Base Rates for Catastrophe

Jacob Steinhardt

To communicate risks, we often turn to stories. Nuclear weapons conjure stories of mutually assured destruction, briefcases and red buttons, and nuclear winter. Climate change conjures stories of extreme weather, cities overtaken by rising sea levels, and crop failures. Pandemics require little imagination after COVID, but were previously the subject of movies like *Contagion*.

Stories are great for conveying concrete risks (I myself recently did this for AI risks), but they're a bad way to predict the future. That's because most stories are far too specific to be probable. More importantly, stories tend to feature short, simple chains of causation while reality is complex and multi-causal.

Instead of using stories, most competitive forecasters start their forecasts by looking at historical reference classes. This works really well, and also makes sense: history pulls us away from the biases of story-telling by grounding in events that actually occurred. While history is filtered through narratives, a good history will contend with the complexity of reality, and we can further strip away narrative by grounding in raw numbers.[1]

In this post, I'll use reference classes to understand the largest risks society faces today. I'll do this by considering two different reference classes for historical catastrophes:

- Events that killed a significant fraction of the global human population (Section 1)
- Extinctions of species, and especially mass extinction events (Section 2)

Looking at these reference classes teaches us two things. First, it gives us a numerical estimate of how rare different catastrophes are. If we define a catastrophe as an event killing 1% of the global population within a decade, then 11 such catastrophes have occurred since 1500, for a base rate of 2% per year. If we raise the bar to killing 10% of the population, the base rate drops by an order of magnitude, to 0.2%.

History also gives us qualitative insights. For instance, all the catastrophes in the previous paragraph were epidemics, wars, or famines. Further, many events were multi-causal—the worst epidemics occurred when populations were already weakened by famine, and many epidemics and famines were precipitated by changes in climate or by political turmoil. Species extinctions are also multi-causal, and the common culprits are climate change, natural disasters, invasive species, and humans.

One argument against using historical base rates is that the present is so different from the past (e.g. due to technology) that base rates are meaningless. While today's world is indeed different

[1] Of course, numbers themselves can be misleading, as many historical numbers are based on guesswork! A lot of the work that went into this post was doing extensive reading to decide which numbers to believe.

from the past, base rates can help sharpen rather than neglect these differences, by clarifying what's actually new. For instance, the mere presence of technology cannot move us far above the base rate, because many technologies have been developed throughout history and none has caused a catastrophe in the sense defined above. Instead, we should look for technology that shares properties with the historical drivers of catastrophe: epidemics, famines, wars, political turmoil, climate changes, natural disasters, invasive species, and humans.

I analyze these drivers in detail (Section 3), and find that they fall into a few core groups:

- Natural events that are global or regional in scale (famines, climate change, natural disasters)
- Novel, highly adapted, self-replicating organisms (epidemics, novel pathogens and predators, invasive species)
- Coordinated groups of humans seeking resources, land, or power (wars, political turmoil, extinctions due to overhunting and habitat destruction)

This list makes sense—to have a global impact, something should either start out with a global scale (large natural events), or have a means to get there (self-replication, coordination).

From this perspective, what are the possible drivers of catastrophe in the 21st century? Some answers are obvious from the list above—pandemics, climate change, and major wars continue to be serious threats. Famines are less obviously threatening, as the last major one was in 1961, though preparing for them may still be prudent. And political turmoil, when not itself catastrophic, creates the conditions for other catastrophes to occur.

Turning to new technologies, engineered pathogens are dangerous because they are novel self-replicators, as are certain types of nanotech. Nuclear weapons are dangerous because they have similar effects to natural disasters, and because they increase the worst-case damage from war.

Finally, AI (my own area of study) unfortunately has properties in common with many drivers of catastrophe. It is a novel self-replicator (it can copy itself) that can quickly adapt to new data. AI systems can be trained to coordinate and may seek power, mirroring the threat of coordinated groups of humans. Finally, AI may exacerbate other drivers of catastrophe if it leads to economic unrest and subsequent political turmoil.

Historical Causes of Human Population Loss

To start our analysis, I looked at the largest historical causes of human population loss, as measured by the fraction of the global population that was killed by a given event. To do so, I combined data from the Wikipedia lists of major wars, slavery and other forced labor, famines, epidemics, and natural disasters. I considered other data sources such as technological disasters, but all of these had much smaller death tolls than the five above. The main exception is genocides, as these often co-occurred with wars and are already included in those death tolls, so I excluded them to avoid double-counting.

I wrote a Python script to scrape these sources and aggregate them into a single Pandas dataframe, then filtered to create two datasets:

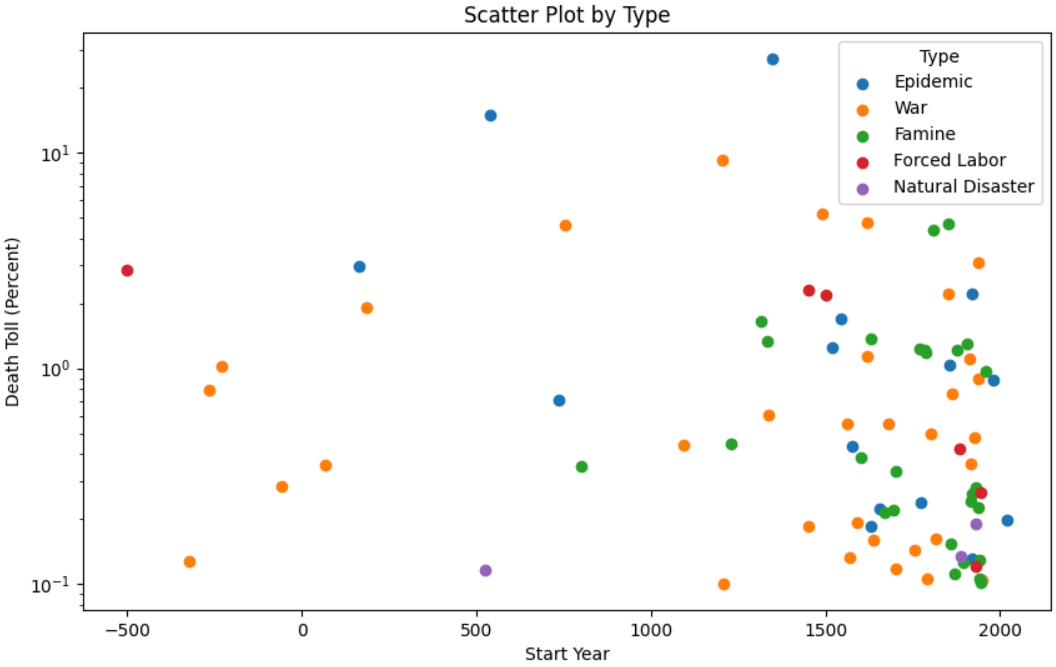
- **Catastrophes**: all events that killed at least 0.1% of the human population, calculated by dividing total deaths by the world population at the start of the event.[2][3]
- **Strict catastrophes:** I further restrict to events that are "fast" (last less than a decade) and in which at least 1% of the human population died.

The set of catastrophes comprises 85 events, of which 80 occurred since 0 CE, and of which 33 were wars, 28 were famines, 15 were epidemics, 6 were forced labor, and 3 were natural disasters. The strict catastrophes comprise 17 events: 5 wars, 8 famines, and 4 epidemics. I include the complete list of strict catastrophes below, as well as a scatter plot of all catastrophes.

