# AGI Catastrophe and Takeover: Some Reference Class-Based Priors

Zach Freitas-Groff

*I am grateful to Holly Elmore, Michael Aird, Bruce Tsai, Tamay Besiroglu, Zach Stein-Perlman, Tyler John, and Kit Harris for pointers or feedback on this document.*

## Executive Summary

### Overview

In this document, I collect and describe reference classes for the risk of catastrophe from superhuman artificial general intelligence (AGI). On some accounts, reference classes are the best starting point for forecasts, even though they often feel unintuitive. To my knowledge, nobody has previously attempted this for risks from superhuman AGI. This is to a large degree because superhuman AGI is in a real sense unprecedented. Yet there are some reference classes or at least analogies people have cited to think about the impacts of superhuman AI, such as the impacts of human intelligence, corporations, or, increasingly, the most advanced current AI systems.

My high-level takeaway is that different ways of integrating and interpreting reference classes generate priors on AGI-caused human extinction by 2070 anywhere between 1/10000 and 1/6 (mean of ~0.03%-4%). Reference classes offer a non-speculative case for concern with AGI-related risks. On this account, AGI risk is not a case of Pascal's mugging, but most reference classes do not support greater-than-even odds of doom. The reference classes I look at generate a prior for AGI control over current human resources anywhere between 5% and 60% (mean of ~16-26%). The latter is a distinctive result of the reference class exercise: the expected degree of AGI control over the world looks to far exceed the odds of human extinction by a sizable margin on these priors. The extent of existential risk, including permanent disempowerment, should fall somewhere between these two ranges.

This effort is a rough, non-academic exercise and requires a number of subjective judgment calls. At times I play a bit fast and loose with the exact model I am using; the work lacks the ideal level of theoretical grounding. Nonetheless, I think the appropriate prior is likely to look something like what I offer here. I encourage intuitive updates and do not recommend these priors as the final word.

### Approach

I collect sets of events that superhuman AGI-caused extinction or takeover would be plausibly representative of, *ex ante*. Interpreting and aggregating them requires a number of data collection decisions, the most important of which I detail here:

1. For each reference class, I collect benchmarks for the likelihood of one or two things:

- Human extinction
- AI capture of humanity’s available resources.

2. Many risks and reference classes are properly thought of as annualised risks (e.g., the yearly chance of a major AI-related disaster or extinction from asteroid), but some make more sense as risks from a one-time event (e.g., the chance that the creation of a major AI-related disaster or a given asteroid hit causes human extinction). For this reason, I aggregate three types of estimates:
    - 50-Year Risk (e.g. risk of a major AI disaster in 50 years)
    - 10-Year Risk (e.g. risk of a major AI disaster in 10 years)
    - Risk Per Event (e.g. risk of a major AI disaster per invention)

    For the latter two types of estimates, see the full document.

3. Given that there are dozens or hundreds of reference classes, I summarise them in a few ways:
    - Minimum and maximum
    - Weighted arithmetic mean (i.e., weighted average)
        - I “winsorise”, i.e. replace 0 or 1 with the next-most extreme value.
        - I intuitively downweight some reference classes. For details on weights, see the methodology.
    - Weighted geometric mean

## Findings for Fifty-Year Impacts of Superhuman AI

See the full document and spreadsheet for further details on how I arrive at these figures.

<table>
<tr><th colspan="3"><b>Reference Class Descriptions and Summaries</b><br><i>Color scale: green = most credible and informative, red = least</i></th></tr>
<tr><td>What I Estimate</td><td>What’s Included</td><td>Summary</td></tr>
<tr><td colspan="3">Emergence of Relatively Superintelligent Species/Genus<br><i>What share of species go extinct because a newly capable species or genus arises?</i><br><i>What share of pre-existing species’ resources do a newly capable species or genera capture?</i></td></tr>
<tr><td>Share of species extinct because of newly superintelligent species</td><td>- Share of megafauna that went extinct shortly after human arrival<br>- Projections of eventual excess mammal species extinction rate in the anthropocene<br>- Adjustments of the above rates to account for the fact that humans are exceptional<br>- Effect of invasive mammal species on island bird extinctions</td><td>Minimum: 0<br>Maximum: 67%<br>Weighted arithmetic mean: 6.69%<br>Weighted geometric mean: 0.524%</td></tr>
<tr><td>Share of resources controlled by newly superintelligent species</td><td>- Share of land modified or used by humans<br>- Share of Earth’s surface used by humans<br>- Share of global or animall biomass consisting of or domesticated by humans<br>- Average population decline across wildlife species in the anthropocene<br>- Adjustments of the above rates to account for the fact that humans are exceptional</td><td>Minimum: 7.72 x $10^{-11}$<br>Maximum: 50%<br>Weighted arithmetic mean: 5.99%<br>Weighted geometric mean: 0.0141%</td></tr>
<tr><td colspan="3">Reasons to believe this reference class:</td></tr>
</table>

<table>
<tr><td colspan="3">- This is a common analogy in arguments about risk from superhuman AGI.<br>- Most arguments about superhuman AGI (e.g. convergence theses) are about the idea of intelligence or discontinuous capabilities and thus apply to intelligent species as well.<br>Reasons not to believe it:<br>- Biological causes of extinction may differ from AGI-related causes.<br>- Intelligent species, including humans may be qualitatively different from superhuman AGI.</td></tr>
<tr><td colspan="3">Known Human Extinction Risks<br><i>What is the chance humanity goes extinct from a plausibly alleged extinction threat?</i></td></tr>
<tr><td>Estimated odds of human extinction</td><td>- Chances of 8 billion deaths from bioterror and biowarfare assuming a power law<br>- Likelihood of mass extinction from an asteroid<br>- Likelihood of mass extinction from a supernova<br>- Likelihood of mass extinction from a gamma ray burst<br>- Yearly chance of “infinite impact” from the Global Challenges Foundation for various causes</td><td>Minimum: 0<br>Maximum: 0.056%<br>Weighted arithmetic mean: 0.00539%<br>Weighted geometric mean: 0.000365%</td></tr>
<tr><td colspan="3">Reasons to believe this reference class:<br>- Since we largely have only intuitive arguments for AGI risk, looking at other “things people argue could cause extinction” offers a natural benchmark.<br>- Extinction from cause X should be less likely the harder it is for humans to go extinct.<br>Reasons not to believe it:<br>- Since AGI is agential, it is likely more damaging than accidental risks.<br>- AGI might be seen as more speculative than the risks included here (and therefore lower).<br>- Observation selection may select for worlds with low natural risks relative to anthropogenic ones.</td></tr>
<tr><td colspan="3">Power of Social Organisations (Governments and Corporations)<br><i>What share of resources are controlled by organised groups of people compared to individuals?</i></td></tr>
<tr><td>Share of resources controlled by social organisations</td><td>- Government or central government spending as share of GDP, US<br>- Share of humans who are citizens of a nation-state<br>- Share of people employed by government in OECD countries<br>- Corporate or government assets as share of global assets</td><td>Minimum: 4.7%<br>Maximum: 50%<br>Weighted arithmetic mean: 20%<br>Weighted geometric mean: 29.4%</td></tr>
<tr><td colspan="3">Reasons to believe this reference class:<br>- Organised of individual humans are in some sense a superintelligent entity relative to individuals.<br>Reasons not to believe it:<br>- It is ambiguous how to distinguish what belongs to a collective and what belongs to individuals.<br>- Socal orgnisations may be less (or more) intelligent than superhuman AGI.</td></tr>
<tr><td colspan="3">Naïve Posteriors from Previous Technologies<br><i>How likely can human extinction be from a threatening invention given prior inventions?</i><br><i>How likely can transformative change be from a major invention given prior major inventions?</i></td></tr>
<tr><td>Chance of extinction from a threatening invention</td><td>- Chance of extinction from a given category of threatening inventions (subjectively defined) given a Beta (0.5, 0.5) prior<br>- Chance of extinction from a given threatening invention (subjectively defined) given a Beta (0.5, 0.5) prior<br><br>In addition, I estimate what rate of extinction would imply a &lt;1% chance of seeing as many threatening inventions as we have seen.</td><td><i>Rate given Beta (0.5, 0.5) prior (&lt;1% likelihood rate in parentheses):</i><br>Minimum: 0.61% (5.53%)<br>Maximum: 6.33% (48.7%)<br>Weighted arithmetic mean: 3.94% (31.4%)<br>Weighted geometric mean: 2.78% (33.4%)</td></tr>
</table>

<table>
<tr><td>Chance of transformation from a major invention</td><td>- Chance an ex-ante potentially transformative invention (subjectively defined) is actually transformative (subjectively defined) given a Beta (0.5, 0.5) prior<br>- Chance a historic invention (subjectively defined) is actually transformative (subjectively defined) given a Beta (0.5, 0.5) prior<br><br>I also compute <1% likelihood estimates as for the extinction measure.</td><td><i>Rate given Beta (0.5, 0.5) prior (<1% likelihood rate in parentheses):</i><br>Minimum: 1.25% (5.1%)<br>Maximum: 17.8% (53.7%)<br>Weighted arithmetic mean: 13.67% (41.6%)<br>Weighted geometric mean: 9.17% (29.8%)</td></tr>
<tr><td colspan="3">Reasons to believe this reference class:<br>- It is perhaps the reference class where it is most obvious superhuman AGI fits in.<br>Reasons not to believe it:<br>- The definition of an invention that would have seemed threatening or major is subjective.<br>- Extinction estimates depend heavily on the prior since we have never observed human extinction.<br>- Here I am taking “chance of transformation” as another estimate of “share of resources controlled”, but it is quite a different way of thinking about that (in probabilities rather than fixed shares).</td></tr>
<tr><td colspan="3">Damages from and Power of AI Systems to Date<br><i>How likely is it that current AI systems would cause human extinction?</i><br><i>What share of current economic activity can be automated by existing AI technologies?</i></td></tr>
<tr><td>Likelihood of a current AI system killing 8 billion people</td><td>- Frequency of “critical” incidents from AI systems<br>- Likelihood a critical incident kills 8 billion people based on various distributions. Note: tenuous and poor fit.</td><td>Minimum: 0<br>Maximum: 0.104%<br>Weighted arithmetic mean: 0.0718%<br>Weighted geometric mean: 0.139%</td></tr>
<tr><td>Forecasted AI share of economy</td><td>Naïve extrapolations of the following:<br>- Share of 2017 work tasks that could be automated<br>- Contribution of automation to GDP</td><td>Minimum: 34.3%<br>Maximum: 69.3%<br>Weighted arithmetic mean: 30.5%<br>Weighted geometric mean: 55.8%</td></tr>
<tr><td colspan="3">Reasons to believe this reference class:<br>- This is perhaps the second-most natural reference class after the previous-technologies one.<br>Reasons not to believe it:<br>- Extinction likelihoods depend on extrapolations and judgment calls that are difficult to defend.<br>- Economic estimates are currently very naïve and likely unrealistic.</td></tr>
<tr><td colspan="3">Rates of Product Defects<br><i>How often do various consumer products exhibit major defects?</i></td></tr>
<tr><td>Share of products with a serious defect</td><td>- Share of cars or car components subject to recall (with or without risk of death)<br>- Share of drugs withdrawn from market or with a post-market safety issue<br>- Share of meat recalled by weight<br>- U.S. standard for acceptable cancer risk</td><td>Minimum: $2.1 \times 10^{-10}$<br>Maximum: 58.5%<br>Weighted arithmetic mean: 2.29%<br>Weighted geometric mean: 0.0645%</td></tr>
<tr><td colspan="3">Reasons to believe this reference class:<br>- This reference class seems somewhat natural, and data is available.<br>Reasons not to believe it:<br>- Determining what counts as catastrophe requires delicate judgment calls.<br>- These sorts of products and defects are likely quite different from AGI misalignment.</td></tr>
</table>

# Overview

In this document, I collect and describe reference classes for the risk of catastrophe from superhuman artificial general intelligence (AGI). On some accounts, reference classes are the best starting point for forecasts, even though they often feel unintuitive. To my knowledge, nobody has previously attempted this for risks from superhuman AGI. This is to a large degree because AGI is in a real sense unprecedented. I argue, however, that it does fit at least loosely into several classes of events where we can observe or bound the frequency of catastrophe. I offer these as a starting point from which one can apply intuitive adjustments.

Each reference class is a set of events that the arrival of superhuman AGI would belong to. The first reference class I cover is the collection of posited human extinction threats where we have some empirical estimate of the likelihood that the threat causes human extinction. Intuitively, this captures how likely it is that humans go extinct from an event that seems threatening. This will not be novel for many, but I argue that it offers a prior for risks where we lack such estimates. The second reference class is the emergence of a species that is much more intelligent ("superintelligent") than other previously existing species. By looking at intelligent species in the past, we can see how often they cause other species to go extinct and how many resources they typically take over. Additional reference classes that are less informative are current AI systems, new products, major historical inventions, and organised states.

