research community.

Bostrom recently suggested another possibility, **the “Vulnerable World Hypothesis,”** Bostrom (2019) where **a technological advance occurs that will by default lead to extinction.** He suggests that this is concerning, but notes that it is unclear if there are in fact such technologies in humanity's future. It is far more clear, however, that **there are already technologies or events which would lead to global catastrophic risks** short of extinction, **and it seems implausible that more will not be found. Such events would create** widespread devastation, could at least severely limit humanity's potential future advancement, and would at the very least cause **tremendous human suffering and death** in the short term. Bostrom, 2003; Bostrom (2003), Parfit (1984)

The class of existential risks addressed in **this paper** is related to several of the previous models, especially Bostrom's vulnerable world hypothesis, but **focuses on systemic risks** rather than individual technologies. Systemic risks are **those that emerge from complex system failure, where the failure of a single component leads to systemic knock-on effects. These** are potentially devastating, and we will suggest that they might **lead to an inevitable extinction** of the type Bostrom suggests. Not only that, but because of the way these risks emerge, they seem more difficult to classify or estimate than typical risks. Tonn and Steifel suggest methods for estimating existential risks which are primarily suitable for risks that are not systemic or unknown. Tonn and Stiefel (2013) These methods are unfortunately appropriate for neither type of systemic risk discussed here, and work on understanding and estimating these risks is an open and important problem. Still, Baum suggests that “threats are rarely completely unknown or unquantifiable,” Baum (2015) and while he notes that risk-based analyses are limited, resilience as a paradigm is useful for cases like this where the risk is underestimated, unknown, or unquantifiable.

Daniel Schmachtenberger suggests that there is a dichotomy of two generating processes that can lead to human-induced existential risk. Schmachtenberger and Thorson (2019) The first, where rivalrous and uncoordinated actors combined with “exponential technology” lead to either collapse, or anti-rivalrous solutions, parallels Bostrom's Vulnerable World Hypothesis scenarios. The second, where complex systems become increasingly complex and fragile, is the case discussed in this paper. We will seek to expand on this, and more clearly describe why these systemic failures are particularly dangerous. In order to discuss this, the paper first reframes Bostrom's argument by replacing the analogy he uses with a more complex but also more complete explanation. This reframing is used to show how the systemic risk cases discussed in this paper are similarly worrying. Following this, the paper presents arguments for why systemic risks and fragility could make complex sociotechnical systems into fragile worlds, and then presents an additional sub-hypothesis about such systems which would entail Bostrom's vulnerable world hypothesis, focused on path dependence, as will be explained. Finally, the paper presents a number of ways in which the expanded view of vulnerable worlds leads to a number of conclusions which are strikingly different from, or even exactly contrary to, the conclusions Bostrom suggested. 2. Novel Technologies Reframed as an Explore/Exploit tradeoff Bostrom's recent “Vulnerable World Hypothesis” lists several ways in which “there is some level of technological development at which civilization almost certainly gets devastated by default, i.e. unless it has exited the 'semi-anarchic default condition.”’ He illustrates this with an analogy modeled on a classic type of explanation in probability theory, that of an urn from which we draw balls. In his analogy, each ball represents a technological advance, and the balls can be white, black, or any shade of gray – representing the impact on humanity. Since we do not know what the future of technology will bring, the distribution of the colors of the balls inside the urn is unknown. He then discusses four scenarios which posit the existence of a “black ball” technology drawn from the urn, which, unlike those found so far with positive effects or mixed effects of humanity, “invariably or by default destroys the civilization that invents it.” Bostrom admits that his analogy to an urn is simplistic, though it is compelling. However, for our argument it is insufficient. Instead of an urn containing discrete technologies, the process of scientific discovery that occurs when certain advances pose risks can be understood with a different parallel from probability theory, one more often used in machine learning, that of the explore-exploit tradeoff. To explain this, we can consider the multi-armed bandit problem, where a gambler possesses some finite number of coins, and is faced with a row of slot machines, each of which has a potentially different payout distribution. For each coin, the gambler is faced with a choice of which machine to play. A conservative gambler might “explore” a bit and try a few machines, and notice a machine that pays out 2:1 just over half the time, and stay there to “exploit” the machine, slowly but steadily making money, uninterested in further exploration. A more adventurous gambler would instead spend more time exploring the other machines, perhaps seeking a machine that pays out less often, but with a far higher reward. If no such machine exists, they lose out from their continued exploration - but if there is a machine they find which pays out, say, 50:1 25% of the time, they will make far more money. Explore-exploit tradeoffs are also discussed in somewhat different terms in machine learning, by analogy to a landscape, rather than an urn or a multi-armed bandit. Here, imagine a robot, which in our analogy will parallel human civilization, in a potentially dangerous environment. There are rewards, that is, new technologies, scattered about. At each stage, the robot can explore in any direction to search the landscape for a new reward (novel technologies), or focus on places where it knows rewards can be found. Because the payoffs of future exploration are unknown, optimal decision making in explore-exploit situations is notoriously hard to analyze. In fact, when the multi-armed bandit problem was first proposed during World War II, allied analysts suggested that perhaps the most useful application was to give the problem to the Axis, so that they would also waste their time on the seemingly intractable problem.Gittins (1979) Unlike the gambler, whose worst outcome is sub-optimal returns on their investment, in our scenario we suppose that the robot might accidentally become trapped or be destroyed by some previously unknown danger. If Bostrom's hypothesis is correct, civilization's exploration of new technologies comes with some probability of a fatal result, and some probability of discovering a highly rewarding new technology. In this new context, the goal is to find a search strategy that maximizes the total benefit humanity ever receives, what Bostrom has called humanity's cosmic endowment. Bostrom (2003) Any strategy, however, risks what Bostrom called astronomical waste - either embracing “technological relinquishment,” by being so risk-averse that advanced technologies that greatly benefit humanity are never explored or developed, or being so risk-accepting that Bostrom's postulated Vulnerable World becomes inevitable, losing the entire potential future benefit to humanity. The first critical question addressed by Bostrom - “Is there a black ball in the urn of possible inventions?” is, to reframe the question, about the existence of fatal dangers in our Robot's exploration landscape. If we could answer that question in the negative, this would seem to refute the informal hypothesis he proposes of a vulnerable world. This is not the only refutation he suggests, however. Bostrom's suggestion of “differential technological development” is, in our terms, to cordon-off or minimize exploration of sections of the landscape or directions most likely to contain fatal traps. If this can be done, it would refute the suggestion that technological development inevitably leads to increased risk of choosing a devastating technology. Falsifying Bostrom's initial hypothesis does not, however, show that the world is not vulnerable in other ways. The fuller statement of the hypothesis in Bostrom's paper is that “[i]f technological development continues then a set of capabilities will at some point be attained that make the devastation of civilization extremely likely, unless civilization sufficiently exits the semi-anarchic default condition.“ This statement, I will argue, cannot be falsified even by proving the impossibility of “a technology that invariably or by default destroys the civilization that invents it”1 . In addition to Bostrom's four vulnerable worlds, I suggest a fifth possibility - that the simple accumulation of white and/or gray balls drawn from the urn can itself lead to fragility and, without strong forces pushing in the opposite direction, collapse of civilization by default. To frame it in our explore/exploit terms, the fifth possibility is that as the robot explores, it may dig too deep, or alter the landscape itself incautiously so that further movement could cause a collapse. The primary way for this to happen which this paper discusses is fragility - in the analogy, our exploration could lead to the equivalent of an avalanche or mine-collapse brought on by incautious behavior. That is, the creation of fragility in the landscape means that a lack of caution when exploring further could end up burying us under the weight of accumulated technologies. 3. Fragility and Systemic Risk