[2] Population sizes were collected from Our World in Data.

[3] The Taiping rebellion is double-counted, once as a War and once as a Famine.

Event	Death Toll (Total)	Start	End	Start Population	Death Toll (Percent)	Type
Black Death	122M	1346	1353	450M	27.2%	Epidemic
Plague of Justinian	39M	541	549	261M	14.9%	Epidemic
An Lushan Rebellion	13M	755	763	282M	4.6%	War
World War II	70M	1939	1945	2.27B	3.1%	War
Spanish flu	41M	1918	1920	1.86B	2.2%	Epidemic
Cocoliztli epidemic of 1545–1548	9M	1545	1548	509M	1.7%	Epidemic
Great Famine of 1315–1317	8M	1315	1317	454M	1.7%	Famine
Deccan famine of 1630–1632	7M	1630	1632	538M	1.4%	Famine
Chinese famine of 1333–1337	6M	1333	1337	452M	1.3%	Famine
Famines in east-central China	22M	1907	1911	1.72B	1.3%	Famine
1520 Mexico smallpox epidemic	6M	1519	1520	506M	1.3%	Epidemic
Great Bengal famine of 1770	10M	1769	1773	815M	1.2%	Famine
Famine in India, China, Brazil, Northern Africa (and other countries)	17M	1876	1879	1.38B	1.2%	Famine
Chalisa famine	11M	1783	1784	903M	1.2%	Famine
Doji bara famine or Skull famine	11M	1789	1793	927M	1.2%	Famine
World War I	20M	1914	1918	1.82B	1.1%	War
Qin's wars of unification	2M	-230	-221	196M	1.0%	War

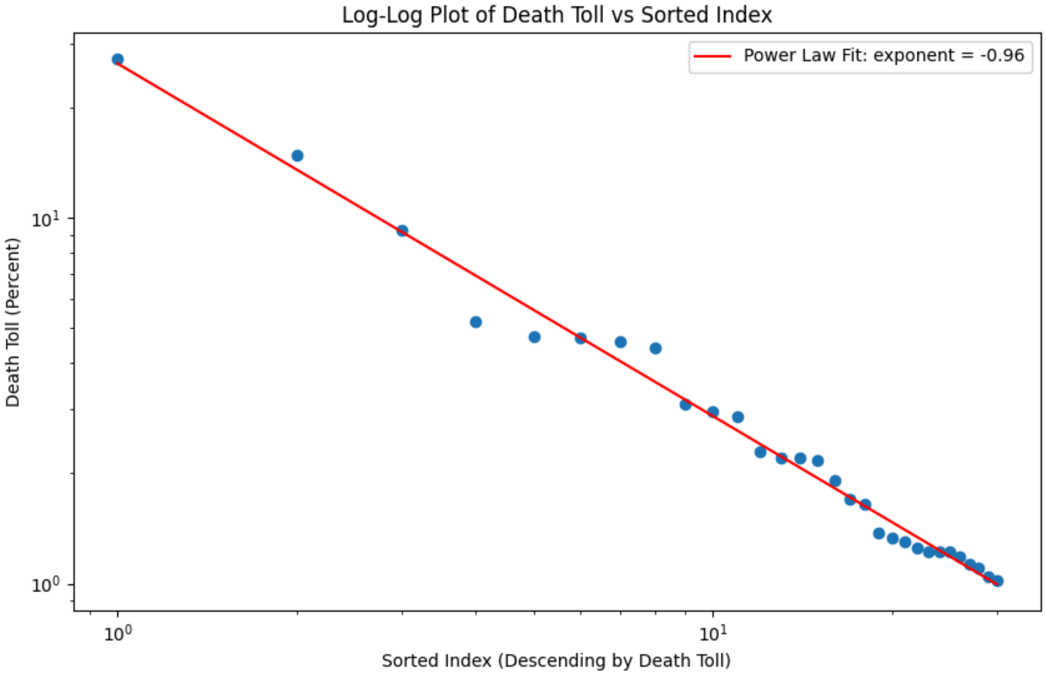


In addition to these historical events, two important prehistoric events are the Toba catastrophe (97% drop in human population, possibly due to a supervolcano) and the 4.2kya event (likely led to global famines, but death toll is unclear).

Reporting bias and base rates. There is very likely reporting bias, as we see the rate of catastrophes "increase" in the 1500s and again in the 1900s, and this happens for all categories including famines (which should decrease over time with better technology). If we start from 1500, there have been 51 catastrophes (0.11/year), and 11 strict catastrophes (0.02/year).

Let's next model how the base rate of (fast) catastrophes[4] varies with their severity. Looking at all catastrophes that lead to at least a 1% population drop, we see an approximately Zipfian distribution: the probability of a catastrophe with a death rate of r is proportional to 1/r.

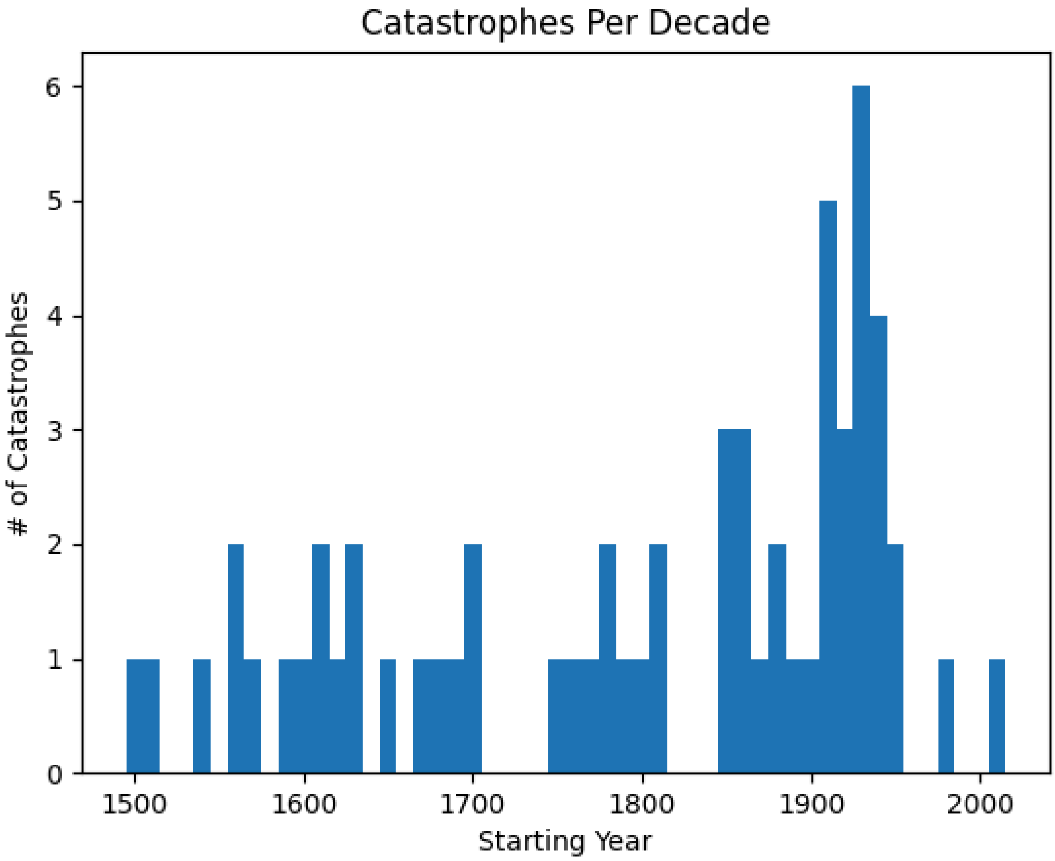
[4] As above, "fast" means taking less than one decade.



