The Optimistic Meta-Induction

Petra Kosonen
Some argue that existential risk mitigation should be one of the most important global priorities.¹

Existential risks are risks that threaten the destruction of humanity’s long-term potential.

Such risks might be posed by, for example, synthetic pathogens, artificial general intelligence, asteroids or climate change.

Extinction risks are one type of existential risk.

---

¹See for example Bostrom (2003), Bostrom (2013) and Greaves and MacAskill (2021).
The justification for existential risk mitigation need not rely on a high subjective credence in these risks.\textsuperscript{2} 

As humanity’s future is potentially very long, even small reductions in the net probability of existential catastrophe may correspond to enormous increases in expected moral value.\textsuperscript{3} 

However, many do, in fact, give relatively high probabilities for these risks materializing.

\textsuperscript{2}In fact, Thorstad (forthcoming) argues a high subjective credence in these risks undermines the case for existential risk mitigation. 
\textsuperscript{3}Bostrom (2013).
Existential risk estimates

- For example, participants of the Global Catastrophic Risk Conference (Oxford, 2008) gave a 19% chance for human extinction prior to 2100.\(^4\)

- The Oxford philosopher Toby Ord estimates that the probability of an existential catastrophe by year 2120 is 1/6—Russian roulette.\(^5\)

- The British Astronomer Martin Rees is even more pessimistic, as he thinks there is only a 50% chance that our present civilization on Earth survives to the end of the present century.\(^6\)

- And the Doomsday Clock now stands at 90 seconds to midnight—closest to global catastrophe it has ever been.\(^7\)

---

\(^4\)Sandberg and Bostrom (2008).
\(^5\)Ord (2020, p. 167).
\(^6\)Rees (2003, p. 8).
\(^7\)See Mecklin (2023).
However, our estimates for the probabilities of existential risks might be undermined by the *Optimistic Meta-Induction:* \(^8\)

The Optimistic Meta-Induction

The history of humanity is full of doomsday predictions that turned out to be wrong, so we have no reason to believe that our current doomsday predictions are approximately right.

In this talk, I explore whether we should lower our subjective credences in current existential risk estimates in light of the historical track record of failed predictions.

Should we turn back the Doomsday Clock?

---

\(^8\)The Optimistic Meta-Induction is analogous to the *Pessimistic Meta-Induction* in philosophy of science: Given that past widely accepted scientific theories were found to be false, we have no reason to think our current scientific theories are true or approximately true.
Here are some examples of predicted doom (see appendix for more):

- **1000**: Some Christian theorists proposed the end would occur 1000 years after the birth of Jesus.
- **1588**: Regiomontanus, a mathematician and astronomer, predicted the end of the world during 1588.
- **1600**: Martin Luther, known for starting the Protestant Reformation, predicted the end of the world would occur no later than 1600.
- **1656**: In his Book of Prophecies (1501), Christopher Columbus predicted that the world would end in 1656.
- **1688**: John Napier, a mathematician, calculated the end of the world would be in 1688 based on calculations from the Book of Revelation.
- **1719**: Jacob Bernoulli, a mathematician, predicted a comet would destroy the Earth on 5 Apr 1719.
Failed doomsday predictions

▶ 1926: Spencer Perceval, a British MP, believed that the world was growing nearer to the Apocalypse due to what he viewed as the rampant immorality of the times in Europe.
▶ 1982: John Gribbin, an astrophysicist, predicted that combined gravitational forces of aligned planets would create a number of catastrophes, including a great earthquake.
▶ 2000: During and before 1999 there was widespread predictions of a computer bug that would crash many computers on midnight of January 1, 2000 and cause malfunctions leading to major catastrophes worldwide.
▶ 30 Oct–Nov 29 2003: Aum Shinrikyo, a Japanese cult which carried out the Tokyo subway sarin attack in 1995, predicted the world would be destroyed by a nuclear war between 30 October and 29 November 2003.
▶ 2013: Rasputin, a Russian mystic who died in 1916, prophesied a storm would take place on 23 Aug 2013 where fire would destroy most life on land and Jesus would come back to Earth.
▶ 2026: In a 1960 Science article, Heinz von Foerster predicted 13 Nov 2026 as a date for the end of humanity due to overpopulation.
Were past doomsday predictions wrong?

- One might object to the Optimistic Meta-Induction by arguing that past doomsday predictions were, in fact, not wrong.
- There is one obvious way in which past doomsday predictions were wrong: the predicted catastrophes never happened.
- As a matter of fact, things turned out such that humanity still exists, contrary to predictions.