**As technologies develop, they often build on one another, so that** the continued operation of **the system depends on a growing set of other systems**. The way that failure occurs and propagates in such systems is non-obvious, but it is largely dependent on the topology of the interdependence between components. Pastor-Satorras, Castellano, Van Mieghem, and Vespignani (2015) Simple dependencies, where a system requires the functioning of every component, can make the resulting system of systems more fragile than the components. A very basic model of this shows that given a system-of-systems with N components, each of which independently can fail at the rate Fi the failure rate is

. While this grows more slowly than the sum of the individual failure rates as new systems are added, it is also far higher than the average failure rate of those individual systems.

**As a concrete example**, the peak of **efficient farming once required family farms** to depend on a family to manage the farm, **a blacksmith** to make horseshoes and plows, **and draft animals** to pull them. **Losing any one** of these **would be enough to** (eventually) **make the system unable to continue**, and there was some risk that this would occur. **Still, the limited number of inputs** and the substitutability of other inputs **made such systems fairly stable** - especially because many risks were uncorrelated across farms, and could often be addressed by borrowing from other farms nearby.

**The farms of today**, of course, **are not nearly as simple. They require** everything from satellite **GPS** systems to pinpoint locations, to the **semiconductors** and **fabrication plants** used to make specific integrated chips used in the farm equipment, to **internet connectivity** to run **machine learning algorithms** using collected data **and satellite imagery on remote servers**. Zubarev, Fomin, and Zubarev (2019) Modern agribusiness depends on everything from finding and hiring skilled laborers to manage complex machinery, to managing regulatory, financial, and other factors critical to farm operation. Kingwell (2011) These are often more tightly correlated across an economy, increasing risk. Beyond that, managing these farms requires understanding “human, technical, economic, financial, risk, institutional and social” issues, as Lewis et al. noted more than a decade ago. Lewis, Malcolm, and Steed (2006)

**The risk is likely not yet critical, but it seems clear that dependence is growing**, and the ability to use backup systems can be lost. For example, if remote servers become unavailable, local corn farms may lack the information needed to decide where to increase and decrease watering levels, or even lose the ability to run their computer-controlled irrigation systems. Similarly, decades ago supplies and ordering were managed on paper, and now, without the servers running Software-as-a-Service supply-chain-management software, the dairy farm down the road may not have access to an inventory of their supplies or know what amounts of products are needed or what they have historically ordered, and end up unable to feed their cows.

The inter-dependencies in such systems are more complex than the above model allows, but more complex analyses, such as those employing percolation analysis to understand mutual interdependency of multiple networks, Buldyrev, Parshani, Gerald Paul, Stanley, and Havlin (2010) show the same trend. That is, interdependent systems where failures can propagate can be far more likely to fail than the average rate at which the individual systems’ components fail. Worse, analysis of “high–value, technology– and engineering–intensive products or systems…used to produce consumer goods and services” has shown that the failure rates are nearly-additive, and worse, are hard to identify. Yeo and Ren (2009)

It should be noted that modern computer networks do not display such fragility, but this is a function of intentional design. Metcalfe and Boggs (1976) Simple network structures, such as lines or rings, are far more prone to failure Clark, Pogran, and Reed (1978). That is, unless a system was designed for resilience, resilience should not be expected. And when technological systems are made efficient and complex, they tend to be tightly coupled - meaning that failure in one place spreads Bookstaber (2007).

3.1. Inevitable Technological Fragility Hypothesis

The proposal of this paper, to provide an addendum to Bostrom's hypothesis, is that **if technological development continues indefinitely, systemic fragility will increase to the point that the possibility of a shock sufficient for complete collapse approaches certainty.**

This hypothesis rests on a number of assumptions, but there are also a variety of reasons to find it plausible and concerning. To lay these out clearly, we will first consider the question of how and why individual systems are fragile, then make an argument that it is at least plausible that **the multiple interconnected systems** and systems-of-systems **which are necessary for** much of **modern civilization not only fail to address this risk, but multiply it**.

