Doomsday argument

DA types and sampling methods

Sampling method

Self sampling (SSA)

All else equal, an observer should reason as if they are randomly selected from the set of **all actually existent observers (past, present and future) in their reference class.**

1/2 in Sleeping beauty

Self indication (SIA)

All other things equal, an observer should reason as if they are randomly selected from the set of all possible observers. (wiki)

"The idea that **your very existence** gives you a reason to think that there **are many observers**" – Bostrom, link

1/3 in Sleeping beauty

Other sampling methods

list link

Gott's formula

$$P(N \leq 20n) = \frac{19}{20}$$

Unconditional approach

- It predicts future duration of human survival based only on current duration of existence of the civilization.
- Proven to work in many cases, such as length of theater shows.

Gott's article

Predictions:
- *95% human extinction at 5000 to 8 mln years, 50% for 200 000 years*
- *50% in next 1000 years if population growth is considered*
- *50% for 200 years if science is referent class*
- *50% for 20 years if DA-specialists are referent class*

Gott's formula requires actually existing observers in future, that is only possible if the many worlds interpretation of QM is not true and future is deterministic (see Leslie's book).

But even if the condition is not met, Gott's formula could be used to estimate median life expectancy.

Cosmology

Sampling order

Sampling order depends of which of my features are by definition in my referent class and which are random: the first ones I use for prior and second ones for Bayesian update.

Core observer and its random properties

Any intelligent observer has inner structure, which consists of his core based on his definition, and random properties, like date of birth or place of birth.

In result this observer could use probability distributions for his random properties.

Such properties behave like random in my case, and I could predict year duration in months based on my month of birth.

Carter's formula

$$\Pr(\mathcal{M}|e) = \frac{\Pr(e|\mathcal{M})\Pr(\mathcal{M})}{\Pr(e)}$$

Conditional approach

Increases Bayesian probability of the expected catastrophe or the fact that we are located before it in time.

It depends of apriori probability of the catastrophe N and total number of the observers if it will not happen.

The update could become arbitrary large if N is large.

But we could take N from Gott formula.

Bostrom's article with formula

Prediction: Rises 1 probability of near-term extinction to 99.9% if N is 1000 times larger (Such N is minimum for galactic civilization)

The SSI increases chance that the referent class is big and it neutralizes DA.

If all possible observers exist, SSI becomes uninformative, and doesn't cancel DA.

A very big actual Universe implies that all possible observers exist, and SSI does not provide any new information in it, so we have to return to SSA.

Bostrom article about SIA

But SIA has its own doomsday: "Everything else being equal, SIA implies that the population growth rate in your past is likely to be higher than the rate in the future; i.e. it predicts an observed decline, not in population, but in population growth rates". Armstrong

ASSA

Absolute self-selection assumption (link)

Observer is randomly selected from all possible observer moments, measured by Solomonov complexity of implied world (less complex are more probable).

SSA+SIA

"All else equal, an observer should reason as if they are randomly selected from the set of all possible observers." Bostrom according to wiki.

Full non-indexical conditioning

FNC says the probability of being in a possible world is proportional to the chance of anyone in that world having exactly your experiences, link.

Another classification (link)

SSA-A : **Actual** observers: "SSA should be applied only to observers in the reference class that actually exist".

SSA-MW : SSA-A but **Many Worlds** observers are treated as actual.

SIA-C : Self Indicating Assumption (Chance version): "SSA should be applied to all observers in the reference class that are **possible**, but do not necessarily actually exist". You are more likely to exist in universe with lots of observers, so your existence favors a realization of the universe with lots of observers.

SIA-U : Self Indicating Assumption (Uncertainty version) "SSA should be applied to all observers **that we think might exist** or could have existed, but do not necessarily actually exist." Subjective uncertainties about which theories are correct are treated in the same way as probabilistic outcomes of a correct theory.

Anthropic shadow

$$\eta = \frac{P(\text{a priori})}{P(\text{a posteriori})},$$

Underestimation of the rate of natural catastrophes because of observation selection

- Anthropic principle worked in the past creating conditions for observers' existence, but will not work in the future.
- Future may be less stable and less anthropic friendly.
- A catastrophe of unknown kind may be long overdue.
- Our world may be more fragile than we know (e.g. Venusian-style runaway global warming is more probable).
- Small anthropogenic action could start unexpectedly large catastrophe (methane hydrates release, supervolcano eruption).