Based on this, catastrophes with 10% death rates have an incidence of 0.002/year (once per 5 centuries) and those with a 1% death rate occur 0.02/year (twice per century). While these numbers might seem low, they imply that there is about a **5% chance of a 10%-death-rate catastrophe in the next 25 years** (since 0.002 * 25 = 0.05).

Below 1% death rates, catastrophes are less likely than Zipf's law predicts (see Appendix). For instance, the empirical incidence for 0.1% death rates is 0.08/year (slightly less than once per decade).

Trends over time. If we compute the catastrophes in each decade since 1500, we obtain the following plot:



There were many more catastrophes in the period 1850-1950, although I suspect this is an artifact of reporting bias. Prior to this period, the rate of catastrophes appears roughly constant over time: neither a Ljung-Box test nor a Wald-Wolfowitz test is able to reject the null that catastrophes were identically distributed across decades from 1500-1900 (p=0.36 and 0.26, respectively).

The most notable change over time is the calm period that we are currently in, starting around 1950-1960. Indeed, catastrophes decreased significantly since the first half of the 20th century:

- 9 famines occurred in the first half of the 20th century but only 1 occurred in the second half (Great Chinese Famine, 1959-1961)
- 5 major wars occurred in the first half but only 1 occurred in the second half (Korean War, 1950-1953)
- Epidemics were more constant, with 2 in the first half and 1 in the second half (plus COVID in 2019).

Famines plausibly decreased due to better food production and storage, which is hopefully a lasting improvement. Wars probably decreased due to the Pax Americana, but that unfortunately may now be unwinding with growing global tensions. Thus epidemics and (possibly) wars are the main modern sources of catastrophe so far.

Qualitative analysis: multi-causality. Many catastrophes have multiple causes. For instance, in the predominant theory of the Black Death, climate change was a driver in two ways. First, climate change in Asia led rodents to migrate from mountainous areas to more populated regions, spreading the disease. Second, the Little Ice Age in Europe led to famines, causing populations to be weak and thus more susceptible to disease.[5] Interestingly, the Black Death may have also intensified the Little Ice Age by causing reforestation where populations decreased, leading to carbon capture and subsequent cooling.

To give several other examples of multiple causes:

- In the European colonization of the Americas, most deaths were due to disease rather than war.
- The transition from Ming to Qing was caused by many factors, including disease and famine; the famine itself was possibly caused by the Little Ice Age.
- The Taiping Rebellion was started due to political unrest from famine, and many of the subsequent deaths were caused by drought, famine, and disease rather than military fatalities.
- In general, many famines were caused by climatic events and/or bad government policies.

Overall, this suggests that to reduce the number or intensity of catastrophes, we should attack not just immediate causes, but also more systemic upstream causes.

Species Extinctions

As a second reference class, I considered extinctions of non-human species[6]. This is more difficult to analyze, for several reasons:

- Most extinctions occurred many millions of years ago, so we have only indirect evidence, and there is significant sampling bias based on which species are more easily preserved.
- A species can go extinct if it gradually adapts into a new species, which we might not want to count as a “catastrophe”.
- Some purported mass extinction events may actually be many smaller events occurring over a period of time.

To reduce these difficulties, I will focus on two relatively recent extinction events:

- The Late Quaternary Extinctions (Koch and Barnosky, 2006), which occurred 10,000-50,000 years ago and led to most large mammals becoming extinct.
- The Holocene Extinctions, occurring over the past 10,000 years (and increasing over the past century), primarily driven by human hunting and habitat destruction.

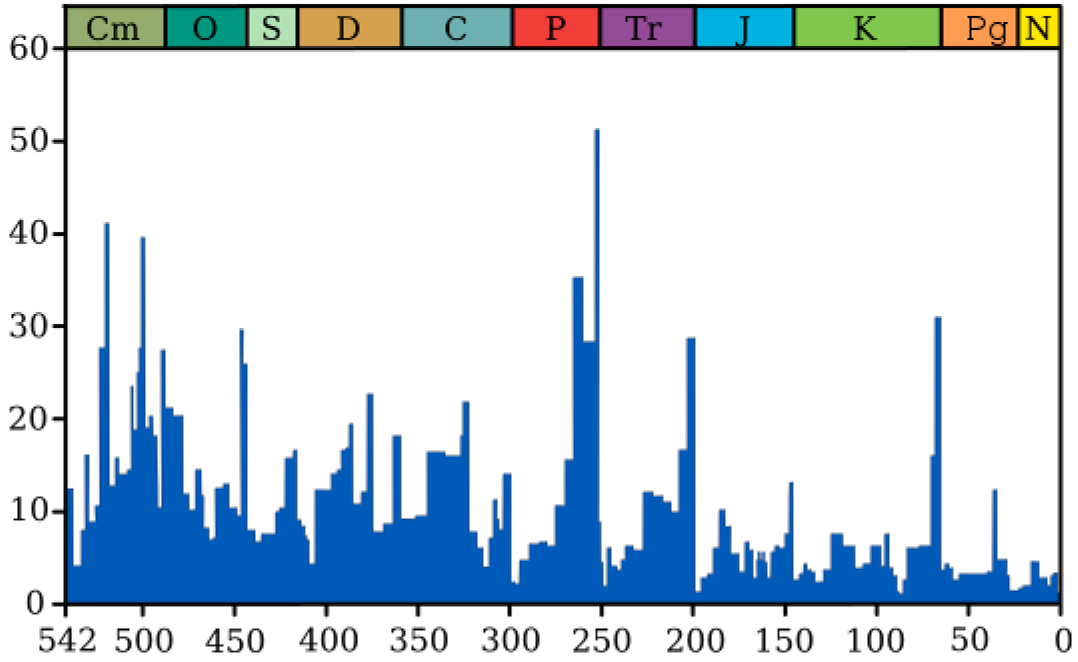
[5] Another theory is that the Mongol invasions (another catastrophe) spread the Black Death, since Mongols threw diseased corpses into cities as a form of biological warfare. This is not currently the predominant theory, but would be another instance of multi-causality, and shows that different major catastrophes can be linked to each other.

[6] Technically speaking, the historical fossil record usually only resolves extinctions at the level of genera rather than species, but I will generally elide this distinction for simplicity.

While most historical mass extinction events were driven by climate change or natural disasters, these two more recent extinctions are argued to have been driven in whole or part by humans. I'll review the evidence and leading theories about both extinction events below.