4. Single-System Complexity and Fragility The key question so far is whether fragility increases over time as systems are built. The answer to that question depends on a combination of factors that can push in either direction. These include increasing complexity of systems, the economic incentives for efficiency over robustness and the resulting levels of investment in resilience, the failure rates of individual components and systems, as well as the way in which systems-of-systems (and systems-of-systems-of-systems) are interrelated, and the extent to which systems and their interdependencies are designed to be robust. Even the claim of inevitable fragility in individual systems makes several assumptions about how fragility increases. Before looking at the systemic question of how fragility could lead to collapse, we will outline these assumptions. Note that these are in fact assumptions, rather than claims - if any one of them is false in ways that are outlined, it would refute the hypothesis. The third assumption is particularly critical, and will be explored further in the next subsection. First, for fragility to matter, the current trend of efficiency-increasing and resilience-decreasing technologies must continue to apply to at least one critical system, such as agriculture, communication, or transport. If this is wrong, and future white-ball / safe exploration technologies are ones that favor robustness over efficiency in all such critical domains, the trend would reverse. For instance, distributed fault tolerant computing arguably increases both efficiency and robustness. Most new technologies move in the opposite direction, but if enough resilience increasing technology is found, the balance could shift. Second, the argument for increasing fragility assumes that economic growth continues to absorb human effort in a way that does not lead to overabundant resources. In Eric Drexler's ’Paretotopia‘ scenario, increased resources are unmatched by increased demand. Drexler (2019) In that future case, resources are abundant enough that robustness is easy to achieve. This second scenario also assumes the absence of supercharged competition that uses the newly abundant resources. This would not occur, for example, in Hanson's proposed default “Em” scenario, where human-based intelligences are simulated computationally, leading to a reduction rather than an increase in surplus that could be redirected to robustness for lack of other needs. Hanson (2016) Third, it assumes that fragility is relatively hard to identify, such that at least some failures will be unanticipated. This has been true historically, but it is possible that future developments would reverse this trend, making the search for increased robustness itself efficient enough to counterbalance the more general and destabilizing increased fragility that new technology allows. If failures do become easy to anticipate, more expensive general resilience can be replaced with more specific redundancies targeted to the exact failure modes identified. 4.1. Non-Obvious Fragility As mentioned, hard-to-identify fragility is a key assumption. Broadly speaking, non-obvious fragility is the result of planning for efficiency, instead of designing for redundancy, fault tolerance, or even provable safety. This is a fairly general fact about any control system. Paattilammi and Makila (2000) The concrete result of the current optimization shows clear signs of producing fragile results. One example is the proliferation of disposable technology, such as fragile smartphones designed to be replaced rather than fixed or upgraded. Failure of these optimized devices is normal, and while mitigating failure is important, it is often the case that risk must be accepted, rather than avoided. Perrow (2011) This type of fragility is obvious and anticipated, rather than non-obvious and worrying. For example, individual computers are fragile, and components fail frequently. For this reason, in high-reliability computer systems, a variety of mechanisms are in place to compensate, including redundant online systems for data storage, Chen, Lee, Gibson, Katz, and Patterson (1994) or methods to address other hardware failures. Wang, Zhang, and Xu (2017) The fact that computer networks are not fragile, and the fragility that does exist is well understood, seems to be a counterexample. But the resilience itself is planned, in contrast to ecological systems where it is emergent - as we will discuss in detail below. This means that fault-tolerant designs are built to be tolerant of expected faults. Not only that, but resilience itself is optimized, for example, to minimize the number of backups or other costs needed to have a planned level of reliability. Rodrigues-da Silva and Crispim (2014) This creates fragility to unexpected faults, and allows the systems to operate through anticipated contingencies, but not to anything beyond that point. 4.2. Sociotechnical Resilience Fragility of systems is not based purely on the lack of resilience of technical systems. In fact, fragility of optimized technical systems is compensated for by the greater robustness of sociological systems. The combined sociotechnical system, then, is the level at which fragility should be considered. To reduce the fragility of sociotechnical systems, organizations can attempt to build more resilience at the organizational level. This can involve information sharing, distributed decision making, and better risk assessment. If done well, these attempts provide a sociotechnical system that compensates for technical and operation risk, but is again very different from emergent resilience. Langeland, Manheim, Mcleod, and Nacouzi (2016) Unfortunately, the interaction between humans and technology can often multiply risks, rather than mitigate them. Yeo and Ren (2009) Another reason to think that sociotechnical resilience will not fully compensate for technological fragility is the reduced human involvement in technical systems. As automation increases, Danzig notes that humans are increasingly necessarily out-of-the-loop. Danzig (2018) He further argues that when there is competition, this dynamic is a necessary result of continued optimization. To conclude the discussion of single-system fragility, we note that inevitable fragility of systems is not actually required for the hypothesis presented. As this section argued, it does seem plausible that in expectation, new technologies will be more fragile than those they replace. However, systemic risk can exist given the much weaker claim that specific critical systems are relied upon, and technological improvements relevant to those systems alone exhibit sufficient fragility to cause a cascading collapse. Before discussing the interaction between systems, however, it is worth considering how these human, technical, or sociotechnical systems differ from naturally resilient biological systems. 4.3. Contrasting (fragile) sociotechnical systems with (resilient) biological systems There is an extensive literature on resilience and complexity in biological systems that contrasts greatly with the arguments in this paper. The analogy between complex systems in biology and complex technological systems is apt in many ways. For example, free markets are often understood as evolutionary, with businesses engaged in Darwinian competition. We will explain, however, that this provides reason to think that the comparison reinforces some of the arguments, rather than solely contradicting them. Resilience was arguably first discussed in ecology by Holling. Holling (1973) In his paper, it is clear that resilience itself is an emergent property of the complex ecological system. Notably, however, the premise of the paper is that at the time, ecology and related disciplines were analyzed along similar lines to technical systems. He suggests that this parallel is unfortunate, because ecological systems fundamentally differ. Engineered systems are complex in ways similar to biological systems. However, even small differences lead to very different higher level properties, which is a general fact about complex or emergent systems. One such difference which is critical relates to the diverse redundancy of ecological systems, where many different species can fill each niche. Technology usually relies on a single system, with a backup that is nearly if not completely identical. Holling made the argument for the difference between the two classes of system clearly, calling for people not to confuse resilience as discussed in ecology and resiliency as discussed in engineering. Holling (1996) To expand on this point, Holling has more recently argued, in Peterson et al., that resilience is a function of ecosystem diversity, with overlapping functions and redundancies across scales. Peterson, Allen, and Holling (1998) In contrast, engineered systems typically reduce redundancy to increase efficiency. For example, a key software design principle is that anything which performs the same task should be “refactored” to use a single function rather than having redundancy. Kerievsky (2004). This will increase maintainability, and likely lead to fewer bugs, but will also increase correlation of failures and vulnerabilities. While there are design patterns that take the opposite tack, Chen and Avizienis (1995) they are not in common use. Another key difference is the type of redundancy, as mentioned above. While having multiple organisms to fill a niche, or even individuals of the same species can increase resilience due to redundancy, having multiple copies of software in a system typically has no benefits, Point (2018) and if anything the necessity to replace or patch software in multiple places makes maintaining a system, or recovery after a failure, even harder. It could be contended that the layer above that of sociotechnical systems is where resilience will emerge. Markets create competition that can provide another layer of redundancy, and it could be argued that they create efficiency in ways similar to evolutionary competition. Unfortunately, there are clear arguments that markets may not evolve towards greater resilience. Tisdell (1999) In addition to theoretical arguments, there are empirical facts that have emerged to support the claims. Global markets are consolidating rather than becoming more diverse, and firms are often collaborating (if not colluding), rather than competing. As banks showed in 2007, large firms with complex businesses can create systemic risks of interdependence, rather than robustness from diversification. Laeven, Ratnovski, and Tong (2014) Recent research shows that this is true not only in finance, but more broadly. Welburn et al. (2020) This implies that systemic risks are in fact critical. 5. Larger Scale Risks from Fragility Failures happen, and as outlined above, we can see at least one plausible future where failures due to fragility cannot be eliminated. The idea that they could lead to global catastrophes or existential risks is perhaps more surprising, and requires a stronger claim than inevitably increasing fragility of individual systems outlined earlier. For that reason, we will now lay out some preconditions for increasing systemic fragility that would imply that larger scale vulnerabilities are also inevitable. • First, larger scale risks leading to collapse requires that technological advances enable or require complex systems that cannot be replaced with simpler alternatives, either because of path-dependent technological dependencies, or lack of availability of such alternatives. • Second, collapse risk requires that failures are rare enough that technological growth allows building systems far more complex than are maintainable or replaceable before failure occurs. • Third, collapse risk requires that failures either exhibit a domino effect, where systems are reliant on one another in ways that causes a chain reaction of failure, or that failures are correlated due to all being reliant on a single point of failure. Despite the above required confluence of uncertain claims, it seems that each of the claims is not only plausible, but at least somewhat supported by evidence. To explain how, it is worth looking at each, and considering how the incentives in the system work, and past experience with how such systems develop. Finally, we will note some ways in which these factors interrelated and increase risk further via incentives for suboptimal mitigation, and the increased risk of moderate investments in resilience. 5.1. Path Dependence and Irreversibility: There's No Way Out But Forward Path dependence leading to irreversible complexity is another oft-seen dynamic in complex systems2 . As the earlier example in agriculture noted, most real-world systems have a limited ability to roll-back to earlier states. Older equipment is discarded, and prior techniques are lost when older workers leave. Even for software, where neither experience or equipment is typically needed to go back to earlier systems, there are limitations in rolling back a single protocol or system once other related systems depend on them3 . Such software obsolescence is a well-understood problem. Sandborn (2007) As a concrete example of how this occurs, software programs cannot always be run on newer versions of operating systems. If you want to run an older program, often the only solution is to run an emulation, i.e. depend on the newer system. Similarly, older operating systems cannot typically run on newer hardware, since they do not have appropriate device drivers for new hardware. And older hardware cannot easily be manufactured, as production lines were long ago scrapped. 5.2. Momentum and Cumulative Complexity