For each reference class, I measure the frequency of extinction, the frequency of a serious malfunction, the share of resources taken over, or the frequency of dramatic social transformation. The former two measures offer priors for the likelihood of human extinction. The latter two offer priors for the fraction of the value of future human civilization we should think would be determined by superhuman AGI. Which measure I can produce depends on the reference class. The following table summarises the measures I produce for each reference class:

| Reference Class | Measures of Catastrophe Likelihood |
|---|---|
| Known human extinction risks | ● Likelihood of human extinction |
| Emergence of relatively superintelligent species/genus | ● Frequency of other species' extinction (e.g. megafauna, mammals, vertebrates)<br>● Shares of resources controlled |
| Power of social organizations (governments and corporations) | ● Shares of resources controlled |
| Current AI systems | ● Likelihood of human extinction<br>● Frequency of malfunction<br>● Shares of resources controlled |
| Rates of Product Defects | ● Frequency of malfunction |

| | |
|---|---|
| | ● Frequency of fatal malfunction |
| Naïve posteriors from historic inventions | ● Likelihood of human extinction<br>● Likelihood of social transformation |

For most measures, I calculate a version that is in terms of annual risk and a version that is in terms of the total risk from an event over the course of its lifetime. The latter will typically be higher than the former. Which one is most useful in a given case is a bit of a judgment call.

I try to present reference classes without intuitive adjustments, but intuition is inevitable in how one calculates the measures and combines different possible measures. Which members of each class to include and exclude, how to weight them, and how exactly to define the measure all require judgment calls. I describe how I calculate each measure, and more detail is available in the spreadsheet here.

This document is rougher than something I would normally want to share, but I thought it made sense on balance to put this version together and share it given the conversation happening now about AI risks and timelines. I expect to revise it and could imagine sharing a more polished version later. I would encourage people to not take this as anything like a final word on even the reference classes I have looked into, and there are likely others I have omitted.

## Why Reference Classes?

The basic case for caring about reference classes comes from research on the psychology of decision-making and forecasting. Forecasting based on historical examples is often described as “outside view” forecasting. In a 1993 essay summarising implications of cognitive psychology for decision-making in organizations, Daniel Kahneman and Dan Lovallo write, “It should be obvious that when both methods are applied with equal intelligence and skill, the outside view is much more likely to yield a realistic estimate.” In *Superforecasting*, Philip Tetlock and Dan Gardner note that when estimating a probability, the best forecasters will typically start from a base rate and then adjust. Reference classes offer some base rates to start from.

When approaching a new problem, it seems best to take an approach that has worked well in problems where you could observe results. An important feature of reference-class-based forecasting is that it often or usually feels unintuitive, so if it feels unintuitive in the case of risks from transformative AI, that should not be *prima facie* surprising. That said, there are a number of intuitive reasons why reference classes are useful.

One intuitive justification for using reference classes is that many claims about the future are empirical, and reference classes offer a test of these claims’ validity. For example, the instrumental convergence theses imply things not only about how a superintelligent AI should behave but also about how other intelligent beings should behave (e.g., seeking self-preservation and resources). Even claims that don’t seem empirical at first glance sometimes have empirical implications (deriving and testing the empirical implications of theories is much of what science is about).

A second intuitive justification for using reference classes is that human decision-making often follows an approach of anchoring on a number and then adjusting; reference classes anchor us in the right ballpark for similar events (where the alternative might be “something I understand is unlikely, e.g. 1%”). Third, most of us struggle to assess probabilities, especially very close to zero or one; reference classes can help us think about probabilities that might be very small.

There are a number of criticisms levied against reference classes in general or in specific cases; I tend to think we should be more skeptical of these criticisms and find many of them lacking. Probably the most dramatic of these is Eliezer Yudkowsky’s “Outside View!" as Conversation-Halter.” Much of that post takes issue with a particular approach to using reference classes, which is (as the name suggests) to end the conversation. The claim that reference-classes work despite seeming unintuitive can then shut down objections. To this concern, I basically would say the right way to use reference classes, like those in this document, is as a start rather than an end to the conversation.[1]

A particular objection to reference-class forecasting is that it just ends in “reference-class tennis”, where two sides of an argument lob different reference classes at each other, and the problem collapses into the same conversation as intuitive forecasting. I find this objection informative but not all that damning. Even with a large number of possible reference classes, some will seem more plausible than others. The distribution of estimated frequencies can still be informative about what seems like a reasonable range. Last, we can at least see which reference classes you have to rely on to get a high or low probability; as we’ll see below, we should be more worried about AI the more we anchor on the case of humans’ effect on animal species.

It’s also worth noting that reference-class forecasting has produced some of the most informative research in longtermism. Ajeya Cotra’s report on bio anchors for AI timelines and Tom Davidson’s report on semi-informative priors both involve reference-class forecasts in some sense. Luisa Rodriguez’s work on US-Russia nuclear war uses historical reference classes. Toby Ord’s chapter in *The Precipice* on natural risks uses several different reference class to benchmark the scale of those risks. Piers Millett and Andrew Snyder-Beattie use historical rates of bio attacks and damages from them to forecast the scale of existential risk from bioweapons. David Roodman forecasts economic growth from past growth rates. To varying degrees, all of these cases involve judgment calls on how to combine different base rates, how to understand what they represent, and in some cases how to extrapolate from them. I do the same in the full analysis, though, again, I take this analysis to be earlier-stage than most of these and would accordingly urge caution.

## Methodology

I assemble reference classes for two types of events: (i) the likelihood of human extinction from superhuman AI and (ii) the share of resources otherwise available to humans we

[1] I would vote for then updating from intuitive arguments by thinking about the odds ratio they imply, as described in this interview.

should expect superhuman AI to ultimately control ("takeover"). I took these to be two ways of capturing core worries around AI that were sufficiently precise to compare to other events. Notably, extinction is narrower than "existential risk," as I understand it (because it does not include permanent disempowerment or population decreases), while takeover is broader (because it could theoretically increase human potential).

I consider three ways of thinking about the likelihood of extinction or takeover from superhuman AI. The first is to consider the chance that the invention of superhuman AI *ever* causes human extinction or takes over a large share of resources available to humans. The second and third are to consider the chances that it does so in the first ten or fifty years. I take these three approaches for a few reasons. First, it is sometimes easier to think in annual terms, but in other cases it is easier to think in terms of how likely a single event is to cause catastrophe, regardless the timing. Second, what ultimately matters is whether superhuman AI ever causes a catastrophe in a way we could prevent, not the timing. But third, nearer-term risks may matter more because we may be more able to prevent more distant ones.

In general, the risk over ten years will be less than the risk over fifty, which will be less than the total risk from an event since that can extend beyond fifty years. Yet in some cases because of challenges with how to define an "event", the event-based estimate is higher. This is far from ideal, but I present this as is so that people can select which estimates and which combination of them seems most credible.

To summarise the reference classes, I compute weighted arithmetic and geometric means. My weights seek to account for the fact that some reference classes seem much more plausible than others, and I invite others to copy the spreadsheet and toy with the weights. My weighting scheme generally follows a few principles. First, I try to be naïve and weight things equally, all else equal. Second, I reduce the weight proportionally for reference classes that are largely redundant (i.e., two approaches to the same estimate). Third, I occasionally deviate from naïve weighting when I think the intuitive case for prioritising a given reference class is sufficiently strong. In my overall weight across reference classes, I give the most weight to the "relatively superintelligent species" reference classes followed by the class of other extinction threats and far less weight to the others, which strike me as much less informative.

In each section, I give an explanation of what the reference class is meant to be and an argument of why I think it is useful. In some cases, I think reference classes are not that useful, but I think they deserve some weight and include them because I expect other readers to find them more useful.

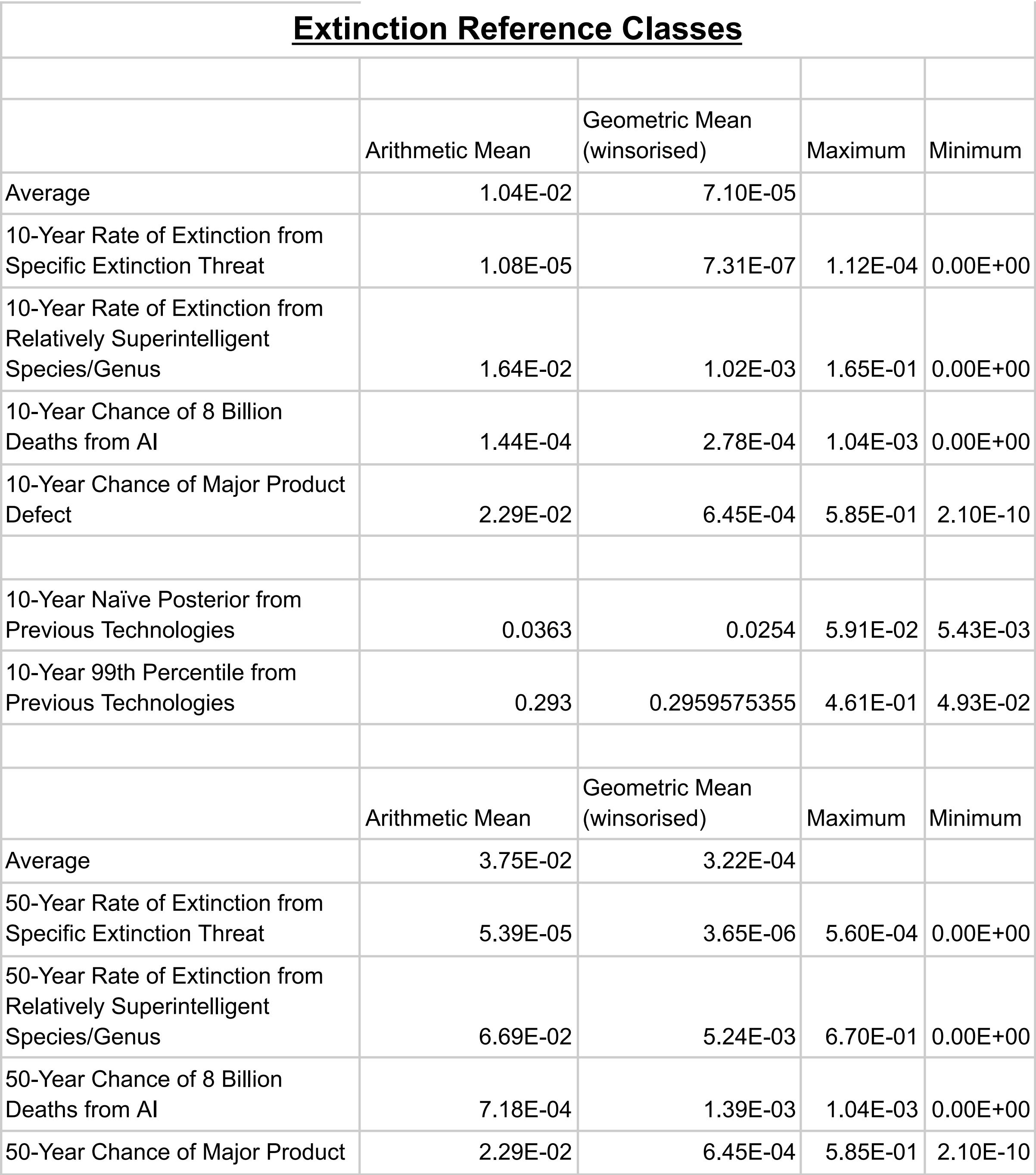
I then include a series of tables (or simple notes substituting for tables) generally taking each of the three approaches to evaluating risk described above (all time for a given AI, ten years, fifty years). In some cases, the reference class is only relevant to extinction or takeover; in others, it is relevant to both. Sometimes for simplicity I include a simplified presentation with a link to the accompanying spreadsheet.

## Summary Statistics

For each table, I give the range that the estimates fall in for that table as well as a weighted arithmetic mean and a weighted geometric mean. The geometric means are computed using the mean of the odds ratio (see Sevilla 2021 for more). I also winsorise all estimates that are zero or one for the geometric mean, meaning I replace them with the minimum and the maximum, respectively.