Historical Base Rate

Before discussing the Quaternary and Holocene extinctions, let's compute a base rate for context. Based on the fossil record, there is approximately one extinction per species per million years on average[7]. However, these extinctions are not constant across time but instead come in "pulses", as shown below (image from Wikipedia):



During these pulses, extinctions per million years are roughly 2-10x the background rate.[8]

Late Quaternary Extinction

The Late Quaternary Extinction spanned a period from around 50,000 to 10,000 years ago. Over this time, around 34% of all mammals went extinct, including most mammals in the Americas and Australia and nearly all large mammals worldwide. This is an order of magnitude higher than the expected background extinction rate (which would be ~4% over 40,000 years).

The tables below (adapted from Wikipedia) document extinctions by geographic region and by size:

[7] Note that this varies by taxon and estimates within a taxon are approximate, with different estimates in the literature varying by a factor of 4 or sometimes greater.

[8] The 2-10x number is when looking at bins of 1 million years. For sudden catastrophic events such as an asteroid strike, the extinction rate over a 1-year time interval would spike much more than that.

Realm	Afrotropic	Indomalaya	Palearctic	Nearctic	Neotropic	Australasia
Areas Included	Southern Africa, Arabia	India, Southeast Asia, South China	Eurasia, Northern Africa	North America	South and Central America, Caribbean	Australia, New Zealand, neighboring islands
Initial Species	136	97	96	86	93	56
Decrease	7	8	22	57	51	38
Decrease (%)	5%	8%	23%	66%	54%	67%

Size	10-50kg	50-150kg	150-400kg	400-1000kg	1000kg+
Initial Species	215	113	86	46	33
Decrease	23	41	47	31	26
Decrease (%)	10%	36%	55%	67%	79%

As the tables show, extinctions were least severe in Africa (where humans originated, so mammals could co-evolve defenses), and were most severe in large mammals.

Causes. Historically, researchers debated whether these extinctions were driven by climate change or human contact. To understand this debate, I read several papers and chose to follow Koch & Barnosky (2006), which systematically reviews a number of competing theories. Koch & Barnosky conclude that the pattern and intensity of extinctions was driven by humans, but that climate change was an important additional contributor:

> “Taken as a whole, recent studies suggest that humans precipitated the extinction in many parts of the globe through combined direct (hunting) and perhaps indirect (competition, habitat alteration and fragmentation) impacts, but that late Quaternary environmental change influenced the timing, geography, and perhaps magnitude of extinction. Put another way, absent the various impacts of *Homo sapiens sapiens*, it is highly unlikely global ecosystems would have experienced a mass extinction of large, slow-breeding animals in the late Quaternary. But, absent concurrent rapid climatic change evident in many parts of the globe, some species may have persisted longer.”

Thus there are several paths through which humans could have driven extinctions:

- Direct hunting
- Indirect hunting (by dogs, rats, and other animals that we brought with us)
- Habitat destruction (e.g. by human-caused fire)

Importantly, different species likely went extinct for different reasons. Koch & Barnosky believe that most extinctions in Eurasia were due to climate change, that those in Australia and on most

islands were due almost entirely to humans, and North America was primarily humans with climate as an exacerbating factor.
Here is one story that illustrates the key points. It is consistent with Koch & Barnosky, but elides uncertainty in favor of simplicity.

- When humans arrived on islands, they brought along pigs, dogs, and rats, all of which preyed on indigenous species. Since island species were evolutionarily naive to these predators, many of them went extinct.
- Habitat destruction due to fire and land clearing also contributed to island extinctions.
- On larger land masses, mammals were not evolutionarily naive to carnivorous predators and so did not go extinct so easily. However, humans were very efficient hunters, enough to drive birth rates below death rates in many species, which eventually led to extinction over several millennia.
- Importantly, humans have a diverse diet, so even as they hunted some mammals to extinction, they gathered enough food from other animals and plants to sustain a large population size, thus avoiding traditional predator-prey cycles.
- In Africa and Eurasia, mammals co-evolved with humans or their predecessors over hundreds of thousands of years or more. They therefore had ample evolutionary time to develop defenses to efficient human hunters, explaining the lower rate of extinctions compared to America and Australia.

Overall, hunting by humans was probably the main driver of non-island extinctions, with other factors like climate change contributing. Importantly, it was not enough that humans were a novel predator, as novel predators do not always lead to extinction. It was also important that we were a particularly efficient predator that could occupy many geographic regions and had a diverse diet.

Holocene Extinction

The Holocene Extinction started around 10,000 years ago and has potentially accelerated recently, with most researchers believing that humans play an important role.

Paradoxically, despite occurring more recently, the extent of the Holocene Extinction is more disputed than the Late Quaternary Extinction, for two reasons. First, most extinction counts rely on the fossil record, but Holocene extinctions are based on present and historical observations by humans; this makes direct comparison hard, as the two methods have different (and large) sampling biases. Secondly, the extent of Holocene extinctions is politicized because it is central to present-day arguments about natural preservation, so it is harder to find neutral sources.

After looking through several papers, I decided to follow Barnosky et al. (2011)[9], which carefully discusses several sources of sampling bias and attempts to correct for them. Barnosky et al. conclude that a few percent of total species have gone extinct in the past 500 years, which is an

[9] This is the same Barnosky as above, though I did not know that when searching for papers—in both cases he happened to write the papers that I found most neutral and persuasive. To my delight, I learned that he is also at UC Berkeley. Go bears!

order of magnitude higher than the expected background rate of extinctions (though lower than other estimates in the literature[10]). Barnosky et al. also concludes that if most endangered species go extinct in the next century and this rate continues, we would lose the majority of all species within several centuries, on par with only 5 historical (and generally slower) mass extinction events.

Causes. Barnosky et al. list several stressors that contribute to these extinctions: “rapidly changing atmospheric conditions and warming[...], habitat fragmentation, pollution, overfishing and overhunting, invasive species and pathogens[...], and expanding human biomass”. Koch and Barnosky (2006) add ecological disruptions from the Quaternary Extinctions as a further stressor.[11]

Unlike past extinctions, we can directly observe the causes of many of the Holocene extinctions as they occur in the present day. Based on Hoffmann et al. (2010), habitat destruction is the largest driver of current extinctions, followed by invasive species (including disease) and overhunting, followed by environmental causes such as climate change and pollution.[12][13]

Summary: What Are Typical Causes of Extinction?

Overall, my analysis of past extinctions point to several ways that a species can go extinct

- A large-scale disaster or climate event, which either directly makes a species unviable or else disrupts ecosystems and leads to later extinctions.
- The introduction of a novel, aggressive organism for which the original species has not been adapted. This includes:
  - An invasive species, which can directly outcompete a species for its niche or disrupt the surrounding ecosystem.
  - A novel pathogen, especially if it has a reservoir species[14].
  - A new, efficient predator. This most affects island species, as continental species have been evolutionarily exposed to a diverse enough set of predators to develop robust counter-strategies. However, very efficient predators with diverse diets can overwhelm these evolved defenses even for non-island species.
- Changes in habitat, often caused by other species.

[10] See e.g. Ceballos et al. (2015), with an estimate closer to two orders of magnitude above the background rate.

[11] “Examples include plants that have lost their primary agents of seed dispersal or that are replete with defenses for herbivores that no longer exist, herbivores that are “overdesigned” for all existing predators, and scavengers such as condors that have no naturally occurring carcasses to eat in continental settings.”