**Evidence 1**
The predicted catastrophes never happened.
One might object that the doomsday predictions were probabilistic: rather than predicting certainty of doom, they gave it a high probability.

And sometimes one might give an event a high subjective probability and not be epistemically wrong (in expectation) even if the event does not happen—sometimes the unlikely happens.

So, the fact that human extinction has never happened does not mean that past predictions were epistemically unjustified.

But, one might argue that the predictions must have been wrong because it would be very unlikely for humanity to still exist had these predictions been right about the magnitude of the risk.

**Evidence 2**

It would be very unlikely for humanity to still exist had these predictions been right about the magnitude of the risk.
Observation selection effects

- But, given that we cannot observe doomsday predictions ever materializing, judgement about their wrongness is subject to observation selection effects.\(^9\)
- Had they been right in the past, there would be no one to observe it.
- It could be that extinction risk was high and most civilizations destroy themselves relatively quickly, but we are the lucky survivors.
- Observation selection effects explain why it is not surprising that humanity exists, even if extinction risk was high.
- So, we cannot use the continued existence of humanity as evidence against past doomsday predictions.

\(^9\)Ćirković et al. (2010).
Lack of near misses

- However, we can change the argument slightly: had they been right about the magnitude of the risk in the past, we would expect to at least find evidence of near misses—but we do not.
- Observation selection effects do not explain the lack of near misses.

Evidence 3

It would be very unlikely for humanity to still exist and not have experienced any near misses had these predictions been right about the magnitude of the risk.
Past predictions were unscientific

And lastly, we have other evidence for past doomsday predictions having been wrong: given what we now know, they could never have happened (or had a tiny probability of happening).

For example, many doomsday predictions were religious stories that conflict with the scientific worldview.

Evidence 4

Given what we now believe, the predicted catastrophes could never have happened (or had a tiny probability of happening).
Does the wrongness of past doomsday predictions mean current doomsday predictions are wrong?

▶ We have four kinds of evidence against past doomsday predictions: the predicted catastrophes never happened, it would be unlikely for humanity to still exist had they been right about the magnitude of the risk, the lack of near misses and, given what we now believe, the predicted catastrophes could never have happened (or had a tiny probability of happening).

▶ One might accept that past doomsday predictions were wrong, but insist that this does not mean our current doomsday predictions are wrong.

▶ Next, I will discuss objections to the Optimistic Meta-Induction of this sort.
The first objection says that past doomsday predictions are the wrong reference class to reason about our current doomsday predictions.

I will discuss three different versions of this objection:

1. Unlike in the past, nowadays doomsday predictions are scientifically informed.
2. Unlike in the past, nowadays we can ask superforecasters to estimate the risks.
3. In the past they predicted global catastrophes, not extinction.
A plausible objection to the Optimistic Meta-Induction is that people who made doomsday predictions in the past did not study the risks rigorously and instead relied on, for example, religious stories.\(^{10}\)

So, even if they were wrong in the past, that does not mean we are wrong now; nowadays we know better.

One can also combine this objection with the objection from observation selection effects: either the doomsday predictions in the past were unscientific (unlike today’s doomsday predictions), so they could never have happened, or they they were plausible risks, but observation selection effects explain the continued existence of humanity.

\(^{10}\)Moynihan (2020, pp. 20–21, 27) points this out.
However, the prevalence of past doomsday predictions might be taken to show that we have a psychological tendency to exaggerate extinction/existential risks.¹¹

Some (or most) people have a psychological tendency to make doomsday predictions, whether or not they are justified.

Given that people have always made doomsday predictions whether or not they were justified, we should expect us to make such predictions whether or not they are justified.

And, the same psychological tendencies that caused people to come up with apocalyptic stories in the past might cause today’s scientists and philosophers to exaggerate the risks we face.

In light of this, we should penalize our credences in near-term existential risks.

¹¹This tendency might be cultural too.
The second objection that I will discuss in this section is that, although there is a history of failed doomsday predictions, we should not conclude current doomsday predictions are wrong because nowadays we can ask talented forecasters who have a good track record to estimate the risks.

And, if these ‘superforecasters’ say there is a high probability of an existential catastrophe, we should believe them.

From a recent paper: “[I]f existential risks are anything like the shorter-run geopolitical forecasting questions that were the focus of past studies, the superforecasters could be a better guide to what will actually happen.”¹²

What do the superforecasters say about existential risks?