## Extinction

The following table, produced in this spreadsheet, includes summary statistics for reference classes for the likelihood of human extinction as a result of superhuman AI:



| **Extinction Reference Classes** | | | | |
|---|---|---|---|---|
| | | | | |
| | Arithmetic Mean | Geometric Mean (winsorised) | Maximum | Minimum |
| Average | 1.04E-02 | 7.10E-05 | | |
| 10-Year Rate of Extinction from Specific Extinction Threat | 1.08E-05 | 7.31E-07 | 1.12E-04 | 0.00E+00 |
| 10-Year Rate of Extinction from Relatively Superintelligent Species/Genus | 1.64E-02 | 1.02E-03 | 1.65E-01 | 0.00E+00 |
| 10-Year Chance of 8 Billion Deaths from AI | 1.44E-04 | 2.78E-04 | 1.04E-03 | 0.00E+00 |
| 10-Year Chance of Major Product Defect | 2.29E-02 | 6.45E-04 | 5.85E-01 | 2.10E-10 |
| | | | | |
| 10-Year Naïve Posterior from Previous Technologies | 0.0363 | 0.0254 | 5.91E-02 | 5.43E-03 |
| 10-Year 99th Percentile from Previous Technologies | 0.293 | 0.2959575355 | 4.61E-01 | 4.93E-02 |
| | | | | |
| | Arithmetic Mean | Geometric Mean (winsorised) | Maximum | Minimum |
| Average | 3.75E-02 | 3.22E-04 | | |
| 50-Year Rate of Extinction from Specific Extinction Threat | 5.39E-05 | 3.65E-06 | 5.60E-04 | 0.00E+00 |
| 50-Year Rate of Extinction from Relatively Superintelligent Species/Genus | 6.69E-02 | 5.24E-03 | 6.70E-01 | 0.00E+00 |
| 50-Year Chance of 8 Billion Deaths from AI | 7.18E-04 | 1.39E-03 | 1.04E-03 | 0.00E+00 |
| 50-Year Chance of Major Product | 2.29E-02 | 6.45E-04 | 5.85E-01 | 2.10E-10 |

| | | | | |
|---|---|---|---|---|
| Defect | | | | |
| | | | | |
| 50-Year Naïve Posterior from Previous Technologies | 3.94E-02 | 2.78E-02 | 6.33E-02 | 6.10E-03 |
| 50-Year 99th Percentile from Previous Technologies | 3.14E-01 | 3.34E-01 | 4.87E-01 | 5.53E-02 |
| | | | | |
| | Arithmetic Mean | Geometric Mean (winsorised) | Maximum | Minimum |
| Average | 4.98E-02 | 1.31E-03 | | |
| Likelihood of Extinction from Occurrence of Specific Extinction Threat | 1.22E-03 | 2.37E-04 | 4.00E-03 | 1.12E-05 |
| Likelihood of Extinction from Lifetime of Relatively Superintelligent Species/Genus | 8.91E-02 | 1.19E-02 | 6.70E-01 | 0.00E+00 |
| Likelihood of Incident that Kills 8 Billion People from AI Invention | 1.12E-06 | 5.75E-10 | 6.24E-06 | 0.00E+00 |
| Likelihood of Defect in a Major Product's Lifetime | 2.29E-02 | 6.45E-04 | 5.85E-01 | 2.10E-10 |

## Takeover

The following table, again produced in this spreadsheet, includes summary statistics for reference classes for the share of the future a superhuman AI controls:

| | | | | |
|---|---|---|---|---|
| **Takeover Reference Classes** | | | | |
| | | | | |
| | Arithmetic Mean | Geometric Mean (winsorised) | Maximum | Minimum |
| Average | 1.41E-01 | 8.29E-03 | | |
| 10-Year Share of Resources Controlled by Relatively Superintelligent Species/Genus | 1.54E-02 | 1.26E-05 | 2.07E-01 | 7.89E-08 |
| 10-Year Superhuman AI Share of the Economy | 2.58E-01 | 2.49E-01 | 3.61E-01 | 5.88E-01 |
| 10-Year Social Organization Share of Resources Controlled | 2.00E-01 | 2.94E-01 | 5.00E-01 | 4.70E-02 |
| 10-Year Likelihood Major Technology is Transformative | 1.50E-01 | 1.00E-01 | 1.95E-01 | 1.37E-02 |
| | | | | |
| | Arithmetic Mean | Geometric Mean (winsorised) | Maximum | Minimum |

| | | | | |
|---|---|---|---|---|
| Average | 1.57E-01 | 2.58E-01 | | |
| 50-Year Share of Resources Controlled by Relatively Superintelligent Species/Genus | 5.99E-02 | 1.41E-04 | 5.00E-01 | 7.72E-11 |
| 50-Year Superhuman AI Share of the Economy | 3.05E-01 | 5.58E-01 | 6.93E-01 | 3.43E-01 |
| 50-Year Social Organization Share of Resources Controlled | 2.00E-01 | 2.94E-01 | 5.00E-01 | 4.70E-02 |
| 50-Year Likelihood Major Technology is Transformative | 1.37E-01 | 9.17E-02 | 1.78E-01 | 1.25E-02 |
| | | | | |
| | Arithmetic Mean | Geometric Mean (winsorised) | Maximum | Minimum |
| Average | 3.30E-01 | 3.32E-02 | | |
| Eventual Share of Resources Controlled by Relatively Superintelligent Species/Genus | 1.20E-01 | 2.82E-04 | 1.00E+00 | 1.54E-10 |
| Eventual AI Share of the Economy | 9.92E-01 | 9.99E-01 | 1.00E+00 | 9.16E-01 |
| Eventual Social Organization Share of Resources Controlled | 4.01E-01 | 5.88E-01 | 9.99E-01 | 9.39E-02 |
| Eventual Likelihood Major Technology is Transformative | 1.37E-01 | 9.17E-02 | 1.78E-01 | 1.25E-02 |

# Emergence of Relatively Superintelligent Species/Genus

In this section, I examine how often new intelligent species cause other species to go extinct and what share of available resources intelligent species typically capture. The reference class is the set of species we consider substantially more intelligent than previously existing species. Depending on one's view, this might only be humans, or it might be a wider set of animals, such as great apes. I also consider invasive species, which may not be more intelligent on every dimension but share the feature that they suddenly introduce new capabilities to ecosystems.

In general, my estimates of the share of other species that go extinct are the estimates for humans divided by the total number of species that are intelligent in the relevant sense (I consider a few definitions). This is because no species other than humans have caused widespread extinction to my knowledge. The estimates for the share of available resources controlled are similar. For available resources, I am specifically interested in a measure of exclusion: how many resources other species have do relatively superintelligent species appropriate?

## What This Reference Class Captures

This reference class capture what typically happens when a relatively superintelligent being emerges. The effect of a new intelligent species on the extinction of other species offers a prior for how likely it is that superhuman AGI will cause humans to go extinct. The share of resources controlled offers a prior for takeover, or what fraction of the value of the future we should expect a superhuman AGI to determine. I say "relatively superintelligent" to include "invasive" species, which are not necessarily more intelligent in an absolute or general sense but have important capabilities foreign to an ecosystem.

The distinction between annualised and total risk estimates is tricky for this reference class. I estimate the total risk as extinction frequency of other species and the eventual control of resources over the lifetime a relatively superintelligent species. This involves forecasts of future extinction rates given historical dynamics. I then recommend adjustments to the extinction frequency or control of resources within ten or fifty adjusted years of a relatively superintelligent species's emergence. The adjustment here is key: I assume a superhuman AGI will move faster than humans do, so I compute the computational equivalent of ten or fifty years' time based on computational benchmarks from Ajeya Cotra's Biological Anchors report and some additional computations.

## Why Is This Reference Class Informative (or Not)?

The first reason to take this reference class seriously is that it already informs worries about superhuman AGI in an informal sense. The thought that humans faced with TAI might be like animals faced with the evolution of humans is at least informally important in many discussions. A preview of Bostrom's *Superintelligence* notes the fate of gorillas in the face of humanity:

> The human brain has some capabilities that the brains of other animals lack. It is to these distinctive capabilities that our species owes its dominant position. Other animals have stronger muscles or sharper claws, but we have cleverer brains. If machine brains one day come to surpass human brains in general intelligence, then this new superintelligence could become very powerful. As the fate of the gorillas now depends more on us humans than on the gorillas themselves, so the fate of our species then would come to depend on the actions of the machine superintelligence.

The idea of humans no longer being the smartest creatures on the planet seems to animate many worries about superhuman AGI. Other species past and present have faced the emergence of a more intelligent species before, so trying to understand what happened should tell us the appropriate scale of this worry.

Second, most arguments about AI risk are arguments about the idea of intelligence.[2][3] They therefore yield predictions for other intelligent species as well. The instrumental convergence theses seem like they should not only apply to superhuman AGI but also to newly intelligent animals. As a result, observing that intelligent species do or do not tend to destroy other species offers some evidence about whether we should expect superhuman AGI systems to do so.

[2] Not all (https://www.cold-takes.com/why-ai-alignment-could-be-hard-with-modern-deep-learning/), but even arguments specific to machine learning seem like they would plausibly apply to other conceptions of intelligence, especially intelligence that is selected for in an indirect way, as biological intelligence is.

[3] Relatedly, the biological theory "competitive exclusion" resembles certain arguments about AGI, namely that two similar species (humans and human-derived AGI) cannot coexist. The principle appears not to have much empirical support, and to some extent this reference class offers some evidence against it. I am grateful to Tamay Besiroglu for pointing this out.

For a (darkly, given the matter at hand) humorous take on the relevance of this reference class, see *The Onion*, “Dolphins Evolve Opposable Thumbs”: “‘I believe I speak for the entire human race when I say, 'Holy fuck,'' said Oceanographic Institute director Dr. James Aoki, noting that the dolphin has a cranial capacity 40 percent greater than that of humans. ‘That's it for us monkeys.’”

A reason one might not expect this to be informative is that the causes of other species’ extinction at the hands of biological species may be very different from the causes of extinction at the hands of superhuman AGI. Disease, predation, and over-harvesting in particular might seem to drive the biological cases but be irrelevant to superhuman AGI. I ignore this issue in this section for two reasons. First, there is a general sense in which all of these impacts (disease, predation, and over-harvesting) are particular cases of “a species expands into and makes use of the ecosystem without regard to neighboring species”. The precise form of this may differ, but the idea that superhuman AGI may use resources without regard to humans falls into the same basket. Second, arguments that we should worry about superhuman AGI because of the human track record are reasonably common and assume that the biological track record is informative even if the details differ.

## Should We Expect These Benchmarks to Be Too High or Too Low?

This is perhaps the only reference class where I do not have a strong view on whether the estimates are too high or too low as benchmarks for the risk from superhuman AGI. In general, it seems like it should be a fairly reasonable prior. If I lean in one direction, I lean toward these estimates being too low as benchmarks, since it seems like absolute and not only relative intelligence should matter, especially for recursive self improvement. On the other hand, we might expect these estimates to be too high because humans potentially get to design AGI.

## Estimates

### Notes on Construction of Estimates

My reasons for which reference classes to include in this section are as follows. First, I restrict this investigation to animals (rather than plants, bacteria, etc.) given the similarity, availability of research, and my ease of comprehension. I consider four different reference classes of relatively superintelligent species: humans, apes, primates, and invasive mammals.

Humans are likely the most useful reference class as they most obviously possess something worth calling general intelligence, but I give serious consideration to the fact that there are other plausibly intelligent species which have not, to my knowledge, caused any extinctions or taken a sizable share of even local resources from other species. I apply this consideration by dividing estimates of human-caused extinctions and human-controlled resources by the numbers of other intelligent species under varied definitions. For simplicity, the two separate numbers I use for these adjustments are the number of primates and the number of great apes. My reasons for using these numbers are that primates, especially great apes, also possess sizable frontal lobes, are frequently cited examples of animals that use tools, and are generally similar to humans in some intuitive sense. The numbers of great ape and primate species also seem to me to offer a reasonable benchmark for the number of species with an encephalization quotient at least as large as that of a whale or a gorilla; I expect some primate species to be much less intelligent but some non-primate species, like marine mammals, to compensate and leave these numbers as fairly accurate adjustments.

The reason I obtain average estimates for primate or ape species by simply dividing the estimates for humans is that it appears exceptionally rare for other any species's emergence to cause another species to go extinct (not counting one species evolving into another one). The background extinction rate for all species, i.e. the rate outside of an extinction event and including all possible causes, is around 100 to 1000 times less than that of modern times. Moreover, as far as I can tell, the evolution of a new species has never caused mass extinction event, except the current one if it qualifies.[4]

As noted above, I also look at invasive species, which may not be more intelligent on every dimension but share the feature that they suddenly introduce new capabilities to ecosystems. It seems like AGI risk stories often see superhuman AGI as somethink akin to an invasive species. There also are other examples of biological innovation possibly causing mass extinction, such as cyanobacteria triggering Snowball Earth and burrowing worms triggering mass extinction of Ediacaran fauna.[5]

See Species in the spreadsheet for more detail.