[12] I follow Figure S7 of Hoffmann et al. (reproduced in the Appendix), which counts endangered species grouped by cause of endangerment. I grouped the rows into the categories “habitat destruction”, “invasive species” (which includes the chytrid fungus disease in amphibians), “overhunting/overfishing”, and “environment” (climate change / pollution / natural disasters). Some categories were ambiguous or did not fit into these 4. Overall I counted approximately ~360 in habitat destruction, ~250 from invasive species (dominated by amphibians), ~130 from overhunting/overfishing, and ~40 from environment.

[13] See also Dirzo & Raven (2003) who similarly claim that habitat destruction is the primary driver.

[14] A reservoir species is a second species in which the pathogen is not deadly, allowing it to multiply more freely, and from which the pathogen can cross over to the target species.

- Follow-on effects from other species going extinct. This partially overlaps with items above: for instance, the extinction of megaherbivores led to the regrowth of forests, thus significantly changing the habitat of other species.

Thus, in general most species extinctions are caused by:

- A second species which the original species has not had a chance to adapt to. This second species must also not be reliant on the original species to propagate itself.
- A catastrophic natural disaster or climate event.
- Habitat destruction or ecosystem disruption caused by one of the two sources above.

Why extinctions are usually rare. Since extinctions typically have a low base rate, causes of extinction must be rare. To better understand what *can* cause extinction, let's understand why most threats to a species do *not* lead to extinctions.

First, most predators do not cause extinctions. This is because prey evolve defenses in tandem with predators' offense, and the better a predator is the more evolutionary pressure on the prey (and so the faster defenses evolve). In addition to this, if prey become too rare then predator populations usually collapse, allowing the prey population to re-grow. Therefore, predators usually only cause extinctions if both (1) they enter a new environment with non-evolutionarily-adapted prey, and (2) they feed on multiple species, such that they can drive one species to extinction without their own population collapsing.

Similarly, novel pathogens do not by default lead to extinction of their hosts, since if they kill too many of the host species they don't have targets to propagate to. Instead, "pathogens are more likely to cause host extinctions if they…[have] long-lived infectious stages, or are multi-host pathogens that can be transmitted between common reservoir hosts and more vulnerable target species" (Kilpatrick and Altizer, 2010).

Humans. Finally, let's analyze why humans in particular were such efficient hunters that we *were* able to drive so many species to extinction. First, we are highly adaptable, thus being able to not just survive but live off multiple food sources in a variety of environments. This let us propagate globally and drive some species to extinction while still having alternate food sources. Second, we can coordinate effectively (Marean, 2015), overwhelming larger prey through better tactics. Finally, we used tools and technology to both increase our hunting ability and shape our environment, magnifying two of the key drivers of extinction discussed above.

Takeaways for Modern Catastrophes and for AI

Taking together all of the drivers for both human catastrophes and non-human extinctions, we see a small number of themes:

- Very large-scale natural events
- Highly adapted, self-replicating organisms, especially ones that the victim is not co-adapted to (epidemics, novel pathogens and predators, invasive species).
- Coordinated groups of humans (wars, hunting, habitat destruction)

- Political repression or disruption (forced labor, bad policies leading to famines)
- Follow-on effects from other catastrophes

Interestingly, technology does not seem to be a direct culprit in most human catastrophes, though it could be in the event of a large-scale nuclear war. For non-human extinctions, it is likely a contributor, since technology improves hunting ability and the ease of habitat destruction.

Looking at modern threats, nanotechnology and biotechnology both threaten to create novel self-replicators, and the inclusion of human design could lead them to be "adapted" in ways that are out-of-distribution relative to our evolutionary defenses.

Nuclear weapons increase the worst-case outcome of wars, and mass surveillance increases the worst-case outcome of political repression.

Climate change is a large-scale natural event. Aside from the direct effects, if it leads to many extinctions of non-human species, or induces political unrest, the follow-on effects could potentially be catastrophic for humans. The loss in biodiversity due to ongoing extinctions could also create bad follow-on effects, though it is happening slowly enough that it is probably not an immediate threat.

Finally, where do we put AI in this equation? Unfortunately, it looks to have many of the properties that underlie other drivers of catastrophes:

- AI is self-replicating in the sense that it can copy itself, and can train itself to adapt to new data quickly. It is therefore an adapted self-replicator that humans are not themselves adapted to.
- AI can likely be trained to coordinate better than humans, as humans evolutionarily were only adapted to coordinate in groups of ~150, whereas AI could be trained to coordinate in arbitrarily large groups if we solve the associated multi-agent RL challenges.
- Economic displacement from AI could lead to political unrest.
- AI is also a contributor to many of the other drivers above (though this is arguably double-counting): it makes mass surveillance easier and might speed the creation of other dangerous technologies such as engineered pathogens.

Overall, then, I expect AI to increase the rate of catastrophes. As calculated above, the base rate of very large (10% death rate) catastrophes over the next 25 years is 5%, and I personally expect AI to add an additional 10% on top of that, as I'll justify in the next post.

Open questions. There are several questions not resolved by this post. First, my analysis was inconclusive on whether or how much the rate of catastrophes changes over time. Data from extinctions suggests that it can vary by an order of magnitude, but it would be better to have data about human events.

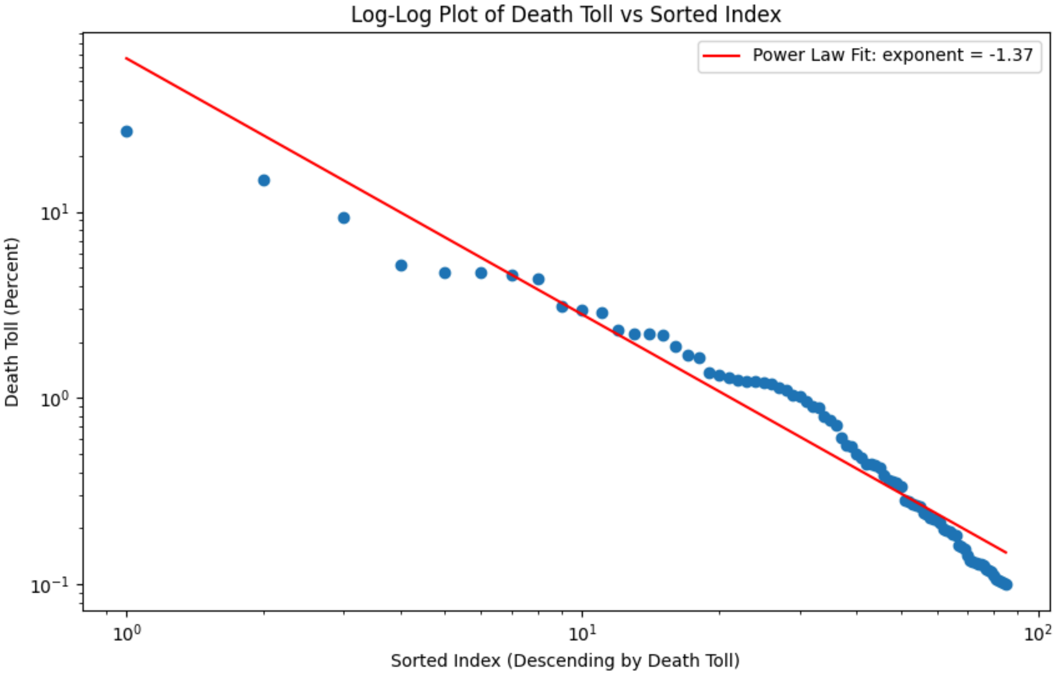
Second, this post says little about the importance of technology and intelligence, even though these are intuitively important. Are technological catastrophes increasing over time, even if right now they are too small to register in the data above? Do more intelligent species often drive less intelligent species to extinction?[15] Base rates on either of these would inform forecasts for AI.