¹²Karger et al. (2023, p. 14).
The median superforecaster predicted a 9% chance of global catastrophe (that kills at least 10%) and a 1% chance of extinction by year 2100.\textsuperscript{13}

Here are some superforecasters’ estimates for extinction risks from various causes by year 2100:

1. AI extinction: 0.38%
2. Engineered pathogen extinction: 0.01%
3. Nuclear extinction: 0.074%
4. Total extinction risk: 1%

If anything, superforecasters seem to have relatively low credences for near-term extinction. So, superforecasters do not support a high credence in human extinction in this century.

\textsuperscript{13}Karger et al. (2023, p. 4). Karger et al. (2023, p. 4, n.4) define a catastrophic event as one causing the death of at least 10% of humans alive at the beginning of a five-year period and extinction as reduction of the global population to less than 5000.
The next objection to the Optimistic Meta-Induction says that past doomsday predictions are not the right reference class to reason about current doomsday predictions because they were about global catastrophes and apocalypses rather than extinction.

Moynihan (2020, p. 32): “Haven’t humans been predicting the end of time since the beginning of history? Certainly—but extinction has nothing to do with religious apocalypse.”

However, the same psychological tendencies might predispose people to believe in both global catastrophes and extinction because global catastrophes and extinction events share similar features.
Finally, I will discuss some reasons to think near-term existential risk is high.

The Optimistic Meta-Induction asks us to be careful before assigning high subjective credences to near-term existential risks.

But what if the evidence for high risks is convincing?

We might have good reasons to think existential risk is high. I will discuss a few:

1. There is evidence for particular risks.
2. In the past civilizations have gone extinct.
3. The Doomsday Argument.
The Optimistic Meta-Induction is higher-order evidence, that is, evidence about the character of first-order evidence.

It seems there is something arrogant about this.

Who are we to say that the carefully formulated existential risk estimates are wrong based on a meta-argument when the scientists and philosophers who made those estimates have studied the topics carefully?

For example, Toby Ord provides detailed descriptions of where the 1/6 existential risk estimate comes from.

The Optimistic Meta-Induction is insensitive to evidence about particular risks.
Some might object that although humans have not gone extinct, many civilizations have, e.g., on Easter island.

And now we have one global civilization, so we should be worried.

Paul Ehrlich, who co-authored the Population Bomb (another failed doomsday prediction), said in an interview: “Civilisations have collapsed before: the question is whether we can avoid the first time [an] entire global civilisation has given us the opportunity of having the whole mess collapse.”
How long have civilizations lasted in the past?

According to a paper on empires from 600 B.C. to 600 A.D., the longest lasting empires were the Parthian-Sassanid empire (lasted 7 centuries) followed by Rome, Andhra, the Byzantine and Ptolemaic empires (lasted around 3 to 4 centuries).\(^{14}\)

According to another paper on empires from 3000 to 600 B.C., the longest lasting empires were Egypt’s New and Old Empires (lasted 5 centuries each) and Hsia-Shang empire (lasted 4 centuries).\(^{15}\)

Even the longest lasting empires seem to be relatively short-lived.

The average lifespan of a civilization is 336 years.\(^{16}\)

However, some individuals often survived the collapse of their civilization, so we should not draw conclusions about extinction risk from these numbers.

---

\(^{14}\) Taagepera (1979, p. 133).
\(^{15}\) Taagepera (1978, p. 191).
\(^{16}\) Kemp (2019).
Lastly, another reason to think near-term existential risk is high is the Doomsday Argument.

Consider the *Self-Sampling Assumption*:

### The Self-Sampling Assumption

You should reason as if you were a random sample from the set of all observers (in your reference class).

Suppose we have two hypotheses:

1. **Doom Early**: humankind goes extinct sometime in this century and the total number of humans that will have existed is, say, 200 billion.
2. **Doom Late**: humankind survives the century and goes on to settle the galaxy; the total number of humans that will ever have lived is, say, 200 trillion.

---

17 The argument presented here is from Bostrom (2008).
▶ Now you find out that your birth rank is about 60 billion (that’s approximately how many humans have lived before you).
▶ This gives you reason to think Doom Soon is more probable than you previously thought: it would be surprising to be so early in the history of humanity if Doom Late was correct.
▶ So, you have reason to expect human extinction happening relatively soon.
Many assign relatively high probabilities for near-term existential risks.

However, our estimates for the probabilities of existential risks might be undermined by the Optimistic Meta-Induction: The history of humanity is full of doomsday predictions that turned out to be wrong, so we have no reason to believe that our current doomsday predictions are approximately right.

First, some might object that past doomsday predictions were not wrong, given observation selection effects.

However, given what we now believe, many past doomsday predictions could never have happened (or had a tiny probability of happening).
Secondly, some might argue that even though past doomsday predictions were wrong, that does not mean we should think current doomsday predictions are wrong. For example, past doomsday predictions might be the wrong reference class to reason about current doomsday predictions because they were not scientifically informed.