### Extinction Rate After Emergence of Relatively Superintelligent Species/Genera

Range of estimates: [0, 0.67]
Weighted arithmetic mean: 1.10 x $10^{-1}$
Weighted arithmetic mean (winsorised): 1.10 x $10^{-1}$
Weighted geometric mean (winsorised): 1.94 x $10^{-2}$

| Measure | Source | Estimate | Notes |
|---|---|---|---|
| Share of megafauna species that went extinct shortly after the arrival of humans (between 38,050 and 2,050 BCE) | Barnosky (2008) | 5.00E-01 | |
| Average share of megafauna species that went extinct shortly after the arrival of one of today's great ape species | Ibid.<br>Wikipedia contributors. "Primate" (2022)<br>Wilson and Reeder (2005) | 6.25E-02 | Previous row divided by eight since there are eight great ape species, and others have not caused extinction to my knowledge (see "Notes on Findings" for how I concluded this).<br><br>I limit this to currently existing species so that I do not consider species which are very |

[4] This relates to a point Joe Carlsmith makes in "Is Power-Seeking AI an Existential Risk?": "some argue that the fate of the chimpanzees is currently in human hands, and that this difference in power is primarily attributable to differences in intelligence, rather than e.g. physical strength… This argument is suggestive, but far from airtight. Chimpanzees, for example, are themselves much more intelligent than mice, but the "fate of the mice" was never "in the hands" of the chimpanzees."

[5] I am grateful to Holly Elmore for pointing me to these examples.

| | | | |
|---|---|---|---|
| | | | short-lived. |
| Average share of megafauna species that went extinct shortly after the arrival of one of today's primate species | Ibid. | 1.11E-03 | Two rows up divided by 450 for the number of primate species. See previous cell for further notes. |
| Share of large and medium-sized mammals that went extinct shortly after the arrival of humans (between 38,050 and 2,050 BCE) | Wikipedia contributors. "Quaternary extinction event" (2022)<br><br>Putshkov (1997) | 3.41E-01 | |
| Average share of megafauna species that went extinct shortly after the arrival of one of today's great ape species | Ibid.<br>Wikipedia contributors. "Primate" (2022)<br>Wilson and Reeder (2005) | 4.26E-02 | See note three rows up. |
| Average share of megafauna species that went extinct shortly after the arrival of one of today's primate species | Ibid. | 7.57E-04 | See note three rows up. |
| Share of megafauna genera that went extinct shortly after the arrival of humans (between 38,050 and 2,050 BCE) | Barnosky (2008) | 6.70E-01 | |
| Average share of megafauna genera that went extinct shortly after the arrival of one of today's great ape genera | Ibid.<br>Wikipedia contributors. "Hominidae" (2022)<br>Wilson and Reeder (2005) | 1.68E-01 | Previous row divided by four since there are four great ape genera, and others have not caused extinction to my knowledge (see “Notes on Findings” for how I concluded this).<br><br>I limit this to currently existing genera so that I do not consider genera which are very short-lived. |
| Average share of megafauna genera that went extinct shortly after the arrival of one of today's primate genera | Ibid.<br>Wikipedia contributors. "Primate" (2022) | 9.31E-03 | Two rows up divided by 72 for the number of primate genera. See previous cell for further notes. |

| | | | |
|---|---|---|---|
| Projected share of all mammal species extinct within 100 years based on current status | Roser et al. (2022).<br>IUCN Red List. (2022). | 1.59E-01 | |
| Projected share of all mammal species extinct attributable to a randomly-selected great ape species | Ibid. | 1.99E-02 | Previous row divided by eight since there are eight great ape species, and others have not caused extinction to my knowledge (see “Notes on Findings” for how I concluded this). |
| Projected share of all mammal species extinct attributable to a randomly-selected primate species | Ibid. | 3.54E-04 | Two rows up divided by 450 for the number of primate species. See previous cell for further notes. |
| Projected share of all mammal species extinct attributable to a randomly-selected great ape genus | Ibid. | 3.99E-02 | Three rows up divided by four since there are four great ape genera, and others have not caused extinction to my knowledge (see “Notes on Findings” for how I concluded this). |
| Projected share of all mammal species extinct attributable to a randomly-selected primate genus | Ibid. | 2.21E-03 | Four rows up divided by 72 for the number of primate genera. See previous cell for further notes. |
| Rough estimate of increase in probability of a given bird species going extinct per mammal predator introduced | Blackburn et al (2004) | 5.84E-03 | |
| Increase in probability of a bird species going extinct per mammal herbivore introduced | Blackburn et al (2004) | 0.00E+00 | |
| Average of minimum and maximum projections of eventual species extinctions attributable to a given invasive mammalian species | Barnosky (2008)<br>Doherty et al (2016) | 3.08E-03 | |

### Extinction Rate Per Relatively Superintelligent Species/Genus, First Ten Years

Range of estimates: [0, 0.165]

Weighted arithmetic mean: 1.98 x $10^{-2}$
Weighted arithmetic mean (winsorised): 1.77 x $10^{-3}$
Weighted geometric mean (winsorised): 2.41 x $10^{-3}$

These estimates come from a weighted adjustment of the previous table that is available in the Species spreadsheet. To produce species-based reference classes for extinction in the first ten years of an intelligent being, I recommend simply adjusting the above table, so I do not present a separate table here. The appropriate reference class depends on a judgment about how quickly we should expect superhuman AI to cause extinctions compared to intelligent species. There are two approaches I consider, yielding three different multipliers I apply to the values in the above table:

- *Biological life years:* multiplier of 0 for all reference classes based on humans (including those based on generally intelligent species); adjust invasive-mammal-driven reference classes using a more detailed assumption.
    - This adjustment assumes superhuman AI would move as fast as a biological species; that is, it would cause as many extinctions in its first ten years as a typical intelligent species does in its first then years of existence.
    - For the human-based reference classes, while a sharp zero is unlikely to be realistic, I do think we should significant weight on an approximate zero to capture the fact that it may take real time for a superhuman AI to move the world in important ways (e.g. to persuade or even kill large numbers of humans).
    - For invasive mammals, I assume they have caused extinctions at a constant rate and have been in their current environments for 200 years, consistent with the data from Blackburn et al (2004). The adjustment is a formula, not a fixed factor.
- *Computations used in training:* multipliers of 24.7% for megafauna extinction reference classes, 21.0%-22.2% [preferred estimate: 22.2%] for current extinction reference classes, and 100% for invasive-mammal-driven reference classes.
    - This adjustment assumes that in a given year, an intelligent being causes extinctions based on the number of computations performed in its lifetime or training.
    - I use estimates of computations per human from Ajeya Cotra's bio anchors report and the human population from Our World in Data to estimate how many computations humans collectively performed up until the megafauna extinction (52,000-9,000) and the modern age (1500-2022).
    - I take from the "best guess" forecast in the bio anchors report as a distribution over when superhuman AI will be developed and how many computations it will have performed as of that time. I use this to estimate how likely it is that within ten years, the total computations available will have increased by at least the number of computations humans had performed up until the megafauna extinction and the modern age, respectively.
    - Performing the computations mammals perform in the typically small amount of time that invasive species have been in their new ecosystems should be trivial relative to performing the computations humans have performed over human history, so I assume 100% of that impact comes in the first 10 computational years.

See Computational Timeline in the spreadsheet for more detail.

### Extinction Rate Per Relatively Superintelligent Species/Genus, First Fifty Years

Range of estimates: [0, 0.67]
Weighted arithmetic mean: 8.24 x $10^{-2}$
Weighted arithmetic mean (winsorised): 8.26 x $10^{-2}$
Weighted geometric mean (winsorised): 9.07 x $10^{-3}$

I approach the fifty years question analogously to the ten years question.

- *Biological life years:* same approach as for ten years.
- *Computations used in training:* multipliers of 100% for everything based on the same approach as for ten years.

See Computational Timeline in the spreadsheet for more detail.

### Share of Resources Eventually Controlled By Newly Capable Species

Range of estimates: [1.54 x $10^{-10}$, 1]
Weighted arithmetic mean: 0.120
Weighted arithmetic mean (winsorised): 0.120
Weighted geometric mean (winsorised): 2.82 x $10^{-4}$

| **Measure** | **Source** | **Estimate** | **Notes** |
|---|---|---|---|
| Share of land modified by humans | Theobald, David M., et al. "Earth transformed: detailed mapping of global human modification from 1990 to 2017." Earth System Science Data 12.3 (2020): 1953-1972. | 1.46E-01 | I am not entirely sure how this differs from the next row; I'm inclined to consider both. |
| Share of land currently used by humans | Hannah Ritchie and Max Roser (2013) - "Land Use". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/land-use' [Online Resource] | 5.00E-01 | The share of global land used by humans has slowed in recent years, so I think both this and the next row are worth considering. |
| Share of land ultimately used by humans, extrapolating out for our expected lifetime | Hannah Ritchie and Max Roser (2013) - "Land Use". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/land-use' [Online Resource] | 1.00E+00 | This growth of land usage overtime easily implies 100% of land used eventually. The previous rows are most useful if we think land usage will slow eventually; this row is useful if we think it will continue apace. |
| Share of Earth's surface area used by humans | Hannah Ritchie and Max Roser (2013) - "Land Use". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/land-use' [Online Resource]<br>LePen, Nicholas. "How much of Earth's surface is covered by each country — in one graphic." World Economic Forum (2021). | 1.50E-01 | 70% of the earth is water which is overwhelmingly unexplored. |

| | | | |
|---|---|---|---|
| Share of Earth's surface area ultimately used by humans, extrapolating out for our expected lifetime | Hannah Ritchie and Max Roser (2013) - "Land Use". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/land-use' [Online Resource]<br>LePen, Nicholas. "How much of Earth's surface is covered by each country — in one graphic." World Economic Forum (2021). | 1.00E+00 | This growth of land usage overtime easily implies 100% of land used eventually. The previous rows are most useful if we think land usage will slow eventually; this row is useful if we think it will continue apace. |
| Share of global biomass consisting of or domesticated by humans | Hannah Ritchie (2022) - "Wild mammals make up only a few percent of the world's mammals". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/wild-mammals-birds-biomass' [Online Resource] | 1.60E-04 | |
| Share of animal biomass consisting of or domesticated by humans | Hannah Ritchie (2022) - "Wild mammals make up only a few percent of the world's mammals". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/wild-mammals-birds-biomass' [Online Resource] | 4.00E-02 | |
| Average population decline across wildlife species since 1970 | Roser, Max, et al. "Biodiversity." Our world in data (2021). | 6.91E-01 | Decline from the Living Planet Index; notably, the decline was steepest in the middle of the sample and fairly recently. |
| Average population decline across wildlife species, extrapolating back to 1500 | Roser, Max, et al. "Biodiversity." Our world in data (2021). | 9.11E-01 | I take the decline in the first four years of the Living Planet Index and estimate the decline since 1800 assuming the rate of decline in 1970-1974 was the same as the rate from 1800-1970. |
| Above rates adjusted to reflect average share controlled by a great ape species | Wikipedia contributors. "Primate." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 27 Nov. 2022. Web. 30 Nov. 2022.<br>Wilson, Don E., and DeeAnn M. Reeder, eds. Mammal species of the world: a taxonomic and geographic reference. Vol. 1. JHU press, 2005. | Divide by 8 | See Species in the spreadsheet for more detail. |

| | | | |
|---|---|---|---|
| Above rates adjusted to reflect average share controlled by a primate species | Ibid. | Divide by 450 | See Species in the spreadsheet for more detail. |
| Above rates adjusted to reflect average share controlled by a great ape genus | Ibid. | Divide by 4 | See Species in the spreadsheet for more detail. |
| Above rates adjusted to reflect average share controlled by a primate genus | Ibid. | Divide by 72 | See Species in the spreadsheet for more detail. |

### Share of Resources Controlled By Newly Capable Species, First Ten Comp. Years

Range of estimates: [0, 0.207]
Weighted arithmetic mean: 1.54 x $10^{-2}$
Weighted arithmetic mean (winsorised): 1.54 x $10^{-2}$
Weighted geometric mean (winsorised): 1.26 x $10^{-5}$

I take the same approach here as for extinction. Most resources captured by humans appear to have been captured since the Industrial Revolution, so I adopt the 22.2% multiplier for computations.[6] See Computational Timeline in the spreadsheet for more detail.

### Share of Resources Controlled By Newly Capable Species, First Fifty Comp. Years

I take the same approach here as for extinction. This amounts to 50% weight on a multiplier of one (computational years) and 50% weight on a multiplier of zero (biological years). See Computational Timeline in the spreadsheet for more detail.

## Known Human Extinction Risks

In this section, I summarise estimates of the likelihood of human extinction from other risks where estimates exist with some empirical basis. Much of this is a recapitulation of Toby Ord's estimates from *The Precipice* but restricted to those that are empirical (thus excluding the intuitive estimates of anthropogenic and future risks). I argue that these estimates should inform our prior for the likelihood of human extinction from superhuman AGI.

[6] From my coarse estimates for how many biological years AGI can simulate within ten years of being invented, the chance that AGI can perform enough computations to complete some but not all of the Industrial Revolution is very small. For this reason, I ignore any chance that it can execute some but not all of the modern extinctions.

## What This Reference Class Captures

This reference class captures the likelihood of human extinction from risks that can be somewhat empirically estimated. Obviously the estimates are not standard frequencies; they do not tell us how often humans go extinct out of the number of times an event occurs. Instead, they either combine assumptions on data distributions or evidence from the natural sciences on what sorts of events would prevent human survival. This reference class excludes risks where nobody has yet been able to perform this sort of analysis. It also excludes supernatural risks.