Finally, one might argue that elapsed time is not the right x-axis, but instead elapsed population growth, economic growth, or technological progress. As one example, take world GDP. There have been as many doublings of world GDP since 1900 as there have been between 1900 and 0CE, so if GDP doublings are the right "clock" to measure against then we might expect many more catastrophes to happen each decade now than in the past. This doesn't seem true to me from the data so far, but I'd like to see it analyzed in more detail.

Appendix

Log-Log Plot of Catastrophes

As noted in Section 1, the distribution of catastrophes no longer follows Zipf's law when we go below death tolls of 1%, as shown below:

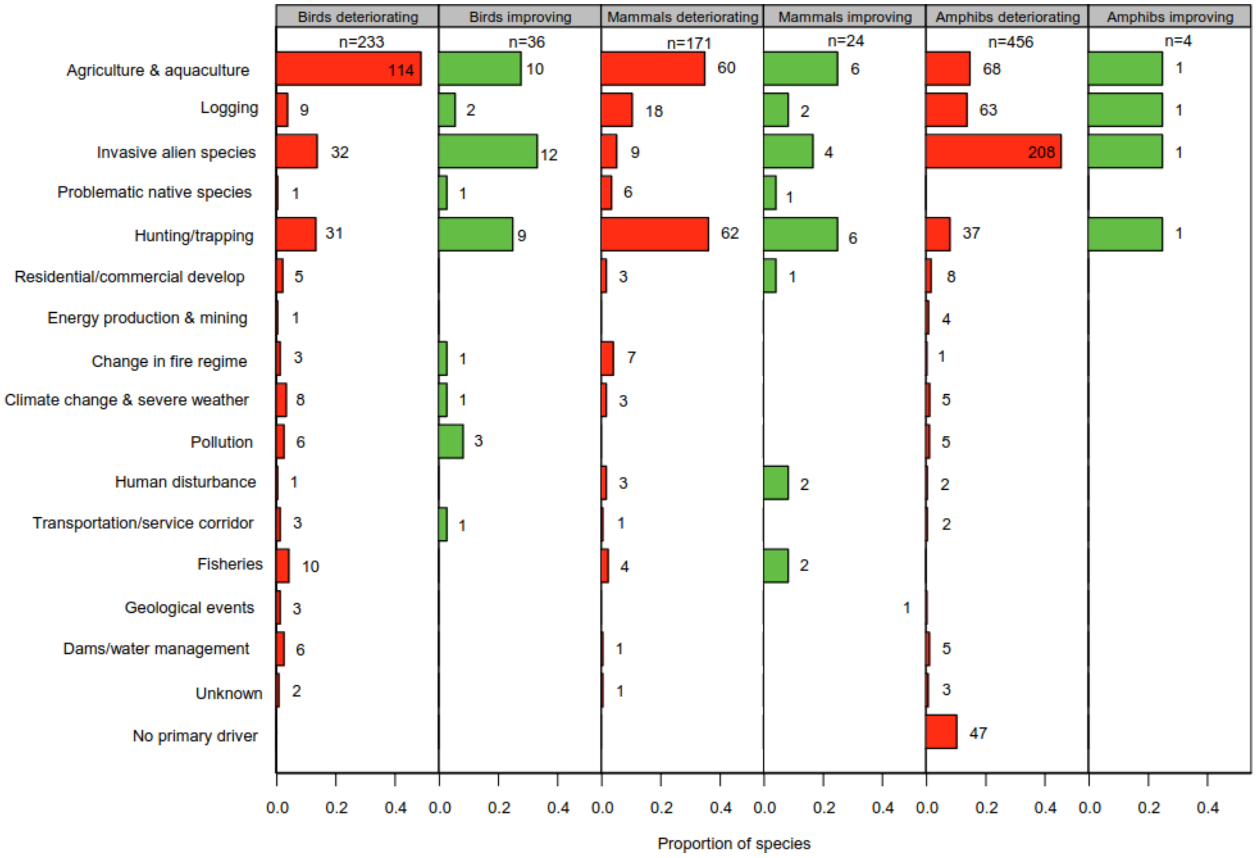


One possibility is that the actual trend is log-normal instead of power law. Another is that less severe events are underreported.

[15] The closest I found was Dembitzer et al. (2022), who claim that more intelligent mammals were less likely to go extinct during the Late Quaternary Extinction. However, we ideally want to study the opposite: are more intelligent mammals more likely to cause *other* species to go extinct?

Species Endangerment by Cause

The following is a reproduction of Figure S7 from Hoffmann et al. (2010).



OLD STUFF

“For large, slow-breeding animals, the most dangerous predators are omnivores who are sustained by small game and gathered food”
https://www.annualreviews.org/doi/pdf/10.1146/annurev.ecolsys.34.011802.132415

Event	Death Toll (Total)	Start	End	Start Population	Death Toll (Percent)	Type
Black Death	122M	1346	1353	450M	27.2%	Epidemic
Plague of Justinian	39M	541	549	261M	14.9%	Epidemic
Mongol invasions and conquests	41M	1206	1405	445M	9.3%	War
European colonization of the Americas	26M	1492	1691	498M	5.2%	War
Transition from Ming to Qing	25M	1618	1683	530M	4.7%	War
Taiping Rebellion + drought + famine	60M	1850	1873	1.28B	4.7%	Famine
An Lushan Rebellion	13M	755	763	282M	4.6%	War
Four famines in China	45M	1810	1849	1.02B	4.4%	Famine
World War II	91M	1939	1945	2.27B	4.0%	War
Antonine Plague	7M	165	180	239M	3.0%	Epidemic
Taiping Rebellion	28M	1850	1864	1.28B	2.2%	War
Spanish flu	41M	1918	1920	1.86B	2.2%	Epidemic
Yellow Turban Rebellion	5M	184	205	240M	1.9%	War
Cocoliztli epidemic of 1545–1548	9M	1545	1548	509M	1.7%	Epidemic
Great Famine of 1315–1317	8M	1315	1317	454M	1.7%	Famine
Deccan famine of 1630–1632	7M	1630	1632	538M	1.4%	Famine

Chinese famine of 1333–1337	6M	1333	1337	452M	1.3%	Famine
Famines in east-central China	22M	1907	1911	1.72B	1.3%	Famine
1520 Mexico smallpox epidemic	6M	1519	1520	506M	1.3%	Epidemic
Great Bengal famine of 1770	10M	1769	1773	815M	1.2%	Famine
Famine in India, China, Brazil, Northern Africa	17M	1876	1879	1.38B	1.2%	Famine
Chalisa famine	11M	1783	1784	903M	1.2%	Famine
Doji bara famine or Skull famine	11M	1789	1793	927M	1.2%	Famine
Thirty Years' War	6M	1618	1648	530M	1.1%	War
Third plague pandemic	13M	1855	1960	1.29B	1.0%	Epidemic
Qin's wars of unification	2M	-230	-221	196M	1.0%	War
World War I	18M	1914	1918	1.82B	1.0%	War
The Great Chinese Famine	29M	1959	1961	2.97B	1.0%	Famine
HIV/AIDS epidemic	42M	1981	2023	4.52B	0.9%	Epidemic
Second Sino-Japanese War	20M	1937	1945	2.22B	0.9%	War
Punic Wars	2M	-264	-146	191M	0.8%	War
Dungan Revolt	10M	1862	1877	1.31B	0.8%	War
735–737 Japanese smallpox epidemic	2M	735	737	281M	0.7%	Epidemic
Hundred Years' War	3M	1337	1453	451M	0.6%	War
Mughal–Maratha Wars	3M	1680	1707	577M	0.6%	War
French Wars of Religion	3M	1562	1598	511M	0.6%	War