**Technology evolves rapidly, and failures are not often seen.** Despite this, problems often persist for years, and even critical software security issues remain undiscovered and can last through many versions and for many years. Ablon and Bogart (2017) Not only could a cause of failure be present in older versions, but it may last to the point where software without the problem is obsolete. This implies that a system might not fail a single step past where it is safe, where rollback to less fragile technology is only difficult, but that it instead ends up 4 or 5 steps past that point, so that replacing systems with earlier versions in time to prevent collapse is completely non-viable.

A related issue is that complexity and failures do not always result from a single system. Often, it is the interplay between systems that creates issues. The entire discipline of systems engineering exists in large part to ensure different systems can work together, and it is rarely a trivial problem. The problems of integrating pairs of systems is only made harder for systems-of-systems with dozens or hundreds of component systems.

When these systems of systems are designed in concert with one another, the problems are managed. More often, the parts were designed fully independently, and engineers build processes that depend on the systems working together. Of course, these systems-of-systems are themselves often then connected. Tonn et al. discuss “Earth Scale convergence systems” Tonn et al. (2013) and note that maintaining such earth scale systems is “a constant and seemingly growing challenge.“ This is not coincidental, or only due to a lack of planning. Instead, the size itself creates challenges.

5.3. Growing complexity and interconnection

A single system-of-systems failure might not lead to collapse, even if it did pose a global catastrophic risk. However, the benefits of efficiency means that firms and individuals will also seek to subcontract or outsource to specialized firms despite the risks of dependency, and create additional interconnections and dependencies. This is already seen as a large source of risk in global supply chains, where companies are exposed to risks by suppliers of suppliers, and those companies’ suppliers as well.

An example of this type of fragility which surprised the automobile industry was the fact that despite attempts to mitigate their supply chain risks for components, they discovered that all of the various suppliers of metallic color paints for their cars actually ordered their pigments from a single factory which was shut down by the 2011 Japan Tsunami. Sheffi and Lynn (2014) While failures of this sort are uncommon, further research found that the dependencies which allow them are typical. Research looking at consumer goods, health care, and manufacturing firms found that between 12% and 13% of the firms’ first-degree suppliers depended on the same second-degree suppliers. Wang, Li, and Anupindi (2015) Similar analyses in other complex domains like software package dependencies show similar shared reliance. Decan, Mens, and Claes (2017)

**The growing interconnection of the world, and the growing dependence on technologies, is likely only going to increase as more** and more new **technologies** that improve human lives and economic success **become available**. The fragility of these systems, of course, will also grow as systems grow ever more efficient, and interconnected. If such systems are fragile, however, people will be motivated to reduce risks. As we will argue next, this is insufficient.