In keeping with my overall approach, I offer estimates of annual risk from the possibility of an event (e.g. the possibility of biowar) and the risk from each time the event occurs (e.g. the risk when there is a biowar event). The former is lower because it accounts for the fact that threatening events do not happen every year. The more we think the risk from superhuman AGI is front-loaded, the more we would want to adopt the second measure. There are some estimates where I only have an annual measure.

I do not consider non-extinction takeover in this section.

## Why Is This Reference Class Informative (or Not)?

This reference class is informative under either of two plausible views.

First, we might see superhuman AI as a special case of “things that people argue can cause human extinction”; a compelling argument is basically all we have to justify the view that superhuman AI threatens human extinction. After reading chapters of *Global Catastrophic Risks* on specific risks (e.g. supervolcanism or asteroids) without social context and without additional review of historical evidence, I would wager many educated readers would evaluate these risks similarly to that of superhuman AI.

Second, we might think that finding the right order of magnitude for the chance of AGI-caused human extinction is primarily a matter of finding the right order of magnitude for human extinction in general. The event “human extinction” requires a highly-specific set of things to all happen. Almost every human must die around the same time, based on my reading of Luisa Rodriguez’s interview on the 80,000 Hours podcast. So we might want to start estimating AI extinction risk from a benchmark for how plausible extinction is in general.

This reference class is uninformative to the extent we think that natural and, to a lesser extent, previous anthropogenic risks differ from the risks from superhuman AGI. A core claim in the literature on existential risk is that man-made risks are much higher than natural ones. This claim seems to require an explanation. My understanding is the best explanation for how these can come apart is that man-made risks, especially superhuman AGI, can be agentic. Killing every single human is difficult by accident, but goal-seeking behavior makes that more likely.

A final issue with this reference class concerns observation selection effects. Observers, i.e., humans, will tend to find ourselves in cases where long-term, i.e., natural, risks are low relative to relatively recent, i.e., anthropogenic risks. This may drive a systematic wedge between the odds of extinction from natural causes and the odds of extinction from more recent causes. This issue does not affect all the examples in this section, however, because biological, nuclear, climate change, synthetic biology, and nanotechnology risks are anthropogenic and recent enough not to be heavily selected.

## Should We Expect These Benchmarks to Be Too High or Too Low?

In general, the benchmark estimates from this reference class seems like they should be too low. The main reason has to do with the agentic nature of superhuman AGI, which makes it potentially more capable of wiping out every single human than an event that is purely accidental.

It is possible, however, that these benchmarks are too high. Because I only include events where we can empirically estimate the scale of the risk, I include threats that are fairly well-vetted in their ability to cause human extinction. One might think that because the risk from superhuman AGI is less empirically demonstrable, we should treat it as being lower.

## Estimates

### Notes on Construction of Estimates

I am indebted to Michael Aird's database of existential risk estimates for many of these sources.

I only include primarily empirical estimates of the extinction risks, i.e. I exclude claims of subjective confidence. I exclude estimates for supervolcanoes from *The Precipice* because it was less clear how to translate their frequency into a risk of extinction. I was hesitant on whether to include Pamlin and Armstrong (2015), which cites a set of papers for each estimate that I believe they aggregate to yield there estimates. However, the citations appear to have a serious empirical basis in each case cited here. I excluded their estimate for "Unknown Consequences," which cites only an expert survey.

See Other_Risks in the spreadsheet for more detail.

### Risk Per Event from Specific Risks

Range of estimates: [$8.7 \times 10^{-5}$, $4 \times 10^{-3}$]
Weighted arithmetic mean: $1.22 \times 10^{-3}$
Weighted geometric mean: $2.37 \times 10^{-4}$

| Metric | Source | Estimate | Notes |
|---|---|---|---|
| Chance of 8 billion deaths from a biological terrorism event assuming a power law distribution | Millett and Snyder-Beattie (2017) | 1.12E-05 | The estimate I include here is different from (higher than) that in the original paper because I drop an intuitive adjustment. |
| Chance of 8 billion deaths from a wartime biological attack assuming a power law distribution | Millett and Snyder-Beattie (2017) | 8.70E-05 | See the previous row. |
| Share of asteroids that threaten mass extinction (i.e. one over 10 kilometers) in an average century | Ord (2020) | 4.00E-03 | |

| | | | |
|---|---|---|---|
| Share of asteroids that threaten mass extinction (i.e. one over 10 kilometers) in the current century | Ord (2020) | 8.00E-04 | |

## Risk Per Year of Possibility from Specific Risks

Range of estimates: [1 x $10^{-660}$, 1.12 x $10^{-5}$]
Weighted arithmetic mean: 1.08 x $10^{-6}$
Weighted arithmetic mean (winsorised): 1.08 x $10^{-6}$
Weighted geometric mean (winsorised): 7.31 x $10^{-8}$

| Metric | Source | Estimate | Notes |
|---|---|---|---|
| Yearly chance of 8 billion deaths from a biological terrorism assuming a power law distribution | Millett and Snyder-Beattie (2017) | 1.12E-05 | The source finds one bio attack per year on average, so this is simply the same as the extinction risk per event. |
| Yearly chance of 8 billion deaths from a wartime biological attack assuming a power law distribution | Millett and Snyder-Beattie (2017) | 2.61E-06 | I count six events over the past two hundred years. |
| Yearly chance of an asteroid that threatens mass extinction (i.e. one over 10 kilometers) in an average century | Ord (2020) | 6.67E-07 | |
| Yearly chance of an asteroid that threatens mass extinction (i.e. one over 10 kilometers) in the current century | Ord (2020) | 6.67E-09 | |
| Yearly chance of a supernova that depletes the ozone by more than 30% in an average century | Ord (2020) | 2.00E-07 | |
| Yearly chance of a supernova that depletes the ozone by more than 30% in the current century | Ord (2020) | 2.00E-08 | |
| Yearly chance of a gamma ray burst that depletes the ozone by | Ord (2020) | 4.00E-07 | |

| | | | |
|---|---|---|---|
| more than 30% | | | |
| Yearly chance of the Earth freezing or boiling or crashing into another planet from stellar disruption of planetary orbits | Ord (2020) | 5.00E-15 | |
| Yearly chance of the vacuum of space collapsing | Ord (2020) | $1.00 \times 10^{-660}$ | This is one of two upper bounds offered; others suggest a sharp zero. |
| Yearly chance of “infinite impact” from climate change | Pamlin and Armstrong (2016) | 5.00E-07 | The report defines an infinite impact as “When civilisation collapses to a state of great suffering and does not recover, or a situation where all human life ends. The existence of such threats is well attested by science.” |
| Yearly chance of “infinite impact” from nuclear war | Pamlin and Armstrong (2016) | 5.00E-07 | |
| Yearly chance of “infinite impact” from a global pandemic | Pamlin and Armstrong (2016) | 1.00E-08 | |
| Yearly chance of “infinite impact” from a major asteroid impact | Pamlin and Armstrong (2016) | 1.30E-08 | |
| Yearly chance of “infinite impact” from a supervolcano | Pamlin and Armstrong (2016) | 3.00E-09 | |
| Yearly chance of “infinite impact” from synthetic biology | Pamlin and Armstrong (2016) | 1.00E-06 | |
| Yearly chance of “infinite impact” from nanotechnology | Pamlin and Armstrong (2016) | 1.00E-06 | |

### Risk Per Year of Possibility from All Natural Risks

Range of estimates: $[1 \times 10^{-8}, 5 \times 10^{-6}]$
Weighted arithmetic mean: $1.19 \times 10^{-6}$
Weighted geometric mean: $2.74 \times 10^{-7}$

| **Metric** | **Source** | **Estimate** |
|---|---|---|

| | | |
|---|---|---|
| Midpoint of “Best Guess” interval based on lifetime of *Homo sapiens* | Ord (2020) | 2.50E-06 |
| Midpoint of “Best Guess” interval based on time since Neanderthal split | Ord (2020) | 1.00E-06 |
| Midpoint of “Best Guess” interval based on lifetime of *Homo* genus | Ord (2020) | 1.50E-07 |
| Estimated survival rate based on lifetime of *Homo neanderthalensis* | Ord (2020) | 5.00E-06 |
| Estimated survival rate based on lifetime of *Homo heidelbergensis* | Ord (2020) | 2.50E-06 |
| Estimated survival rate based on lifetime of *Homo habilis* | Ord (2020) | 2.00E-06 |
| Estimated survival rate based on lifetime of *Homo erectus* | Ord (2020) | 6.00E-07 |
| Estimated survival rate based on lifetime of typical mammal species | Ord (2020) | 1.00E-06 |
| Midpoint of estimated survival interval for all species | Ord (2020) | 5.50E-07 |
| Estimated rate of mass extinction events | Ord (2020) | 1.00E-08 |

# Power of Social Organizations (Governments and Corporations)

In this section, I look at the share of global resources and economic activity controlled by agents more intelligent than humans, specifically corporations and governments. The central idea is that corporations and governments, by combining the minds of a large number of individuals, are able to achieve superhuman intelligence.

## What This Reference Class Captures

This reference class captures the extent to which the only superhuman intelligent agents we know of (plausibly) have gained power in the world. I use various measures of economic activity, culture, and human autonomy to assess the degree to which governments and corporations control resources otherwise available to humans. I think about this similarly to the case of newly intelligent species or genera.

I do not look at extinction in this section. I considered including figures on the vanishing of indigenous people (e.g. 90% of minority languages projected to go extinct according to Graddol, David. "The future of language." Science 303.5662 (2004):

1329-1331.) and self-employment (currently 10.1% of people in the U.S. but virtually 100% before the Industrial Revolution). I concluded that this would not really reflect something like extinction, however, since much of what it captures is absorption or assimilation.

## Why Is This Reference Class Informative (or Not)?

I find this reference class somewhat less informative than the reference class of relatively intelligent species and genera but still somewhat informative (more so than the reference class of current AI systems). The main problem with this reference class is that the superintelligence consists of intelligent agents, so distinguishing governments' and corporations' resources from individuals' resources is difficult to do. I offer my best working attempt at this, but it is highly imperfect and even more subject to interpretation than the species reference class.

## Should We Expect These Benchmarks to Be Too High or Too Low?

I do not see much reason to expect these estimates to be too high or too low. They would be too low if we think corporations or governments are less intelligent than a superhuman AI would be. They might be too high since governments and corporations consist of individuals, so individuals willingly yield their resources to governments and corporations in a way humans might not with superhuman AI.

## Estimates

### Share of Resources Controlled by Social Organizations

Range of estimates: [0.0939, 0.999]
Weighted arithmetic mean: 0.401
Weighted geometric mean: 0.588

| **Measure** | **Source** | **Estimate** | **Notes** |
|---|---|---|---|
| Government spending by GDP, US | Roser Esteban Ortiz-Ospina and Max Roser. "Government Spending." Our world in data (2013). | 4.37E-01 | |
| Central government spending by GDP, US | Roser Esteban Ortiz-Ospina and Max Roser. "Government Spending." Our world in data (2013). | 3.30E-01 | Plausibly, the central government is a better analogue to a superintelligent being than government as a whole. |
| Share of humans under who are citizens (i.e. not stateless) of a nation-state | United Nations High Commissioner for Refugees. "UNHCR Global Trends 2014: World at War". Refworld (2015).<br>UN News. "'12 million' stateless people globally, warns UNHCR chief in call to States for decisive action." UN (2018). | 9.99E-01 | Averaging two sources on statelessness given that the higher one says "may be" stateless, and the lower one is undercounted. |

| | | | |
|---|---|---|---|
| Share of people employed by government in OECD countries as of 2019. | “Employment in government as share of total employment in OECD countries in 2007, 2009, 2017, and 2019.” Statista (2022). | 1.79E-01 | |
| Corporate assets as share of global assets | Woetzel et al, "The rise and rise of the global balance sheet," McKinsey, November 2021. | 5.36E-01 | Exhibit 14 |
| Government assets as share of global assets | Woetzel et al, "The rise and rise of the global balance sheet," McKinsey, November 2021. | 9.39E-02 | Exhibit 14 |
| Corporate assets as share of global assets | Woetzel et al, "The rise and rise of the global balance sheet," McKinsey, November 2021. | 6.30E-01 | Exhibit 14 |

### Share of Resources Controlled by Social Organizations, First Ten Years

I recommend the same approach here as for the species-based reference classes, 50% weight on clock time and 50% weight on computational time. Government capture of resources happened very quickly in computational time, however, so the computational weight is one rather than 22%.

### Share of Resources Controlled by Social Organizations, First Fifty Years

This estimate is the same as that for ten years.