Event	Death Toll (Total)	Start	End	Start Population	Death Toll (Percent)

Mongol invasions and conquests	41M	1206	1405	445M	9.3%
European colonization of the Americas	26M	1492	1691	498M	5.2%
Transition from Ming to Qing	25M	1618	1683	530M	4.7%
An Lushan Rebellion	13M	755	763	282M	4.6%
World War II	91M	1939	1945	2.27B	4.0%
Taiping Rebellion	28M	1850	1864	1.28B	2.2%
Yellow Turban Rebellion	5M	184	205	240M	1.9%
Thirty Years' War	6M	1618	1648	530M	1.1%
Qin's wars of unification	2M	-230	-221	196M	1.0%
World War I	18M	1914	1918	1.82B	1.0%
Second Sino-Japanese War	20M	1937	1945	2.22B	0.9%
Punic Wars	2M	-264	-146	191M	0.8%
Dungan Revolt	10M	1862	1877	1.31B	0.8%
Hundred Years' War	3M	1337	1453	451M	0.6%
Mughal–Maratha Wars	3M	1680	1707	577M	0.6%
French Wars of Religion	3M	1562	1598	511M	0.6%
Napoleonic Wars	5M	1803	1815	995M	0.5%
Chinese Civil War	10M	1927	1949	2.01B	0.5%
Crusades	2M	1095	1291	394M	0.4%
Russian Civil War	7M	1917	1921	1.85B	0.4%
Jewish–Roman wars	837K	66	136	235M	0.4%
Gallic Wars	632K	-58	-50	222M	0.3%
Japanese invasions of Korea	1M	1592	1598	516M	0.2%

Conquests of Mehmed II	873K	1451	1481	472M	0.2%
Mfecane	2M	1816	1828	1.06B	0.2%
Wars of the Three Kingdoms	876K	1639	1651	545M	0.2%
Seven Years' War	1M	1756	1763	766M	0.1%
Eighty Years' War	678K	1568	1648	512M	0.1%
Kalinga War	235K	-321	-261	183M	0.1%
War of the Spanish Succession	707K	1702	1714	597M	0.1%
French Revolutionary Wars	1M	1792	1802	942M	0.1%
Korean War	3M	1950	1953	2.5B	0.1%
Albigensian Crusade	447K	1208	1229	446M	0.1%

Event	Time Span	Death Toll (Total)	Death Toll (5-year span)	Total Population	Source	Cause
World War II	1939-1945	91M (4%)	>= 3.3%	2.26B	[1]	War
Mongol invasion and conquests	1206-1405	41M (9.2%)	>= 0.3%[16]	445M	[1]	War, Disease
Taiping rebellion	1850-1864	28M (2.2%)	>= 0.8%	1.28B	[1]	War
European colonization of the Americas	1492-1691	26M (5.2%)	>= 0.1%	498M	[1]	Disease
Manchu conquest	1618-1683	25M				

- Crusades (3-5%?)

[16] Based on the idea that most of the conquests seemed to happen in a 133-year span, so we multiply 9.2% by 5/133.

- Mongol invasion (200 years?)
- Three Kingdoms War (10% over 60 years?)
- Anlushan Rebellion
- Toba catastrophe (97% reduction in human population, around 75,000 years ago)
    - Population went from around 200,000 to <10,000 and stayed that way for around 20,000 years.
- Anlushan Rebellion (8 years)
- Black Death (30% of world population)
- Great Bengal Famine of 1770 (~1.5% of world population)
- I think Mongol invasions if we restrict to the single worst 5-year period goes above 1% but probably not above 10% (I'd guess 1.5%-2%)
- Three Kingdoms and Taiping rebellion are similarly going to be above 1% but below 10% if we pick the worst 5-year period. I think there's actually a bunch of these.

So seems like we have one at 30%, one at 4%, and a lot in the 1%-2% range (I'd guess maybe 5 if we collate all the examples?).

That suggests a power law with exponent ~2 (1/4 as many things in the 2%-4% range as the 1%-2% range).

    - Give list / table
    - Talk about overall rate per year
    - Group into categories of things
    - Have more happened recently? (for deciding what actual base rate should be)
- Historical extinctions of species
    - Say sources I used to assemble this
    - Provide general caveats
        - What counts as an extinction
        - General differing theories
        - Difficulty of measuring
    - Main causes historically
    - Main causes recently
    - Can we say anything more specific / quantitative about recent stuff?
        - Give all the caveats here
- Overall takeaways
    - What stuff can cause large population drops / extinctions
    - [stuff that seems maybe important, possibly out of place here]
        - What factors potentially mediate it? [e.g. co-adaptation]
        - What about new technologies?
    - What this says for current threats
        - Climate change
        - Biorisk
        - Nuclear

        - AI
        - [anything else…?]

For **base rates**, the base rate for “human extinction” should be pretty low, given that we have been around for a while and haven’t gone extinct yet. But it’s important to choose the right reference class. If we use the reference class of “new technologies that seem like a big deal”, the base rate is still low, but there’s other natural reference classes that make it seem higher. For instance, I think AI likely will end up in the class of “changes that are as big a deal as the industrial revolution”, and there have been only a handful of those throughout history.

More importantly, AI belongs simultaneously to the reference classes of “rapidly-adapting self-organizing systems” and “things that are smarter than humans”, and each of those contain several examples (e.g. viruses, governments) that are among the leading historical causes of mass deaths. There are few (perhaps no?) examples of things that are both of these at once. From this perspective, it’s reasonable to be uncertain a priori whether AI could cause extinction (rather than being highly confident that it won’t). Overall, combining these different reference classes, I’d assign a base rate of 10% to “things that are like AI” causing human extinction. (See Appendix B for more on this.)

, and Appendix C computes a “base rate” for extinction events (relevant to step 3)

Nothing has caused extinction so far — should have pretty low base rate
Argument is strongest for natural disasters
But still has force for human-induced threats—lots of new technologies, many people have worried about them, none have killed us yet
However, two key reference classes:

- Things that are smarter than humans
- Rapidly-evolving self-organizing systems

Things smarter than humans:

- Corporations, governments, the economy, the Internet
    - Genocide averages 150k/year worldwide
    - Millions in worst case
    - Instagram global effects on mental health
    - Significant consequences that are hard to easily stop

Rapidly-evolving self-organizing systems

- Bacteria, viruses, cancer
    - 600k/year malaria
    - millions/year

Automobiles: 1M

If we add both together, what happens? 500k -> 10B?
30M seems like better base estimate (say 80th percentile), 600M for 90th percentile, 6B for 95th percentile, so maybe ~2% base rate of extinction

looks like crusades has been updated to 3-5%!