All in all, I think there is something to the Optimistic Meta-Induction.

At least we should be a bit careful before assigning a high subjective credence to near-term existential risks.
Here are some examples of predicted doom:

- 365: Hilary of Poitiers, an early French bishop, announced the end of the world would happen during this year.
- 375–400: Martin of Tours, a French bishop, stated that the world would end before 400 AD.
- 482: Hydatius, a bishop of Aquae Flaviae, wrote his chronicle (c. 469) in his firm belief that humanity was living in the end times, marching towards its certain doom on this day, when Jesus would come back and the world would end.
- 793: Beatus of Liébana, a Spanish monk, prophesied the Second Coming of Christ and the end of the world on that day in front of a large crowd of people.
- 799–806: Gregory of Tours, a French bishop, calculated the end would occur between 799 and 806.
- 992–995: Good Friday coincided with the Feast of the Annunciation; this had long been believed to be the event that would bring forth the Antichrist, and thus the end-times, within three years.

Failed doomsday predictions: 11th–15th centuries

- **1000:** Christian theorists proposed the end would occur 1000 years after Jesus’ birth.
- **1033:** Following the failure of the prediction for 1 January 1000, some theorists proposed that the end would occur 1000 years after Jesus’ death, instead of his birth.
- **1290 and 1335:** After his 1260 prediction failed, the followers of Joachim of Fiore rescheduled the end of the world to 1290 and then again to 1335.
- **1346–1351:** The Black Death spreading across Europe was interpreted by many as the sign of the end of times.
Failed doomsday predictions: 16th century

- 1524: A group of astrologers in London predicted the world would end by a flood starting in London.
- 1528: Hans Hut, a German Anabaptist (Christian movement), predicted the end would occur on 27 May 1528.
- 1528: A revised date from Johannes Stöffler after his 1524 prediction failed to come true.
- 1588: Regiomontanus, a mathematician and astronomer, predicted the end of the world during this year.
- 1600: Martin Luther, a German priest and professor of theology known for starting the Protestant Reformation, predicted the end of the world would occur no later than 1600.
Failed doomsday predictions: 17th century

- **1624**: The same astrologers who predicted the deluge of 1 February 1524 recalculated the date to 100 years later after their first prophecy failed.
- **1654**: Helisaeus Roeslin, a physician, made a prediction that the world would end this year based on a nova that occurred in 1572.
- **1656**: In his Book of Prophecies (1501), Christopher Columbus predicted that the world would end during this year.
- **1666**: Following his failed prediction of 1648, Sabbatai Zevi (a Jewish mystic and rabbi) recalculated the end of the Earth for this year.
- **1688**: John Napier, a mathematician, calculated the end of the world would be this year based on calculations from the Book of Revelation.
- **1697**: Cotton Mather, a Puritan minister, predicted the world would end this year. After the prediction failed, he revised the date of the end two more times.
1705–1708: Camisard prophets (French Protestants) predicted the end of the world would occur in either 1705, 1706 or 1708.

1719: Jacob Bernoulli, a mathematician, predicted a comet would destroy the Earth on 5 Apr 1719.

1700–1734: Nicholas of Cusa, a cardinal, predicted the end would occur between 1700 and 1734.

1736: William Whiston, a theologian, predicted a comet colliding with the Earth this year.

1792 and 1794: The Shakers, a Christian sect founded in 18th century England, predicted the world would end in 1792 and then in 1794.
Failed doomsday predictions: 19th century

- 1805: Christopher Love, a Presbyterian minister, predicted the destruction of the world by earthquake in 1805.
- 1843 and 1847: Harriet Livermore, a preacher, predicted the world would end.
- 1862: John Cumming, a Scottish clergyman, stated it was 6000 years since creation in 1862, and that the world would end.
Failed doomsday predictions: 20th century

▶ 1910: Camille Flammarion, a French astronomer and author, predicted that the 1910 appearance of Halley’s Comet would ‘possibly snuff out all life on the planet’, but not the planet itself. ‘Comet pills’ were sold to protect against toxic gases.

▶ 1926: Spencer Perceval, a British MP, believed that the world was growing nearer to the Apocalypse due to what he viewed as the rampant immorality of the times in Europe.

▶ 1934: Walter Marks, an Australian MP, told the House of Representatives that Armageddon would occur in 1934 and culminate with the Royal Navy bringing Christ’s chosen people to Jerusalem.

▶ 1935: Wilbur Glenn Voliva, an evangelist, announced that “the world is going to go ’puff’ and disappear” in September 1935.