## Naïve Posteriors from Historic Inventions

This section is a bit different from previous sections. In this section, I look at what we can say about the impacts of superhuman AI based on previously existing technologies given no other information. I look at what different evaluations of the history of technology would imply for how high the likelihood of extinction an be from a threatening technology. I also consider the likelihood of AI transforming the world in a way on par with the Industrial Revolution or the invention of writing.

## What This Reference Class Captures

For extinction, this reference class essentially tells us what plausible views are on the likelihood of extinction from superhuman AI based on the fact that other technologies have been invented which an observer at the time might reasonably have feared would cause extinction. I call these reference classes naïve because I incorporate no other information except a Jeffreys prior over the likelihood of extinction from superhuman AI.

For the likelihood of transformation, the reference class performs a similar role, but given that we have seen technologies transform the world, the prior ends up being less relevant. As a result, this reference class is similar to other reference classes above for transformation: it gives us a base rate for the likelihood that a potentially transformative technology actually does transform the world.

## Why Is This Reference Class Informative (or Not)?

I generally find this reference class informative for what it is. I take the estimates for the likelihood of transformation to be fairly informative. For extinction, I think the naïve prior should perhaps play the role of an upper bound since it incorporates no other information (e.g. the fact that humans have been around for hundreds of thousands of years). The 99th percentile estimates definitely appear to be upper bounds.

## Should We Expect These Benchmarks to Be Too High or Too Low?

I would expect the likelihood of transformation to be about right, perhaps a bit low since it seems like there is potentially more information to indicate superhuman AI is transformative than I have already incorporated. I expect the likelihood of extinction to be too high for a reference class forecast since it incorporates none of the information from previous sections and starts from a 50/50 prior.

## Estimates

### Chance of Extinction Within 10 Years of a Threatening Invention

Range of estimates: [.005, .059]
Weighted arithmetic mean: .0363
Weighted geometric mean: .0254

| Measure | Source | Estimate | Notes |
|---|---|---|---|

| | | | |
|---|---|---|---|
| Chance of extinction from a given category of threatening inventions given a Beta (0.5, 0.5) prior | Timeline of historic inventions, Other Inventions Worksheet | 0.045 | I pulled out all inventions where it seemed like I could construct a story for how they would risk extinction-like catastrophe and counted the number of fairly distinct categories, i.e. categories within which one invention did not render an earlier one obsolete.<br>The estimate is just one divided by [two plus two times the number of inventions]. |
| Chance of extinction from a given threatening invention given a Beta (0.5, 0.5) prior, first definition | Timeline of historic inventions, Other Inventions Worksheet | 0.005 | Similar to the previous row but per invention, not category. |
| Chance of extinction from a given threatening invention given a Beta (0.5, 0.5) prior, second definition | Timeline of historic inventions, Other Inventions Worksheet | 0.059 | I counted the share of emerging technologies with plausible claims of extinction risk. I then used this to estimate how many past technologies I would have found to be plausible sources of extinction.<br>The estimate is just one divided by [two plus two times the number of inventions]. |
| Chance of extinction from a given category of threatening inventions, 99th percentile | Timeline of historic inventions, Other Inventions Worksheet | 0.369 | See rows up for the distribution.<br>The estimate is just one minus 0.01 to the power of one over the number of inventions. |
| Chance of extinction from a given threatening invention, 99th percentile, first definition | Timeline of historic inventions, Other Inventions Worksheet | 0.049 | See rows up for the distribution.<br>The estimate is just one minus 0.01 to the power of one over the number of inventions. |
| Chance of extinction from a given threatening invention, 99th percentile, second definition | Timeline of historic inventions, Other Inventions Worksheet | 0.461 | See rows up for the distribution.<br>The estimate is just one minus 0.01 to the power of one over the number of inventions. |

### Chance of Extinction Within 50 Years of a Threatening Invention

Range of estimates: [.006, .063]
Weighted arithmetic mean: .0394
Weighted geometric mean: .0278

| Measure | Source | Estimate | Notes |
|---|---|---|---|
| Chance of extinction from a given category of threatening inventions given a Beta (0.5, 0.5) prior | Timeline of historic inventions, Other Inventions Worksheet | 0.050 | See previous table.<br>The estimate is just one divided by [two plus two times the number of inventions]. |
| Chance of extinction from a given threatening invention given a Beta (0.5, 0.5) prior, first definition | Timeline of historic inventions, Other Inventions Worksheet | 0.006 | See previous table. |
| Chance of extinction from a given threatening invention given a Beta (0.5, 0.5) prior, second definition | Timeline of historic inventions, Other Inventions Worksheet | 0.063 | See previous table. |
| Chance of extinction from a given category of threatening inventions, 99th percentile | Timeline of historic inventions, Other Inventions Worksheet | 0.401 | See previous table. |
| Chance of extinction from a given threatening invention, 99th percentile, first definition | Timeline of historic inventions, Other Inventions Worksheet | 0.055 | See previous table. |
| Chance of extinction from a given threatening invention, 99th percentile, second definition | Timeline of historic inventions, Other Inventions Worksheet | 0.487 | See previous table. |

### Chance of Transformation Within 10 Years of a Major Invention

Range of estimates: [.014, .195]
Weighted arithmetic mean: .1497
Weighted geometric mean: .1003

| Measure | Source | Estimate | Notes |
|---|---|---|---|
| Chance of world alteration from a potentially transformative invention given a Beta (0.5, 0.5) prior | Timeline of historic inventions, Other Inventions Worksheet | 0.195 | I pulled out all inventions where it seemed like I could construct a story for how they would fundamentally alter the world and counted the number of fairly distinct categories, i.e. categories |

| | | | |
|---|---|---|---|
| | | | within which one invention did not render an earlier one obsolete.<br>The estimate is one half plus the number of actual historical transformations to date divided by [one plus the number of inventions]. |
| Chance of world alteration from a historic invention given a Beta (0.5, 0.5) prior | Timeline of historic inventions, Other Inventions Worksheet | 0.014 | I counted the share of emerging technologies with plausible claims of ability to fundamentally alter the world.<br>The estimate is one half plus the number of actual historical transformations to date divided by [one plus the number of inventions]. |
| Chance of world alteration from a potentially transformative invention, 99th percentile | Timeline of historic inventions, Other Inventions Worksheet | 0.575 | See previous table. |
| Chance of world alteration from a historic invention, 99th percentile | Timeline of historic inventions, Other Inventions Worksheet | 0.057 | See previous table. |

### Chance of Transformation Within 50 Years of a Major Invention

Range of estimates: [.013, .178]
Weighted arithmetic mean: .1367
Weighted Weighted geometric mean: .0917

| Measure | Source | Estimate | Notes |
|---|---|---|---|
| Chance of world alteration from a potentially transformative invention given a Beta (0.5, 0.5) prior | Timeline of historic inventions, Other Inventions Worksheet | 0.178 | See previous table. |
| Chance of world alteration from a historic invention given a Beta (0.5, 0.5) prior | Timeline of historic inventions, Other Inventions Worksheet | 0.013 | See previous table. |
| Chance of world alteration from a potentially transformative invention, 99th percentile | Timeline of historic inventions, Other Inventions Worksheet | 0.537 | See previous table. |

| | | | |
|---|---|---|---|
| Chance of world alteration from a historic invention, 99th percentile | Timeline of historic inventions, Other Inventions Worksheet | 0.051 | See previous table. |

## Damages from and Power of AI Systems to Date

In this section, I look at what we can say about the likelihood of extinction from or takeover by a major new AI system. I draw heavily on the AI Incidents Database from the Center for Security and Emerging Technology for the extinction risk work.

I do not find this reference class particularly informative for extinction. It is uninformative because there have not been all that many incidents involving AI systems killing people—no more than a few dozen, most reasonably fewer than ten. With such sparse data, there is no real way to fit a distribution and extrapolate. I do my best to fit a square peg into a round hole in this section, but *all extrapolation in this section merits great skepticism.*

For takeover, this reference class basically tells us AI systems already have the potential to automate much of the economy if we extrapolate current trends. The share of resources that are in a sense controlled by automation could be quite large just based on current trends.

### What This Reference Class Captures

This reference class tries to say what we should predict if we take future AI systems as similar to today's. If we think of damages from future systems as drawn from the same distribution as damages from today's current ones, and if this reference class is informative about the distribution, it will capture the distribution of future damages. The key assumption that I do not think holds for thinking about extinction with this reference class is that it is informative about the distribution.

Another issue with this reference class is figuring out what exactly is equivalent to a major AI system. The annual measure avoids this issue since I can simply look at the distribution of damages over each year since 1982 (when the first incidents in the dataset are recorded). For the event-specific measure of extinction risk, I define a major AI system in a few different ways, from narrowest to broadest "any major AI company over its lifetime", "any AI company over its lifetime," "any AI patent", "any AI journal article". I ask what the rate of critical incidents is for companies, patents, or journal articles. I also offer estimates for the likelihood of a critical incident if we randomly redrew as many companies, patents, or journal articles as have existed since 1982 as a rough measure of AI risk from the creation of a type of system that then gets built by a large number of different actors. The latter naturally yields much higher estimates.

Last, it is unclear what incident from an AI system we should take to capture the sort of incident that would generate catastrophic risk. CSET classifies incidents in terms of severeity, with critical meaning "many humans were or were almost killed, or that financial, property, social, or political interests were seriously disrupted at a national or global scale (or nearly so disrupted)". On some views, this might be sufficient for catastrophe, but I think that is a bit too broad. I also offer extrapolations to the likelihood of an incident killing 8 billion people, but these have little support in the data.

With regard to the possibility of takeover, this reference class tells us what share of the economy might be automated on reasonable timeframes. That also relies on the current distribution, but extrapolated naïvely based on growth rates.

## Why Is This Reference Class Informative (or Not)?

For the probability of extinction, I find this reference class uninformative. Mostly this is because the data are sparse, as noted above. We are toying with a handful of examples. Any distributional "fit" is tenuous at best. This problem is illustrated by something that will look puzzling: the annual likelihood of 8 billion people dying from AI is higher than the likelihood of 8 billion people dying from any particular company, patent, or even all companies or patents since 1982. This is basically a fluke of distributional fit.

For takeover, the precise estimates are more informative. Roughly, there is good reason to think based on current trends that future AI systems could control much of the economy. That is what this reference class seeks to capture. In the long run, naïve extrapolations appear to run up against the limits of the methodology because automation should be at least somewhat constrained by factors other than AI capabilities, but strictly naïve extrapolations of positive growth in the share of the economy that is automated eventually get to 100%.

## Should We Expect These Benchmarks to Be Too High or Too Low?

If these estimates were correct for what they are (i.e. the extrapolations were valid), we should expect estimates to be low relative to what the actual frequency would be fore more advanced AI systems. Since AI systems will likely grow more powerful in the future, these estimates should be too small.

However, for current AI systems, the frequency of "critical" incidents is an upper bound on the likelihood of an extinction-scale outcome outcome since "critical" is well short of extinction. For that reason, those estimates, without multiplying by the likelihood that a critical incident kills 8 billion people may be too high relative to what the likelihood of extinction would be from a more advanced system.

## Estimates

I downloaded the AI Incidents database, which reports problems with AI systems. I imputed the numbers of deaths directly attributed to AI systems, deaths linked to AI systems, and incidents linked to AI systems. I also gathered a reports that quantified the economics of AI in ways that indicated what share of resources or global activity AI accounts for or will account for soon.

See AI_Incidents in the spreadsheet for more detail.