Bostrom-Circovic article

Prediction: Rises probability of natural catastrophes maybe up to 10 times, significantly lowers threshold of stability of the Earth's atmosphere.

Cosmological implications

- We live in a universe fine-tuned for sentient beings (anthropic principle).
- Among all possible fine-tuned universes, we ought to assume we live in the most abundant class of them (SSA).
- Most of them may not be exactly fine tuned and have a lot of catastrophic events.
- The share of large universes with "rare earths" dominates on the share of completely fine tuned universes where life is on each planet (because not be exactly fine tuned universes are more widespread).
- This means the risk of catastrophes is higher in them.

Simulation argument

$$f_{\text{sim}} = \frac{f_p N H}{(f_p N H) + H}$$

Simulation argument for Earth

1) OUR civilization is very likely to go extinct before reaching a "posthuman" stage.

2) or: WE will not create simulations.

3) or: WE are almost certainly living in a computer simulation.

- Smaller, cheaper and "earlier" one-person-simulations are more probable
- Simulation may be switched off after reaching certain threshold

link on Bostrom's site
link on my Simulation map

Simulation argument for all possible civilizations

1) ALL possible civilizations are very likely to go extinct before reaching a "posthuman" stage (which is unlikely).

2) WE are almost certainly living in a computer simulation.

- A lot of the simulations will presumably be created for the purpose of estimating the probability of existential risks (in order to solve Fermi paradox and probability of meeting other AI in space), and may be dominating class of the simulations.
- If so we are likely to find ourselves near x-risk events and in position to influence them (true).
- If I prove that I am in a simulation it could be switched off.
- Or we go extinct in the simulation or it will be switched off after we escape risks.

Prediction: we are in the simulation.

Doomsday simulation argument

If both DA and SA are true then we live in a simulation and it will be switched off or have existential catastrophe soon.

If we are in a simulation, we don't know when it began. It could have been as late as e.g., the year 2000. If the total simulation is short, and we assume that we are roughly in the middle of it, it could end very soon.

2) Aranyosi: DA and SA cancel each other out (it may be it a result from his assumptions)

Fermi paradox

$$N = R^* \cdot f_p \cdot n_e \cdot f_l \cdot f_i \cdot f_c \cdot L$$

Where are they?

The absence of visible aliens or alien signals may suggest that there is a "Great Filter" which wipes out life before it can spread across space.

We may have already passed this (Rare Earth theories) or it may be in the future (extinction)

Even if the Great Filter is in the past, this fact increases our estimation of the probabilities of natural catastrophes, which were hidden until now from us by observation selection.
wiki

Great filter ahead

SSI tells us that the Great Filter is more likely to be ahead of us (Katja Grace theory) and it is not an AI which will occupy all cosmic space

Future of the Universe

- We could estimate future time of existence of the habitable universe based on our time location (it is billions, not trillions of years).
- We can estimate the approximate upper bound of the probability of cosmic extinction level events by the fact that complex life has existed for upwards of 600 million years (Bostrom estimates the upper possible bound of frequency of these events as 1 every billion years).
- We can estimate "rare Earth" vs. "Earth is first" hypothesis. Rare Earth (or inevitable extinction) is more likely as we rather late in the universe.

Anthropic decision theory

http://lesswrong.com/r/discussion/lw/89q/anthropic_decision_theory_iii_solving_selfless/
Anthropic Decision Theory (ADT): An agent should first find all the decisions linked with their own. Then they should maximise expected utility, acting as if they simultaneously controlled the outcomes of all linked decisions, and using the objective (non-anthropic) probabilities of the various worlds.

Under SIA, the argument goes away, as it would seem that ADT must behave oddly: depending on the selfishness and selflessness of the agents, they would give different probabilities to the extinction of the human race. This is a very counterintuitive result that under ADT, decisions matter more in probable worlds. And agents that are either selfish or selfless enough would act as if they were concerned with the extinction of the human race, so would not act in bets to prevent this.