5.4. Economic Incentives for Failure

A penultimate point is that systems will not only continue to have some risk of failure, but will also systemically underinvest in reliability. A straightforward microeconomic argument provides a reason to expect this. Experience with safe engineering has led to a clear principle in engineering that for any given project, there is a cost/safety tradeoff. A system can always be made safer, but past a certain point, it is not reasonable to pay for marginal improvements in safety. As argued above, this tradeoff by default leads to not only unsafe individual firms, but also to increasing systemic fragility. This implies that systemic resilience is a public good, in the economic sense.

**A sufficient level of investment in** this **public good is unlikely**, as is typical for public goods. This parallels Bostrom's example of climate change, where individual actors each contribute to the problem. Bostrom (2019) To see why, we note that resilience and robustness in self-governed systems is limited by a local/global risk tolerance mismatch. **Firms will obviously invest in risk mitigation, but the extent to which they do so is limited by their economic benefit.**

This is lower than would be hoped for in general because **the worst outcomes possible for an individual firm** or individual, **of failing completely, imposes additional negative externalities on the wider system.** That is, because a firm or individual has downside limited to its own existence, the amount of mitigation that they find to be valuable in expectation will be lower than the level which is socially optimal. **This points to the economic “bad” of complexity and fragility** - i.e. an economic good which is harmful rather than beneficial, parallel to pollution in Bostrom's example of climate change. For this reason, unless we see that firms and individuals are more risk-averse about systemic risks than their local preferences imply are rational, we expect that local decisions towards sufficient robustness will undervalue robustness, and instead be more complex and fragile than ideal.

5.5. Failure Correlation

A number of arguments have now been put forward as to why systems would fail, and these failures would likely be large and pose at least global catastrophic risks. However, the risk of any one system-of-systems failure at a given time is high enough that cumulative complexity and momentum is somewhat unlikely, as long as these failures are uncorrelated, and the confluence of several such failures is then vanishingly unlikely. For that reason, the hypothesis of inevitable complete collapse depends on growing correlation of failures, or an increased risk of larger failures.

Correlated failures can result from having a single point of failure across systems. As noted above, outsourcing is more efficient than distributing tasks, and firms will attempt to find ways to outsource to the lowest cost, most efficient provider or system. Because of economies of scale, it will sometimes be the case that a natural monopoly will form for certain services or products, and when this occurs, any failures will occur across the economy rather than locally. This is not only speculative. Failures in key services already have ripple effects across many industries, as in the case of Salesforce, a widely used customer contact system, which led to a brief shutdown of sales and related activities across industries when it went offline for even a short time. Claburn (2019)

A similar phenomenon occurs in computing, as noted in analyses of cybersecurity risk. Geer, Jardine, and Leverett (2020) Operating systems for servers, for example, have a market dominated by Linux. Desktop computer operating systems are similarly dominated by Microsoft Windows. There are a variety of reasons that such near-monopolies exist, but they are found often. In the same way, many key services for computer systems are complex, and only a few implementations exist. Bugs or failure points that are found in such key services are therefore not always confined to a single system, but are actually universal, by virtue of the universal usage of a single point of failure.

Even where no monopoly exists, convergent needs and the use of related techniques across an industry can lead to convergent failure modes, as was found in the case of Spectre and related security vulnerabilities for computer processors. In that case, all processors which used the general technique of branch prediction for speculative execution were vulnerable to a set of very closely related flaws. Johnson and Davies (2019) The shared failure mode was not due to a single system being relied on, but rather to the fact that a single method was used across the industry.

A final argument can be made for collapse risk rather than simply danger from individual failures that relates the systematic underinvestment to the risk of overly complex systems. In short, rare failures can create additional fragility. The argument is that if systemic fragility is identified due to failures, the larger risks will be apparent. Less frequent failures, as argued above, will lead to growth in complexity and greater fragility. The fact that firms are motivated to invest moderate amounts in reducing fragility is likely to make failures less frequent, but not eliminate the risk - so that when failures occur, they will be failures of more complex and more closely inter-related systems.

6. Conclusion and recommendations

The paper argues that **there is a significant and growing risk of global catastrophe due to technological complexity**, and the resulting fragility of systems. **Individual actors** (at the company, state, or regional level) may **benefit from technological races that promote economic growth over systemic safety and robustness**, but the growing interdependence of international systems makes this risky. This implies that continuing **the current trend of investment based primarily on the promised advantages of new technologies is a significant concern**. The paper then presents a hypothesis that **this is an inevitable result of** a certain type of **technological innovation.** If the hypothesis is true, it would mean that continued technological innovation leads to what Bostrom refers to as a “vulnerable world,” one **that** inevitably **leads to catastrophe.**

#### **Core degasation.**

**Turchin 22** [Alexey Turchin, B.S. in Physics from Moscow University, National Scientific Complex IASA in Kyiv State Polytechnical Institute; May 21, 2022; “The Future of Nuclear War,” https://www.lesswrong.com/posts/G6JMpyMeMyAsNnkTj/the-future-of-nuclear-war]

**Core degasation as** a **Doomsday weapon**

The nuclear-powered probe **could be used to reach Earth's core: a very hot and heavy nuclear reactor could melt through the crust.**

The dangers of such probe and of the core’s degasation were discussed by Circovic in “Geoengineering gone awry” (Cirkovic & Cathcart, 2003). **Such probe may cause degasation of the core, which will kill all life on Earth.** Thus, **such a probe could be used as a Doomsday weapon** for global blackmail.