### Likelihood of an AI System Killing 8 Billion People

*These numbers come from the numbers in the tables in the following two tables multiplied together, divided by 0.1333 and then averaged. The first table captures the likelihood of a "critical" incident, and the second the likelihood of a "severe" incident killing 8* *billion people. I fit a distribution for severe rather than critical to increase the amount of data used, but I use "critical" in the first table because it offers a narrower upper bound on the likelihood of catastrophe than "severe" does.*

Range of estimates: [0, 6.24 x $10^{-6}$]
Weighted arithmetic mean: 1.12 x $10^{-6}$
Weighted arithmetic mean (dropping extreme values): 2.29 x $10^{-6}$
Weighted geometric mean (dropping extreme values): 2.74 x $10^{-8}$

### Likelihood of an AI System Having a Critical Incident

*A critical incident is almost certainly an extremely high upper bound on the likelihood of an incident that would be catastrophic, i.e. kill 8 billion people.*

Range of estimates: [0, 1]
Weighted arithmetic mean: 0.477
Weighted geometric mean: 0.969

| Metric | Source | Estimate | Notes |
|---|---|---|---|
| Share of post-1982 AI companies with a critical AI incident reported | Analysis using AI Incidents Database<br>Daniel Zhang et al, "The AI Index 2022 Annual Report," AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University, March 2022. | 3.44E-04 | See AI_Incidents worksheet for more detail. |
| Share of major AI companies with a critical AI incident reported | Ibid. | 1.11E-01 | |
| Share of times there is a critical AI incident if we rerun all post-1982 companies' histories using inferred probability and assume across-company independence | Ibid. | 9.93E-01 | |
| Share of times there is a critical AI incident if we rerun major companies' histories using inferred probability and assume across-company independence | Ibid. | 6.54E-01 | |
| Annual rate of critical AI incidents per AI-related publication, post-1982 | Ibid. | 1.57E-02 | |

|  |  |  |  |
|---|---|---|---|
| weighted average |  |  |  |
| Share of times there is a critical AI incident if we regenerate as many AI-related publications as there have been post-1982 using inferred probability and assume across-company independence | Ibid. | 1.00E+00 |  |
| Annual rate of critical AI incidents per AI-related patent, post-1982 weighted average | Ibid. | 3.96E-02 |  |
| Share of times there is a critical AI incident if we regenerate as many AI-related patents as there have been post-1982 using inferred probability and assume across-company independence | Ibid. | 1.00E+00 |  |

### Likelihood of a Critical Incident Killing 8 Billion People

*Multiply these numbers with the previous subsection for an estimate of total risk from a new AI system.*
Range of estimates: [0, 1.31 x $10^{-5}$]
Weighted arithmetic mean: 1.12 x $10^{-6}$
Weighted arithmetic mean (winsorised): 1.12 x $10^{-6}$
Weighted geometric mean (dropping extreme values): 8.36 x $10^{-10}$

| Metric | Source | Estimate | Notes |
|---|---|---|---|
| Extrapolating using normal distribution over number directly killed | Analysis using AI Incidents Database | 0.00E+00 | NOTE: These estimates are very tenuous. No tests for distributional fit pass for any row in this table. |
| Extrapolating using normal distribution over number directly killed, dropping zeroes | Ibid. | 0.00E+00 | NOTE: These estimates are very tenuous. No tests for distributional fit pass for any row in this table. |
| Extrapolating using lognormal distribution over number directly | Ibid. | 0.00E+00 | NOTE: These estimates are very tenuous. No tests for distributional fit |

| | | | |
|---|---|---|---|
| killed, dropping zeroes | | | pass for any row in this table. |
| Extrapolating using Pareto distribution over number directly killed, dropping zeroes | Ibid. | 4.87E-09 | NOTE: These estimates are very tenuous. No tests for distributional fit pass for any row in this table. |
| Extrapolating using normal distribution over number directly or indirectly killed | Ibid. | 0.00E+00 | NOTE: These estimates are very tenuous. No tests for distributional fit pass for any row in this table. |
| Extrapolating using normal distribution over number directly or indirectly killed, dropping zeroes | Ibid. | 0.00E+00 | NOTE: These estimates are very tenuous. No tests for distributional fit pass for any row in this table. |
| Extrapolating using lognormal distribution over number directly or indirectly killed, dropping zeroes | Ibid. | 0.00E+00 | NOTE: These estimates are very tenuous. No tests for distributional fit pass for any row in this table. |
| Extrapolating using Pareto distribution over number directly or indirectly killed, dropping zeroes | Ibid. | 3.28E-07 | NOTE: These estimates are very tenuous. No tests for distributional fit pass for any row in this table. |
| Extrapolating using normal distribution over number directly or indirectly killed or repressed | Ibid. | 0.00E+00 | NOTE: These estimates are very tenuous. No tests for distributional fit pass for any row in this table. |
| Extrapolating using normal distribution over number directly or indirectly killed or repressed, dropping zeroes | Ibid. | 0.00E+00 | NOTE: These estimates are very tenuous. No tests for distributional fit pass for any row in this table. |
| Extrapolating using lognormal distribution over number directly or indirectly killed or repressed, dropping zeroes | Ibid. | 1.21E-10 | NOTE: These estimates are very tenuous. No tests for distributional fit pass for any row in this table. |
| Extrapolating using Pareto distribution over number directly or indirectly killed or | Ibid. | 1.31E-05 | NOTE: These estimates are very tenuous. No tests for distributional fit pass for any row in this |

| | | | |
|---|---|---|---|
| repressed, dropping zeroes | | | table. |

## Annual Likelihood of 8 Billion Deaths from AI

Range of estimates: [0, 1.04 x $10^{-4}$]
Weighted arithmetic mean: 1.44 x $10^{-5}$
Weighted arithmetic mean (winsorised): 3.18 x $10^{-5}$
Weighted geometric mean (winsorised): 2.78 x $10^{-5}$

| Metric | Source | Estimate | Notes |
|---|---|---|---|
| Extrapolating using normal distribution over number directly killed | Analysis using AI Incidents Database | 0.00E+00 | NOTE: These estimates are very tenuous. No tests for distributional fit pass for any row in this table. |
| Extrapolating using normal distribution over number directly killed, dropping zeroes | Ibid. | 0.00E+00 | NOTE: These estimates are very tenuous. No tests for distributional fit pass for any row in this table. |
| Extrapolating using lognormal distribution over number directly killed, dropping zeroes | Ibid. | 0.00E+00 | NOTE: These estimates are very tenuous. No tests for distributional fit pass for any row in this table. |
| Extrapolating using Pareto distribution over number directly killed, dropping zeroes | Ibid. | 2.32E-05 | NOTE: These estimates are very tenuous. No tests for distributional fit pass for any row in this table. |
| Extrapolating using normal distribution over number directly or indirectly killed | Ibid. | 0.00E+00 | NOTE: These estimates are very tenuous. No tests for distributional fit pass for any row in this table. |
| Extrapolating using normal distribution over number directly or indirectly killed, dropping zeroes | Ibid. | 0.00E+00 | NOTE: These estimates are very tenuous. No tests for distributional fit pass for any row in this table. |
| Extrapolating using lognormal distribution over number directly or indirectly killed, dropping zeroes | Ibid. | 0.00E+00 | NOTE: These estimates are very tenuous. No tests for distributional fit pass for any row in this table. |
| Extrapolating using | Ibid. | 4.55E-05 | NOTE: These estimates |

| | | | |
|---|---|---|---|
| Pareto distribution over number directly or indirectly killed, dropping zeroes | | | are very tenuous. No tests for distributional fit pass for any row in this table. |
| Extrapolating using normal distribution over number directly or indirectly killed or repressed | Ibid. | 0.00E+00 | NOTE: These estimates are very tenuous. No tests for distributional fit pass for any row in this table. |
| Extrapolating using normal distribution over number directly or indirectly killed or repressed, dropping zeroes | Ibid. | 0.00E+00 | NOTE: These estimates are very tenuous. No tests for distributional fit pass for any row in this table. |
| Extrapolating using lognormal distribution over number directly or indirectly killed or repressed, dropping zeroes | Ibid. | 0.00E+00 | NOTE: These estimates are very tenuous. No tests for distributional fit pass for any row in this table. |
| Extrapolating using Pareto distribution over number directly or indirectly killed or repressed, dropping zeroes | Ibid. | 1.04E-04 | NOTE: These estimates are very tenuous. No tests for distributional fit pass for any row in this table. |

## Forecasted AI Share of the Economy within Ten Years of Superhuman AI

Range of estimates: [0.361, 0.588]
Weighted arithmetic mean: 0.258
Weighted geometric mean: 0.249

| Measure | Source | Estimate | Notes |
|---|---|---|---|
| Share of 2017 work activities automated | Manyika et al. "Harnessing automation for a future that works," McKinsey, January 2017. | 3.61E-01 | See AI_Incidents worksheet for more detail. |
| Share of 2017 work activities automated, adjusted for alternative report | Manyika et al, "Harnessing automation for a future that works," McKinsey, January 2017.<br>Briggs, Joseph and Devesh Kodnani, “The Potentially Large Effects of Artificial Intelligence on Economic Growth," Goldman Sachs, April 2023 | 1.63E-01 | See AI_Incidents worksheet for more detail. |

| | | | |
|---|---|---|---|
| Share of GDP from automation in FTEs, US | Manyika et al, "Harnessing automation for a future that works," McKinsey, January 2017. | 2.97E-01 | See AI_Incidents worksheet for more detail. |
| Share of GDP from automation in FTEs, China | Manyika et al, "Harnessing automation for a future that works," McKinsey, January 2017. | 2.27E-01 | See AI_Incidents worksheet for more detail. |
| Share of GDP from automation in FTEs, Brazil | Manyika et al, "Harnessing automation for a future that works," McKinsey, January 2017. | 3.04E-01 | See AI_Incidents worksheet for more detail. |
| Share of GDP from automation in FTEs, Saudi Arabia | Manyika et al, "Harnessing automation for a future that works," McKinsey, January 2017. | 3.08E-01 | See AI_Incidents worksheet for more detail. |
| Share of GDP from automation in FTEs, Nigeria | Manyika et al, "Harnessing automation for a future that works," McKinsey, January 2017. | 3.09E-01 | See AI_Incidents worksheet for more detail. |

### Forecasted AI Share of the Economy within Fifty Years of Superhuman AI

Range of estimates: [0.343, 0.693]
Weighted arithmetic mean: 0.305
Weighted geometric mean: 0.558

| Measure | Source | Estimate | Notes |
|---|---|---|---|
| Share of 2017 work activities automated | Manyika et al, "Harnessing automation for a future that works," McKinsey, January 2017. | 6.93E-01 | See AI_Incidents worksheet for more detail. |
| Share of 2017 work activities automated, adjusted for alternative report | Manyika et al, "Harnessing automation for a future that works," McKinsey, January 2017.<br>Briggs, Joseph and Devesh Kodnani, "The Potentially Large Effects of Artificial Intelligence on Economic Growth," Goldman Sachs, April 2023 | 5.28E-01 | See AI_Incidents worksheet for more detail. |
| Share of GDP from automation in FTEs, US | Manyika et al, "Harnessing automation for a future that works," McKinsey, January 2017. | 5.84E-01 | See AI_Incidents worksheet for more detail. |
| Share of GDP from automation in FTEs, China | Manyika et al, "Harnessing automation for a future that works," McKinsey, January 2017. | 5.25E-01 | See AI_Incidents worksheet for more detail. |
| Share of GDP from automation in FTEs, Brazil | Manyika et al, "Harnessing automation for a future that works," McKinsey, January 2017. | 5.93E-01 | See AI_Incidents worksheet for more detail. |

| | | | |
|---|---|---|---|
| Share of GDP from automation in FTEs, Saudi Arabia | Manyika et al. "Harnessing automation for a future that works," McKinsey, January 2017. | 5.22E-01 | See AI_Incidents worksheet for more detail. |
| Share of GDP from automation in FTEs, Nigeria | Manyika et al. "Harnessing automation for a future that works," McKinsey, January 2017. | 6.35E-01 | See AI_Incidents worksheet for more detail. |

#### Forecasted Long-Term AI Share of the Economy

Naïve extrapolations indicate total control by AI over the economy in the long run given positive growth rates now, i.e. an estimate of 100%. In the AI_Incidents worksheet, you can see figures for the year 2173, which are very close to one.

I suspect this is unrealistic and would not hold in a more refined model. For that reason and given that the results are all essentially one, I do not include a table of estimates here.

## Rates of Product Defects

In this section, I look at how often products have serious or catastrophic defects. I could not find any overall data that captured something like all consumer products, so I look at four cases I can find measures for: cars, drugs, meat, and overall standards for acceptable cancer risk, all in the U.S. I try to find the frequency of serious defects and also to estimate the likelihood a product causes death via a defect.

### What This Reference Class Captures

This reference class captures how often products go seriously awry measured either by society's standards or by their ability to kill their users. Society's standards are a tricky measure because they are endogenous: the more concerned society is with safety, the higher this number will be. Indeed, in the case of cars we see a steady rise in the share of cars with reported defects over time, though there are hints of a plateau since 2000. This appears to reflect increases in safety. Nevertheless, if we expect that catastrophe from superhuman AI would involve its violating society's standards, this offers a measure of that. Ability to kill users is plausibly not a standard that is endogenous to society's views, but it requires some rough adjustments to estimate.

### Why Is This Reference Class Informative (or Not)?

I do not find this reference class particularly informative, perhaps the second-least informative after the reference class of current AI systems. As with that reference class, there is a difficulty in figuring out what sort of defect is bad enough to be similar to be catastrophic, or to be extinction-level if found in superhuman AI.

### Should We Expect These Benchmarks to Be Too High or Too Low?

I would expect the estimates based on recalls with risk of death to be upper bounds on the sort of defect that could cause extinction. A defect in an advanced AI system that would be catastrophic seems much more severe than a defect in a product that kills its user. I think there are plausible views, though, on which any serious defect in a superhuman AI causes catastrophe, so I place some weight on these estimates.

I expect estimates based on risk of death to be perhaps too low since these products do not seem as threatening as advanced AI, and they seem easier to monitor, so they probably are less lethal.