Copernican mediocrity principle

"If an item is drawn at random from one of several sets or categories, it's likelier to come from the most numerous category than from any one of the less numerous categories" (wiki)

Median age of earth-like civilizations

- We could estimate median life expectancy in given class of objects based on the age of one random sample. (If my age is 40 years, it means that median age of humans is not one millisecond and is not one million years, but tens of years).
- It gives median age of any human-like civilization and ours should be the same (it means thousands of years and early extinction).
- It means that Great Filter from Fermi paradox is ahead.
- It means in particular that most civilizations are pre-Singularity (but this may be explained by abundance of simulations).
- It means that my IQ is median and no many different superAIs exist.

Prediction: We are in middle of tech. civilisation existence and will go extinct in 100 years

Other DA-style ideas

Thermodynamic DA

We live in a thermodynamic fluctuation and as smaller and simpler fluctuations are more probable, there should be a force against complexity, AI development or our existence in general (Strugatsky, "Definitely Maybe").

Quantum Doomsday argument

If number of worlds is growing exponentially, I should find myself near the end of my personal existence in almost all worlds, which should consist of global catastrophe of some kind in the next moment.

This requires many world interpretation of QM (but real Everett's interpretation means globally continuous wave function without separate worlds).

If quantum immortality is true or if the number of worlds is not growing, QDA does not work.

link on Wilson article

Similar to DA problems and ideas

- Sunrise problem by Laplace.
- Induction in general.

DA and Copernican principle in other prediction domains

- Personal life expectancy.
- Market predictions.
- Bus waiting and other problems.
- Future age of the universe.
- Median intelligence and other typical features of intelligent beings and civilizations.

wiki

DA as measure of stability

DA makes near term catastrophes less probable (not tomorrow).

Solutions

Different solutions of the reference class problem

Time

- Duration of Homo sapience (200 000 years).
- Duration of human civilization's existence 5000 years).
- Duration of probability theory and science (200 years).

Observers

- Animals (trillions).
- Humans (100 billion), birth rank.
- Observer-moments.
- Any sentient beings, including robots, ems and aliens.

Self-referencing

- Beings capable of understand DA (millions).
- Beings already knowing DA (tens of thousands).
- Beings already knowing this self-referencing DA idea (hundreds?).

Reference class and its end

- End of the reference class is not the end of the world – each reference class has its own end.
- Reference class is defined by how it ends (like humans become non-humans).
- DA applies to all reference classes, from which I an randomly chosen, but provides different ends in different classes.
- "End" does not mean extinction or death.
- "End" may mean solving DA, or losing interest in it, or become non-human being.

Indifference principle

Bland indifference principle

"Bland indifference principle prescribes indifference only between hypotheses about which observer you are, **when you have no information** about which of these observers you are".

A) **Very weak – only exact copies may be counted as equal** observers. Best for simulation argument, if there is no real difference between sims and real people.

B) Stronger: "The second case is where the minds are "like" each other only in the **loose sense of being the sort of minds that are typical of human creatures**, but they are qualitatively distinct from one another and each has a distinct set of experiences". Bostrom, link.

Strong indifference principle

"The Doomsday argument rests on a much stronger and more controversial premise, namely that one should reason as if one were a random sample from the set of all people who will ever have lived (past, present, and future) **even though we know that we are living in the early twenty-first century rather** than at some point in the distant past or the future" – Bostrom.

DA rebuttal arguments

Non-randomness

I am not a random observer.

Sampling method

SIA – the fact that I exist proves that many more observers exist and exactly cancels out the DA.

wiki

Self rebuttal

DA is randomly chosen from all math theories and if it is true it will be refuted soon. This is paradoxical so DA can't be true.

wiki

Every possible observer exists

If all possible observers exist, the existence of one particular observer does not provide any evidence.

Non-random examples

Many try to refute DA based on some non-random examples (like a child for the case of predicting human life expectancy – Caves)

SIA exactly compensates SSA

DA in the timeless universe

In the timeless universe there is no special NOW moment which is randomly put in the timeline and from which we could conclude about the timeline structure.

DA requires deterministic world

Leslie claims that for DA to be valid the world should be deterministic (and so our future existence time is defined by now) and result is like timeless theory

Reference class is defined by having a question

If I am interested in DA I belong by definition to the class of observers who are surprised to be so early in their civilizations

End of the world is special case

for which DA-style arguments can't be applied

Rebuttal of the rebuttals

Observer has random properties

If we correctly define referent class I will be random in it. This class is all those who know and understand the DA.