#### **Quantum vacuum mining.**

Tim **Folger**, 7/18/20**08**, Nothingness of Space Could Illuminate the Theory of Everything, Discover, https://www.discovermagazine.com/the-sciences/nothingness-of-space-could-illuminate-the-theory-of-everything)// JZ

When the next revolution rocks physics, chances are it will be about nothing—the vacuum, that endless infinite void. In a discipline where the stretching of time and the warping of space are routine working assumptions, the vacuum remains a sort of cosmic koan. And as in the rest of physics, its nature has turned out to be mind-bendingly weird: **Empty space is not really empty because nothing contains something, seething with energy and particles that flit into and out of existence.** Physicists have known that much for decades, ever since the birth of quantum mechanics. But only in the last 10 years has the vacuum taken center stage as a font of confounding mysteries like the nature of dark energy and matter; only recently has the void turned into a tantalizing beacon for cranks. As one blond celebrity heiress and embodiment of emptiness might say, nothing is hot.

To investigate the mysteries of the void, some physicists are using the biggest scientific instrument ever built—the just-completed Large Hadron Collider, a huge particle accelerator straddling the French-Swiss border. Others are designing tabletop experiments to see if they can plumb the vacuum for ways to power strange new nanotech devices. “The vacuum is one of the places where our knowledge fizzles out and we’re left with all sorts of crazy-sounding ideas,” says John Baez, a mathematical physicist at the University of California at Riverside. Whether in the visionary search for the engine of cosmic expansion or the near-fruitless quest for perpetual free energy, **the vacuum is where it’s happening. By mining the vacuum’s riches, a true theory of everything may yet emerge.**

Empty space wasn’t always so mystifying. Until the 1920s physicists viewed the vacuum much as the rest of us still do: as a featureless nothingness, a true void. That all changed with the birth of quantum mechanics. According to that theory, the space around a particle is filled with countless “virtual” particles rapidly bursting into and out of existence like an invisible fireworks display.

Those virtual quantum particles are more than a theoretical abstraction. Sixty years ago a Dutch physicist named Hendrik Casimir suggested a simple experiment to show that virtual particles can move objects in the real world. What would happen, he asked, to two metal plates placed very close together in a complete vacuum? In the days before quantum mechanics, physicists would have said that the plates would just sit there. But Casimir realized that the net pressure of all the virtual particles—the stuff of empty space—outside the plates should exert a minuscule force, a nudge from nothing that would push the plates together.

Physicists tried for decades to measure the Casimir force with great precision, but it wasn’t until 1997 that technology caught up with theory. In that year, physicist Steve Lamoreaux, now at Yale, managed to detect the feeble Casimir force on two small surfaces separated by a few thousandths of a millimeter. Its strength was about equal to the force that would be exerted against the palm of one’s hand by the weight of a single red blood cell.

At first most physicists regarded the Casimir force as a quantum oddity, something of no practical value. Now that has changed: Forward thinkers see it as an important energizer for the tiniest of machines, devices on the nano scale, and a few labs are working on ways to use the force to defy the conventional limitations of mechanical design. Federico Capasso, a physicist at Harvard, leads a small team that is trying to create a repulsive Casimir force by tinkering with the shapes of plates or with the coatings used to cover them. His entire set of experiments fits on a desktop, and the objects he works with are so small that most of them cannot be seen without a microscope.

“Once you have a repulsive force between two plates, you should be able to eliminate static friction,” Capasso says. That could lead to a host of useful applications, including tiny frictionless bearings or nanogears that spin without touching. “But the experiments are enormously difficult, so I cannot tell you when and how.”

For all its strangeness, the Casimir force may be the one property of empty space that does not baffle today’s physicists. It is garden-variety quantum mechanics, weird but not unexpected. The same can’t be said about dark energy, a truly astonishing discovery made by astronomers a decade ago while observing distant exploding stars. The explosions revealed a universe expanding at an ever-faster rate, a finding at odds with previous expectations that the expansion of the cosmos should be slowing down, braked by the collective gravitational pull of all the matter out there. Some unknown form of energy—physicists call it dark energy simply for lack of a more descriptive term—appears to be built into the very fabric of space, countering the gravitational pull of matter and pushing everything in the universe apart. Some theorists speculate that dark energy might cause a runaway expansion of the universe, resulting in a so-called Big Rip some 50 billion years from now that would tear the cosmos to pieces, shredding even atoms.

The observations have allowed physicists to estimate the quantity of dark energy by deducing the force needed to produce the accelerating effect. The result is a minuscule amount of energy for every cubic meter of vacuum. Since most of the cosmos consists of empty space, though, that little bit adds up, and the total amount of dark energy completely dominates the dynamics of the universe.

With the discovery of dark energy came difficult questions: What is this energy, and where does it come from? Physicists simply do not know. According to quantum mechanics, **the energy of empty space comes from the virtual particles that dwell there. But when physicists use the equations of quantum theory to calculate the amount of that virtual energy, they get a ridiculously huge number—about 120 orders of magnitude too large. That much energy would literally blow the universe apart: Objects a few inches from us would be carried away to astronomical distances; the universe would literally double in size every 10-43 second, and it would keep doubling at that rate until all the vacuum energy was gone.** This may be the most colossal gap between observation and theory in the history of science. And it means that physicists are missing something fundamental about the way the universe works.

“We’ve made a prediction on the basis of our best theories, and it is wrong, wildly wrong,” says Sean Carroll, a theoretical physicist at the California Institute of Technology. “That means we don’t just tweak a parameter here and there; we really have to think deeply about what our theories are.”

Even if no one knows where the energy of empty space comes from or why it has the value it does, **there is now no doubt that it exists. And if there is energy to be had, there is inevitably somebody out there thinking of how to exploit it. The notion of limitless energy from empty space has inspired legions of wannabe physicists** who dream of developing the ultimate perpetual-motion device, a machine that would solve the world’s energy problems forever. A quick Internet search for the words free energy and vacuum turns up pages and pages of schemes for tapping the vacuum’s energy. I ask John Baez if such efforts are as hopeless as previous perpetual-motion machines. Are they equally crazy and doomed to failure?

“Perhaps not as doomed as trying to prove the world is flat,” Baez says. “One thing I can say is that I sure hope it doesn’t work, because if you could extract energy from the vacuum, it would mean that the vacuum is not stable. For normal physicists,” he adds with a laugh, “the definition of the vacuum is that it’s the lowest-energy situation possible—it has less energy than anything else.” In short, Baez says, while we may be able to get energy from the vacuum, success “would mean the universe is far more unstable than we ever dreamed.”