▶ 1941: A prediction of the end from the Jehovah’s Witnesses.
Failed doomsday predictions: 20th century

- 1954: Dorothy Martin (leader of a UFO cult called Brotherhood of the Seven Rays) claimed the world was to be destroyed by terrible flooding on 21 Dec 1954.
- 1962: Jeane Dixon, an American psychic and astrologer, predicted a planetary alignment on 4 Feb 1962 was to bring destruction to the world. Mass prayer meetings were held in India.
- 1967: Jim Jones, the founder of the People’s Temple, stated he had visions that a nuclear holocaust was to take place in 1967.
- 1974: David Berg, the leader of Children of God, predicted that there would be a colossal doomsday event heralded by Comet Kohoutek.
- 1975: From 1966 on, Jehovah’s Witnesses published articles which stated that the fall of 1975 would be 6000 years since man’s creation, and suggested that Armageddon could be finished by then.
- 1976: The Brahma Kumaris founder, Lekhraj Kirpalani, has made a number of predictions of a global Armageddon. All of the rest of humanity would be killed by nuclear or civil wars and natural disasters which would include the sinking of all other continents except India.
Failed doomsday predictions: 20th century

- 1980: In 1978 Leland Jensen predicted that there would be a nuclear holocaust in 1980, followed by two decades of conflict, culminating in God’s Kingdom being established on Earth.

- 1982: John Gribbin, an astrophysicist, co-authored the 1974 book The Jupiter Effect which predicted that combined gravitational forces of aligned planets would create a number of catastrophes, including a great earthquake on the San Andreas Fault.

- 1982: In late 1976 on his 700 Club TV programme, Pat Robertson predicted that the end of the world would come in this year.

- Lester Sumrall, a pastor and evangelist, predicted the end in this year, even writing a book about it entitled I Predict 1985.

- 1986: Leland Jensen predicted that Halley’s Comet would be pulled into Earth’s orbit on 29 Apr 1986, causing widespread destruction.

- 1991: Louis Farrakhan, the leader of the Nation of Islam, declared that the Gulf War would be the ‘War of Armageddon which is the final war.’
Failed doomsday predictions: 20th century

► 1994: Neal Chase, a Bahá’í sect leader, predicted that New York City would be destroyed by a nuclear bomb on March 23, 1994, and the Battle of Armageddon would take place 40 days later.

► 1998: Hon-Ming Chen, leader of the Taiwanese cult Chen Tao —‘The True Way’—claimed that God would come to Earth in a flying saucer at 10:00 am on 31 Mar 1998.

► 1999: Charles Berlitz, a linguist, predicted the end would occur in this year. He did not predict how it would occur, stating that it might involve nuclear devastation, asteroid impact, pole shift or other Earth changes.

► 2000: During and before 1999 there was widespread predictions of a Y2K computer bug that would crash many computers on midnight of January 1, 2000 and cause malfunctions leading to major catastrophes worldwide, and that society would cease to function.
Failed doomsday predictions: 21st century

- **2001**: Tynnetta Muhammad, a columnist for the Nation of Islam, predicted the end would occur in this year.
- **30 Oct–Nov 29 2003**: Aum Shinrikyo, a Japanese cult which carried out the Tokyo subway sarin attack in 1995, predicted the world would be destroyed by a nuclear war between 30 October and 29 November 2003.
- **2006**: Yisrayl Hawkins, pastor and overseer of The House of Yahweh, predicted in his February 2006 newsletter that a nuclear war would begin on September 12, 2006.
- **2012**: The 2012 phenomenon predicted the world would end at the end of the 13th b’ak’tun (21 Dec 2012). The Earth would be destroyed by an asteroid, Nibiru, or some other interplanetary object; an alien invasion; or a supernova.
- **2013**: Rasputin, a Russian mystic who died in 1916, prophesied a storm would take place on 23 Aug 2013 where fire would destroy most life on land and Jesus would come back to Earth.
- **2026**: In a 1960 Science article, Heinz von Foerster predicted 13 Nov 2026 as a date for the end of humanity due to overpopulation.
References I


**URL:** https://globalprioritiesinstitute.org/hilary-greaves-william-macaskill-the-case-for-strong-longtermism-2/
URL: https://static1.squarespace.com/static/635693acf15a3e2a14a56a4a/t/64abffe3f024747dd0e38d71/1688993798938/XPT.pdf


Mecklin, J. (2023), ‘A time of unprecedented danger: It is 90 seconds to midnight’.
URL: https://thebulletin.org/doomsday-clock/current-time/