If I am not a random observer, I still have many random properties, like name, date of birth etc.

Aliens

SIA implies existence of very large or infinite universe there all possible observers exist.

But it does not provide information about the size of any particular civilization.

Also, it increases the probability that the Great Filter is in our future.

Different outcomes are possible

The end of any math theories is not rebuttal.

Math theories do not necessary die through explicit rebuttal. They may be forgotten, as when its originator dies. So DA could be true and never explicitly rebutted. Even if it seems that many salient mathematical theories are refuted, the majority are just forgotten and their actual truth value is never evaluated.

So DA could be true and means extinction.

Random distribution is not affected by the fact that everything exist

Even if everything exist, this does not prevent random distribution of observer properties, which may be demonstrated in other domains.

DA-like predictions work in other domains, even if everything exist: birth date.

Even if all observers exist, they can exist in different probability distributions, so that does not refute DA.

DA is statistical

DA is a statistical argument and it may not work in some non-random examples (like how temperature is statistical measure of the speed of molecules and some molecules may be slower).

SIA self rebuttal

If many different universes exist, and one of them has infinite number of all possible observers, SIA imply that I must be in it.

But if infinite number of all possible observers exist, the condition that I may not be born is not working in this universe and I can't apply SIA to the Earth's fate.

But DA-style theories work in other domains

In other situations the result is as if NOW moment is random.

Median life expectancy version of DA works in indeterministic world

DA works for other reference classes too

No evidence for that

Positive DA solutions

DA will be proven to be false

Most promising: non-randomness of our position: even if "immortal" civilizations exist, they will be worrying about DA in the beginning, so some worrying observers do exist even in best scenarios.

DA narrow escape

Most technological civilizations go extinct but some do not, and we will do everything possible to be the one that doesn't.

Change of the reference class

DA is true but we will move to another reference class:
- will become uploads,
- will create AI,
- change birth rate,
- stop thinking about DA.

Observer selection effects result in survival

Many world quantum immortality on civilizational level.

DA is false but it can't be proven

If we exist long enough we will prove empirically that DA does not work, but we will not know why.

DA may be proved to be unprovable. We may prove that we will never know was it right or not.

DA is true but does not mean global catastrophe

It will be forgotten, or solved, or reference class will change, or birth rate decline.

It all adds up to normality

DA is true but is compensated by other theory like SA, multiverse, SSI, BBs, FAI.

DA is non-informative

We already know that dangerous x-risks exist soon ahead, and DA only confirms it or changes their probability in insignificant and unmeasurable way.

Manipulating the extinction probability using Doomsday argument

- Decision to create more observers in case that unfavourable event X starts to happen, thereby lowering its probability (method UN++ by Bostrom).
- Lowering birth's density to get more time for the existence of the civilization.
- Change the way of the end of our reference class.
- Try to be unique (and non-random) to escape DA logic.
- Many worlds immortality on civilizational level is used to create universal probability manipulator.

Meta-Doomsday argument

- Bayesian DA:
We don't know truth about DA but there is a probability that one of the versions is true
- Reverse DA:
The fact that we see global risks soon ahead says that some kind of DA is probably true
- Two-level DA:
Use Gott formula to estimate future existence time, then use Carter formula to upgrade estimation of future x-risks using Gotts result as prior.
- Decision theory DA approach:
How DA could influence our decisions – be more cautious? (Armstrong)

Future research plan

We should make some research to prove or disprove DA. The main line of such research would be to try similar to DA logic in other domains, starting from most mundane, like human age, and up to most complex and civilizational level ones.

For example, it is possible to measure size of the Earth using only Copernican mediocracy principle. All we need to know is distance from my birth place to equator, and assumption that human birth places are distributed on randomly distances from equator. I was born 6000 km from equator, and it may be used to conclusion that Earth's radius is around 7000 km which is almost true (real radius is near 6300) Exact calculations here should be more complex, as it must take in account spherical distribution of the observers and will result in bigger radius.

(c) Alexey Turchin, 2015

© Alexey Turchin, 2015, GNU-like license. Free copying and design editing, discuss major updates with Alexey Turchin.
http://immortality-roadmap.com/ - last updated version here
Proofreading and advise: Michael Anissimov