The reasoning goes like this: **If the vacuum is not at the lowest energy state possible, then at some point in the future, the vacuum could fall to a lower state, pulsing out energy that would threaten the very structure of the cosmos. If some clever engineer were ever to extract energy from the vacuum, it could set off a chain reaction that would spread at the speed of light and destroy the universe.** Free energy, yes, but not what the inventors had in mind.

#### **Particle accelerators.**

Toby **Ord**, 10/30/20**08**, Future of Humanity Institute at Oxford. "Probing the Improbable: Methodological Challenges for Risks with Low Probabilities and High Stakes," arXiv.org, https://arxiv.org/abs/0810.5515)// JZ

Particle physics is the study of the elementary constituents of matter and radiation, and the interactions between them. A major experimental method in particle physics involves the use of particle accelerators such as the RHIC and LHC to bring beams of particles to near the speed of light and then collide them together. This focuses a large amount of energy in a very small region and breaks the particles down into their components, which are then detected. **As particle accelerators have become larger, the energy densities achieved have become more extreme, prompting some concern about their safety. These safety concerns have focused on three possibilities: the formation of ‘true vacuum’, the transformation of the earth into ‘strange matter’, and the destruction of the earth through the creation of a black hole.**

**4.1 True vacuum and strange matter formation**

**The type of vacuum that exists in our universe might not be the lowest possible vacuum energy state.** In this case, the vacuum could decay to the lowest energy state, either spontaneously, or if triggered by a sufficient disturbance. This would produce a bubble of ‘true vacuum’ expanding outwards at the speed of light, converting the universe into different state apparently inhospitable for any kind of life (Turner and Wilczek 1982).

Our ordinary matter is composed of electrons and two types of quarks: up quarks and down quarks. **Strange matter also contains a third type of quark: the ‘strange’ quark. It has been hypothesized that strange matter might be more stable than 11 normal matter**, and able to convert atomic nuclei into more strange matter (Witten 1984). It has also been hypothesized that particle accelerators could produce small negatively charged clumps of strange matter, known as strangelets. If both these hypotheses were correct and the strangelet also had a high enough chance of interacting with normal matter, it would grow inside the Earth, **attracting nuclei at an ever higher rate until the entire planet was converted to strange matter — destroying all life in the process. Unfortunately strange matter is complex and little understood, giving models with widely divergent predictions about its stability**, charge and other properties (Jaffe, Busza et al. 2000).

One way of bounding the risk from these sources is the cosmic ray argument: the same kind of high-energy particle collisions occur all the time in Earth’s atmosphere, on the surface on the Moon and elsewhere in the universe. The fact that the Moon or observable stars have not been destroyed despite a vast number of past collisions (many at much higher energies than can be achieved in human experiments) suggest that the threat is negligible. This argument was first used against the possibility of vacuum decay (Hut and Rees 1983) but is quite general.

#### **Death stars.**

**Turchin 22** [Alexey Turchin, B.S. in Physics from Moscow University, National Scientific Complex IASA in Kyiv State Polytechnical Institute; May 21, 2022; “The Future of Nuclear War,” https://www.lesswrong.com/posts/G6JMpyMeMyAsNnkTj/the-future-of-nuclear-war]

Giant gas planets detonation

**Another idea**, which I had discussed elsewhere, **is the use of nuclear weapons to start explosive nuclear reactions inside gas planets**, which has an inner layer rich in lithium and deuterium.

In the article by Weaver and Wood Necessary conditions for the initiation and propagation of nuclear-detonation waves in plane atmospheres about the possibility of ignition of oceans was concluded that in Earth's oceans deuterium has 20 times smaller concentration that is needed for self-sustained nuclear reaction, and secondly, that to start such reaction, a very large bomb is needed, because of complicated radiation absorption effects. However, inside giant gas planets, the conditions may be more favorable for such reactions, because of the higher pressure and higher concentration of deuterium (and lithium). **While gas planets are remote, humanity already managed to put some encapsulated plutonium inside Jupiter and Saturn** during the destructive diving of Galileo and Cassini spacecrafts.

The plutonium inside thermogenerators of these stations is Pu-238, and **pressures inside giant planets could collapse a core of such material**. The sources differ on the possibility to make a nuclear bomb from Pu-238. It is not listed as fissile material, but it is listed to have a critical mass of 9-10 kg, wiki, article. I do not claim here that the explosion was possible: I just want to show that it is not that far from our current capabilities.

In Quora is a good answer: “Pu-238 is fissile, in principle capable of nuclear criticality. But it would make a poor nuclear explosive because its spontaneous fission neutron yield is high, virtually guaranteeing pretrigger “fizzle” yields during attempts to explosively compress it. It also has high specific activity, creating an abundance of decay heat that would preclude most practical nuclear uses, and lots of (a,n) neutrons if used as an oxide or other refractory compound.

In a remote **future**, the theoretical **possibility to detonate a gas planet could be used as a “Death Star” class weapon. Such an explosion may release energy equal to Sun’s luminosity for 10 000 years in a few seconds.** It is still far from a supernova explosion, but **enough to destroy anything on the surfaces of all planets of the Solar system.**

#### **Passive SETI.**

**Barnett 22** [Matthew Barnett, expert predictor and commentator on potential existential risks; March 15, 2022; “My current thoughts on the risks from SETI,” https://www.lesswrong.com/posts/DWHkxqX4t79aThDkg/my-current-thoughts-on-the-risks-from-seti]

**SETI stands for the search for extraterrestrial intelligence. A few projects**, such as Breakthrough Listen, **have secured substantial funding** to observe the sky and crawl through the data **to look for extraterrestrial signals.**

A few effective altruists have proposed that **passive SETI may pose an existential risk to humanity** (for some examples, see here). The primary theory is that **alien civilizations could continuously broadcast a highly optimized message intended to hijack or destroy any other civilizations unlucky enough to tune in**. Many alien strategies can be imagined, **such as sending the code for an AI that takes over the civilization** that runs it, **or sending the instructions on how to build an extremely powerful device that causes total destruction.**