## Estimates

### Share of Products with a Serious Defect

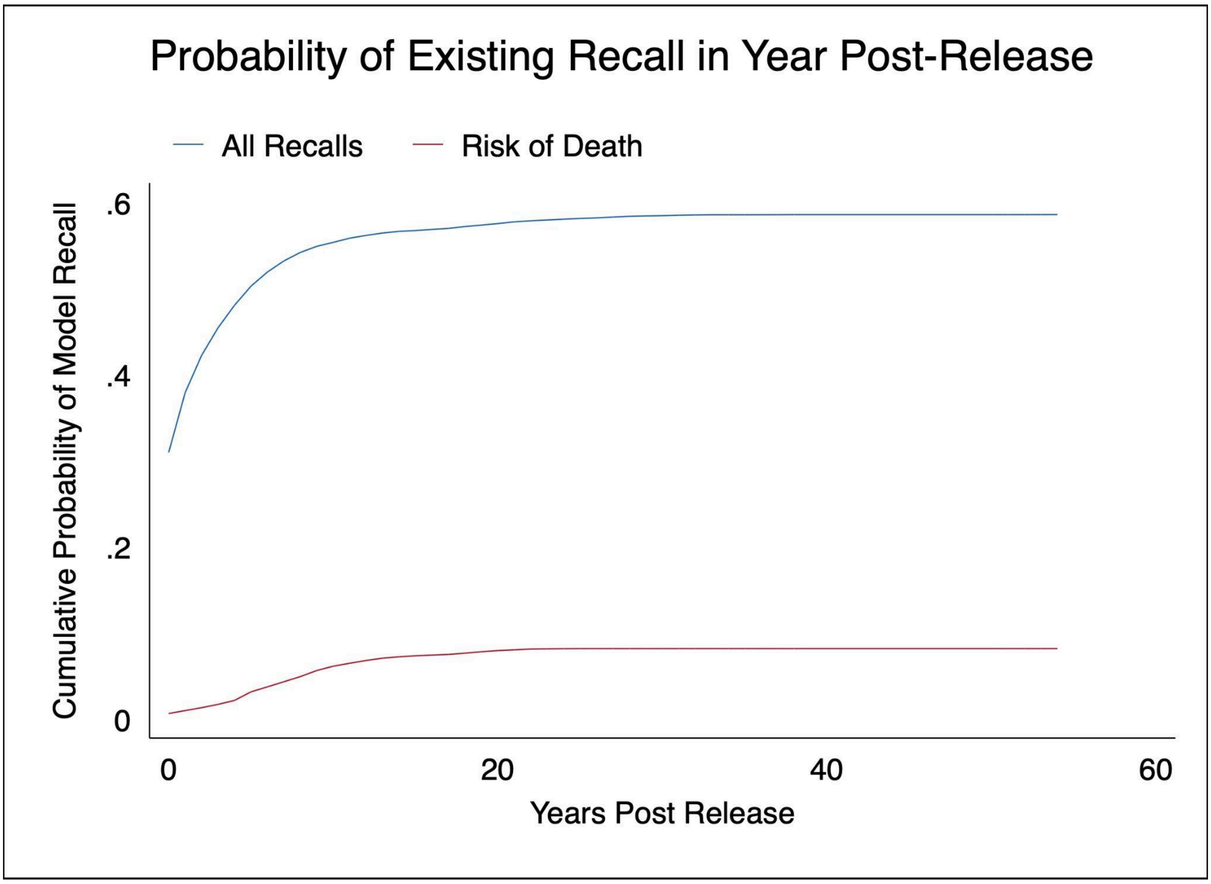
Range of estimates: [$2.1 \times 10^{-10}$, $5.88 \times 10^{-1}$]
Weighted arithmetic mean: $2.29 \times 10^{-2}$
Weighted geometric mean: $6.45 \times 10^{-4}$

| **Measure** | **Source** | **Estimate** | **Notes** |
|---|---|---|---|
| Share of car components in use subject to recall with risk of death, average of maximum and minimum estimates post-2000 | Analysis using Department of Transportation data | 3.26E-03 | |
| Share of car components in use subject to recall, average of maximum and minimum estimates post-2000 | Analysis using Department of Transportation data | 2.21E-02 | Risk of death seems like a strictly narrower upper bound on the risk of a catastrophic defect, but I include this here for completion. |
| Share of cars on the road under a recall notice | Recall Masters, "State of Recalls 2021." May 5 2022. Mar 22 2023. | 2.50E-01 | Same comment as previous cell. |
| Share of cars on the road under a recall notice with risk of death | Estimated by adjusting the Recall Masters report by the first row above divided by the second row. | 3.68E-02 | |
| Likelihood car component owner dies by virtue of defect in that component, best guess | Analysis using Department of Transportation data and news reports | 3.26E-09 | First row above divided by one million, which is my estimate for likelihood of death from a recalled product with risk of death. |
| Likelihood car owner dies by virtue of any defective component, best guess | Estimated by adjusting from two rows up. | 3.68E-08 | Two rows up divided by one million, which is my estimate for likelihood of death from a recalled product with risk of death. |
| Share of car models subject to recall within 50 years of production | Analysis using EPA and NHTSA data | 5.85E-01 | |

| | | | |
|---|---|---|---|
| Share of car models subject to recall with risk of death within 50 years of production | Analysis using EPA and NHTSA data | 8.16E-02 | |
| Share of drugs withdrawn from market because of a safety issue | Downing, Nicholas S., et al. "Postmarket safety events among novel therapeutics approved by the US Food and Drug Administration between 2001 and 2010." Jama 317.18 (2017): 1854-1863. | 1.35E-02 | |
| Share of drugs with a post-market safety event. | Downing, Nicholas S., et al. "Postmarket safety events among novel therapeutics approved by the US Food and Drug Administration between 2001 and 2010." Jama 317.18 (2017): 1854-1863. | 3.20E-01 | |
| Share of meat recalled by weight | Food Safety and Inspection Service. "Summary of Recall Cases in Calendar Year 2021." USDA (2021). | 2.10E-04 | |
| Benchmark US government standard for acceptable cancer risk | Graham, John D. "The legacy of one in a million." Harvard Center for Risk Analysis (1993). | 1.00E-06 | |
| Chance of death from a drug because of a safety issue, assuming withdrawal indicates risk equal to benchmark U.S. cancer risk. | Downing, Nicholas S., et al. "Postmarket safety events among novel therapeutics approved by the US Food and Drug Administration between 2001 and 2010." Jama 317.18 (2017): 1854-1863. | 1.35E-08 | Four rows up multiplied by one-in-a-million. |
| Chance of death from a drug because of a safety issue, assuming post-market safety event indicates risk equal to benchmark U.S. cancer risk. | Downing, Nicholas S., et al. "Postmarket safety events among novel therapeutics approved by the US Food and Drug Administration between 2001 and 2010." *Jama* 317.18 (2017): 1854-1863. | 3.20E-07 | Four rows up multiplied by one-in-a-million. |
| Chance of death from meat , assuming recall indicates risk equal to benchmark U.S. cancer risk. | Food Safety and Inspection Service. "Summary of Recall Cases in Calendar Year 2021." USDA (2021). | 2.10E-10 | Four rows up multiplied by one-in-a-million. |

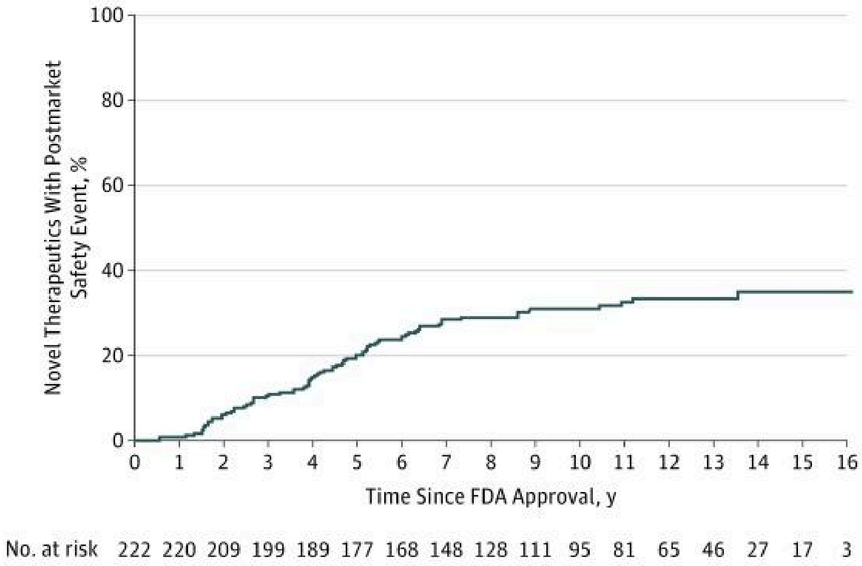
### Share of Products with a Serious Defect in the First 10 or 50 Years

Essentially all defects come out in the first ten years, as the following two graphs show:



Source: analysis using NHTSA and EPA data

Source: Downing, Nicholas S., et al. "Postmarket safety events among novel therapeutics approved by the US Food and Drug Administration between 2001 and 2010." *Jama* 317.18 (2017): 1854-1863.



## Final Lessons

### Overall View

The main overall conclusion of this reference class work is that a wide range of probabilities are consistent with historical reference classes; the playing field for reference class tennis is large. Estimates of existential risk in the single digits or low double digits look perfectly consistent with a reference-class approach.

Some high probabilities require a very large departure from or a very specific selection of reference classes. Reference classes offer some reason to think the likelihood of extinction is not in the mid-to-high double digit percentage. This would be surprising given the past invention of technologies that seem like they would have appeared very threatening *ex ante*. The only individual reference class yielding an extinction probability in the mid-to-high double digits is the share of megafauna wiped extinct by humans. The lower (though significant) extinction rate of other species at the hands of humans and the fact that no other species has caused many extinctions offers reason to think this may be a special case. On the other hand, the fact that the special case is also likely the case of the most intelligent species, this could suggest the threat from superintelligent AGI is in fact larger than the threat from humans to other species.

At the same time, a reference-class approach to thinking about AI risk does not imply that risk from superhuman AI is a Pascalian threat; in fact, there is good reason to think the risk is significant, and there are many analogies consistent with a large chance of catastrophe.

What perhaps most surprised me was the seemingly high probability of takeover-like events relative to human extinction in the reference classes I looked at. Intuitively, this comes from a few places. First, AI does look primed to run much of what would be the domain of humans in the medium term. Second, both humans and human social organizations offer examples of intelligence and have captured a staggering share of global resources. Third, while we have not witnessed human extinction from the hundreds or dozens of significant technologies in history (depending on the count), we have witnessed transformation, so the possibility that AI will transform the world is significantly less surprising. This supports worries about lock-in from advanced AI and Will MacAskill's argument in *What We Owe The Future*. I was surprised to happen upon this.

That said, it is worth keeping in mind that there may be things that we should view as functionally similar to human extinction that are not captured under the likelihood of extinction proper. Permanent disempowerment of humanity that leaves some humans alive would not be extinction the way I am considering it. A world where humans are earthbound by force or manipulation seems like it could be an order of magnitude more likely than extinction. It seems to imply a similar curtailing of human potential, though.

## Some Inside-View Lessons

Looking at reference classes for superhuman AI is a quintessential outside-view exercise, but it suggests some interesting inside-view takeaways.

First, the extinction of megafauna offers a sort of historical analogy to worries about fast AI takeoff. Humans seem to have wiped out 62%-80% of megafauna in the Americas and Australia, where they did not evolve and instead arrived suddenly, compared to 11%-26% in Eurasia and Africa. That is, sudden arrival of this new intelligent species increased the likelihood a megafaunal species would go extinct by four to six times.

Second, it might be possible to make progress benchmarking the cost-effectiveness of generic existential risk reduction by looking at some of the reference classes I considered. Asteroid risk is an obvious one, but biodiversity efforts might also offer a benchmark for how cost-effective preventing extinction is. Preventing nonhuman species from going extinct requires some level of global coordination to prevent a sort of lock-in event. The cost-effectiveness of that work might offer some guidance for thinking about the human case.

Third, toying with risks from AI incidents underlines a very basic mathematical downside of a world with a large number of superhuman AI systems compared to just one that at least I had not really considered. If there is some additional risk from each additional system deployed, the more systems that are deployed, the higher the risk goes.[7] I considered the case where the risk from each additional system was statistically independent. That is an unrealistic simplification, but it serves to illustrate the point. When I take there to be a constant risk from each additional AI system and forecast based on taking each company, patent, or publication as an additional system, I get an estimate of extinction risk from superhuman AI that is one or two orders of magnitude higher than when I take the risk to be a function of whether any superhuman AI is developed or not.

## Future Directions

This document offers a fairly rough, non-academic exploration of reference classes for risk from superhuman AI. For each of the reference classes I considered or any I missed, more work seems valuable. I had to make a number of simple assumptions in this project. It would be good to tweak them, scrutinise the numbers, and see what does and does not hold up. I look forward to seeing what comes of that.

[7] Joe Carlsmith makes a similar point on page 36 of his report.