So Mongol invasion I think is all that remains

https://en.m.wikipedia.org/wiki/List_of_anthropogenic_disasters_by_death_toll

Looks like Black Death is 200M according to that, so that’s 2!

started looking in detail at Base Rates for Death. The issue is how do we count -

let's start here: https://en.wikipedia.org/wiki/List_of_anthropogenic_disasters_by_death_toll
and here: https://en.wikipedia.org/wiki/Estimates_of_historical_world_population

The Black plague definitely counts (100-200MM dead, when the population was 500MM). So thats 1 in the last 1000 years

The issue gets much hairier with War. Do we count the start of the war, or the entire period. Most old wars lasted forever. The three kingdoms was probably 10% of the population - but it took 60 years! Which makes it about 0.2% per year --

The Mongol wars are similarly large but they took 200 years!

Im not happy with this - but we could use the tech argument to say things wouldnt take that long in the modern world

Looks like that gives 1/1000-2/1000, so rule of 3 (3/1000) is probably the upper CI

https://i.redd.it/3xhifaevabm71.jpg

The Anlushan Rebellion was 8 years - so that gives 0-1000 AD 1 event and 1000-2000 1 event, so 1/1000 probably is the strawman

https://ourworldindata.org/grapher/population?time=800..1949&country=~OWID_WRL

If we do everything above 1%, we'd also want to add:

* Black Death (30% of world population)
* Great Bengal Famine of 1770 (~1.5% of world population)
* WW2 (4% of world population)
* I think Mongol invasions if we restrict to the single worst 5-year period goes above 1% but probably not above 10% (I'd guess 1.5%-2%)
* Three Kingdoms and Taiping rebellion are similarly going to be above 1% but below 10% if we pick the worst 5-year period. I think there's actually a bunch of these.

So seems like we have one at 30%, one at 4%, and a lot in the 1%-2% range (I'd guess maybe 5 if we collate all the examples?).

That suggests a power law with exponent ~2 (1/4 as many things in the 2%-4% range as the 1%-2% range).

I think another way to get at base rates for extinction would be to look at extinctions of all species over time. Here we can get much more solid base rates. This article seems like a good starting point: https://royalsociety.org/topics-policy/projects/biodiversity/decline-and-extinction/. I'll copy the key graph:
image.png

This suggests a randomly chosen species has a 0.1% chance per year of going extinct. Should humans be higher or lower than the base rate? Can think of arguments in both directions, so I don't think we should obviously privilege either direction.

This would give 2.7% chance of extinction by 2050 without any inside view. I think AI increases the base rate significantly (even forecasters who are skeptical of x-risk from AI often agree it's one of the *most likely* causes of x-risk--they just think overall risk is low). So I'd double this and say that the chance of *extinction from AI by 2050* is 5.4% using a base rates-focused argument. (My inside view is higher and still in the 10%-15% range.)

Here is a 97% reduction in human population, around 75,000 years ago:
https://en.wikipedia.org/wiki/Toba_catastrophe_theory

Population went from around 200,000 to <10,000 and stayed that way for around 20,000 years.

I guess I also don't view extinction of species as being a good model for human existential risk. When a species goes extinct, it usually is that it has basically evolved into a new form. (Or more accurately, a splinter group was separated off, diverged, and when they recombined took over the old niche.) So rarely is it actually catastrophic. So typically it is more like the extinction of the neanderthals (who are currently alive in our modern genome) or the american indians (who are now basically extinct as far as looking at fossils in NA would show in a million years).

Now a confidence bound that I have computed many times in life, is based on the "rule of three." It says that if you have NEVER seen an event out of a total of N trials, your point estimate is zero, but your confidence interval is [0, 3/N]. So if we estimate the risk of WW III, we have not see it during the 80 years that necular weapons have existed. So the chance of it occuring next year is [0, 3/80]. This unfortunate is kinda wide. But it is all statistics can promise. Further, if we apply it to AI, we'd have to say that in the 13 months since PALM came out, we haven't been destroyed by an AI, so the chance of distruction NEXT MONTH is [0, 3/13]. Likely quite a bit higher than you would give it in your most depressed days!

Hi all--

Sorry to be so slow to reply.

I don't doubt the main qualitative conclusion--humans create an environment that's hostile to many other mammals--but I think papers like this are basically clickbait. None of the numbers means anything. Even the premise that exactly 351 mammalian species went extinct since the Pleistocene is pretty strained. The fossil record has huge uncertainties; new species are being discovered regularly (this from 2013 https://www.smithsonianmag.com/science-nature/for-the-first-time-in-35-years-a-new-carnivorous-mammal-species-is-discovered-in-the-americas-48047/ and this from 2021: https://www.nature.com/articles/nindia.2021.65#:~:text=Scientists%20at%20the%20Zoological%20Survey,nov.); and this analysis is an inchoate jumble of frequentist and Bayesian terminology and methods with made-up models. It's not at all clear how to properly quantify human activity in a relevant way. The assumption of exponential lifetimes for species is pretty strained. I'd ignore all the "confidence intervals," rates, and probabilities.

For an even worse example, see https://www.science.org/doi/10.1126/science.aaa4984
Thanks both. Philip, regarding your qualms with the article, what I'm really interested in is understanding two things:

(1) How much the rate of extinctions have risen since pre-historic times.
(2) How much they're risen (or stayed the same) in the last, say, 50-100 years.

I'd be happy to put a confidence interval on these rather than a single number. It seems like even accounting for uncertainty in the fossil record this should be possible. For instance, regarding (1), the linked paper claims 1000x increase as their point estimate, while other work tries to be fairly conservative and gets a lower bound of around 8x. My personal subjective credible interval given what I've read so far is maybe [100x, 3000x], but I'm not confident yet that I understand all the possible sources of uncertainty.

Regarding gaps in the fossil record, this is a good point and perhaps it would be better to have a model that accounts for some number of not-yet-discovered species, which could probably be estimated with some assumptions.

Best,
Jacob

Hi Jacob--

I don't think the questions you ask have meaningful/useful/defensible quantitative answers.

The sampling process has enormous built-in bias.

The "chance" (loosely speaking) that there is a discovered fossil record of a given species depends on many things, including the animal's size, habitat, how quickly it was buried, etc. See, e.g., https://www.theguardian.com/science/lost-worlds/2012/aug/17/bias-fossil-record

The "chance" (loosely speaking) that a non-extinct species is discovered at all also depends on many things.

+ There could be isolated places where mammals had speciation pressure/opportunity like Darwin's finches, and went extinct. What's the chance we'd know about it?
+ There are still very remote and inaccessible places about which little is known. How many endemic mammalian species went extinct there, and when? How many remain?

Best wishes,
Philip

This is apropos and interesting:
https://www.washingtonpost.com/climate-environment/interactive/2023/anthropocene-geologic-time-crawford-lake/

Looks like interpretable, quantitative data. It's only one spot, but the data quality looks good.

—-

In my last post, I provided several short stories of how advanced AI systems could create catastrophic outcomes for humanity, by exploiting a combination of cyberattack, scientific, and persuasion capabilities along with the ability to create copies of themselves, act in coordination, and think quickly.