Note that this theory is different from the idea that active SETI is harmful, ie. messaging aliens on purpose. I think **active SETI** is substantially less likely to be harmful, and yet it **has received far more attention in the literature.**

#### **Chemical contamination.**

**Turchin 20** [Alexey Turchin, B.S. in Physics from Moscow University, National Scientific Complex IASA in Kyiv State Polytechnical Institute; 2020; “Global Catastrophic Risks by Chemical Contamination,” http://dx.doi.org/10.13140/RG.2.2.17515.98082]

Chemical and biological weapons are completely different. Biological weapons, such as smallpox, may be self-replicating, whereas **chemical weapons** are not. VX gas, one of the most persistent chemical warfare agents could remain stable in environment below 0˚C (WHO 2004). In the context of global catastrophic risks, much of the globe would need to be saturated with VX in a quantity sufficient to kill a large fraction of people within ten days or less. **Given currently available delivery technology**, this **would not be feasible.**

**One could imagine swarms of robot-sprayers**, similar to slaughterbot drones (Oberhaus 2017; Turchin and Denkenberger 2018a), **which would deliver the dangerous agent worldwide. It does seem to be theoretically possible in the distant future**, if robotics, nanotechnology, and artificial intelligence continue to advance, and if progress in offensive applications outpaces defensive applications.

There are agents more lethal than VX gas, such as **botulinum toxin**, for which the lethal dose is only 0.1 micrograms, but it is very unstable in the environment. Such toxins may one day be used as the lethal payload for military microrobots (Phoenix 2004). **A relatively small amount** of this toxin, less than a kilogram, **would be sufficient to kill everyone on Earth**, if delivery were feasible.

## **2NC**

#### **Suffering risks outweigh by billions of times.**

**Baumann 22** [Tobias Baumann, former researcher at University College London focusing on Cooperative Artificial Intelligence, co-founder of the Center for Reducing Suffering; April 22; Avoiding the Worst: How to Prevent a Moral Catastrophe]

**The scope of s-risks is**, by definition, **astronomical. It is difficult to intuitively grasp the staggering scale of cosmic outcomes. Compared to present-day sources of suffering** like factory farming or wild animal suffering, **s-risks are not just** twice as large or even **10 times as large. Instead**, they are larger **by a factor of** thousands or millions, perhaps even **billions**. Thus, unless the probability of occurrence is vanishingly small, the expected value of s-risks is enormous. To avoid that conclusion, **one would need to be extremely confident that s-risks will not happen.**

I will discuss reasons for optimism and pessimism about the likelihood of s-risks in detail later. For purposes of this simple argument, it suffices to note that **one can hardly justify an extremely low probability. To give a specific number**, I claim that **the probability** of an s-risk materialising **is not less than 1 in 1000**, in light of contemporary analogues (like factory farming) and a range of plausible mechanisms for how s-risks could come about. This lower bound and the vast scope of s-risks suggest (in the EV framework) that averting s-risks should be a priority.

#### **Brain reference architecture replicates human intelligence.**

**Yamakawa 21** [Hiroshi; December; MA, engineering, PhD, Aeronautics, professor of the Graduate School of Engineering, Kyoto University, former professor at the Institute of Space and Astronautical Science; Neural Networks, “The whole brain architecture approach: Accelerating the development of artificial general intelligence by referring to the brain,” vol. 144 p. 478-495]

5.4. Applications of AI systems based on BRA

**AI** systems that are **developed based on BRA can be expected to replicate human** cognitive and behavioral **capabilities** almost **exactly**. Therefore, BRA offers several practical applications. **It enables** the construction of an **AI that exhibits familiarity with humans when communicating** with them. Furthermore, **it can be applied computationally** to research fields that deal with mental illness and cognitive impairment. Conversely, findings regarding human cognitive impairment may be used for problematic behavior that is observed in brain-inspired AI. Moreover, we believe that **this** approach **can** also **be used as** a computational **model that will serve as** a **device for mind uploading**.

6. Conclusions

In this paper, the current WBA approach has been introduced and BRA-driven development to accelerate brain-inspired AGI has been discussed. The **BRA includes standardized data that reflect** the **brain** architecture **for the purpose of limiting** the large design **space that is required for** a human-level **AGI** that cannot be grasped by the cognitive ability of an individual. **Even developers who do not have a deep understanding of the brain can develop** brain-inspired **software based on BRAs** that are **designed by people with expertise in neuroscience**. We explained that the BRA is a description consisting of a BIF supported by a mesoscopic neural circuit and an HCD that is consistent with the BIF. Subsequently, to compensate for the lack of neuroscientific findings, **we introduced** the **SCID method**, which formulates the creation of an HCD that is consistent with the anatomical structure of the brain. Furthermore, even if a BRA is used for development, individual development results tend to diverge depending on the diversity of the target tasks. To address this problem, **integration development is planned, which will move AGI** closer to the functioning of the brain. Moreover, we discussed the evaluation of biological plausibility using BRA to prevent the developed software from veering away from the brain.

The main contribution of this study on BRA-driven development, with the following features, is the establishment of a methodology for accumulating data on brain constraints in a form that can be used for software development.

1. Separation of design information: BRA data can be used in various development projects because they are described in a standard format for software development, which is not dependent on a particular development environment.

2. Standardization of description granularity: As a rule, the description of BRA data at a coarser granularity than the mesoscopic level reduces the possibility that the development will focus on details that are unnecessary for the realization of the target cognitive behavioral level.

3. BRA design: The method of designing computational functions according to anatomy (the SCID method) enables BRAs to be created while compensating for the lack of neuroscientific knowledge in a wide range of brain areas.

4. Tolerance of diversity: Even BRAs that contain mutually contradictory HCDs can be registered if they exhibit a certain level of validity, thereby reducing the risk of overly narrowing the considered design space.

The above features of the **BRA will provide a foundation for** large-scale whole-brain **software** development **as** the **comprehensiveness of its data increases**. Thus, the **brain architecture** will **provide** an **anchor for** the efficient convergence and **eventual completion** **of** the development of human-like **AGI**, whereas the development results in this field tend to diverge at present.