Quantum Immortality

A Perspective If AI Doomers Are Probably Right

Alexey Turchin[1] and James D. Miller[2]

1. Digital Immortality Now, Open Longevity, alexeiturchin@gmail.com.
2. Department of Economics, Smith College, jdmiller@smith.edu.

ABSTRACT

What should you believe will be the likely cause of your survival if you are an AI Doomer who believes in quantum immortality? An AI Doomer believes that if their understanding of AI existential risks is correct, humanity must get extremely lucky to survive. A person who believes in quantum immortality thinks that the many-worlds interpretation of quantum physics implies that their consciousness will necessarily continue in some branch of the multiverse. This article shows that combining these two beliefs requires rejecting or going beyond simple Bayesian reasoning when estimating one's most likely cause of survival. Such estimates may have practical implications for investment decisions, reputation management, and anthropic shadows.

Introduction

There are several ways to combine an AI Doomer's view and Quantum Immortality (QI):

1. QI is true and I alone will survive AI apocalypses for some strange reasons.
2. Most of my survival probability mass will occur in timelines where some random event has prevented AI Doom globally, e.g., Taiwan war.
3. Most of my survival probability mass will be in logically possible timelines where AI alignment is easy universally for some logical reasons.
4. We have to abandon using probability estimates when we combine QI and the AI Doomer's view.

We, moreover, claim that the first three variants can be observed beforehand – can have observable harbingers – as a form of "future anthropic shadow". We illustrate this with a thought experiment where a person can correctly bet on guessing an unknown digit of the number π if he will be killed with some probability in the case of wrong guesses. Such possible harbingers for our three variants are: my unique position for future survival, international tensions around Taiwan, and the apparent easiness of Large Language Model (LLM) alignment.

Ultimately, we generalize the notion of the future anthropic shadow into a broader principle centered on observer fate uniqueness, predicting elevated chances of immortality possibly by merger with superintelligent AI.

Motivating Analogy

Before sleeping, I try to guess the 10th digit of π, presently a mystery to me. After falling asleep, seven coins will be flipped. Assume quantum uncertainty affects how the coins land. I survive the night only if I correctly guess the 10th digit of π and/or all seven coins land heads, otherwise I will be killed in my sleep.

Convinced of quantum immortality (Randall, 2004; Turchin, 2018), I am confident of surviving the night. How then should I expect my future self to likely rationalize this survival? According to simple Bayesian reasoning, the most probable cause for my survival would be accurately guessing the 10th digit of π. However, this suggests that before sleeping, I ought to consider my guess regarding this digit as probably correct, a concept that appears nonsensical.

Quantum immortality should influence my belief about whether the future me will think all the coins came up heads because my consciousness is more likely to persist in the branches of the multiverse where this happens. But quantum immortality should not affect whether future me thinks I have already guessed the 10th digit of π correctly because the accuracy of my guess is consistent across the multiverse. By this chain of logic, if I am convinced future me will survive, I should think it far more likely I will survive because of the coin flips than guessing the 10th digit of πcorrectly. But this goes against normal Bayesian reasoning.

Now imagine that I am an AI doomer who thinks there are two ways I will survive: (a) if I am wrong about AI existential risk, (b) if humanity gets extremely lucky. Furthermore, assume that (a) is not influenced by quantum luck, but (b) is. Imagine I estimate (a) at 10% and (b) at 1/64. If I am convinced of quantum immortality it is not straightforward whether I should believe my reason for survival is more likely to be (a) or (b).

Giving Up on Estimate Probabilities

Perhaps the notion of quantum immortality makes it impossible to estimate probabilities, and so AI doomers who believe in quantum immortality should not seek to estimate their likely cause of survival. But giving up has significance compared to going with straightforward Bayesian probabilities.

Assume I am very status-conscious and would only publicly support AI doomers if it bolsters my future reputation for wisdom. If humanity survives just due to quantum luck, validating the AI doomers' accuracy, future generations may well perceive them as wise, as it will be apparent that we only survived because of amazing luck. On the other hand, if AI doomers are proven incorrect, they will be deemed foolish by posterity. Thus, demonstrating that simplistic Bayesian estimation often overreaches might persuade status-conscious individuals to endorse the AI doomer viewpoint.

This issue might also be relevant to investment strategies. Imagine that, assuming the AI doomers are right, AI will likely become much more powerful in the short run. This is largely due to a primary reason for the doomers' potential miscalculation: AI might only reach human-level intelligence in several vital areas. Assuming the doomers are correct yet

humanity survives through quantum luck, a long-term investment in an AI-heavy company like Microsoft would yield the highest returns. Since I will only benefit from my long-term investment if humanity survives, giving up on estimating the likely causes for my survival would make it near impossible to develop an optimal investment strategy.

Scenarios of survival via quantum immortality of AI Doom

I will be the only survivor of misaligned AI

If we count AI doom as a type of risk similar to an asteroid impact, there will always be timelines where I survive due to a combination of random circumstances, and because of quantum immortality, I will find myself in such a timeline. For example, in the case of an asteroid impact, there could be a place where different impact waves compensate for each other, and I would find myself alive in such a place, though likely badly injured.

In the case of AI Doom, it could turn out that the AI has a fluctuation in its utility function which requires the preservation of just one person for research purposes, and it turns out to be me.

While this looks like a way of survival under quantum immortality, if we consider the whole set of possible ways I could survive AI, say, 10 years from now, it would be only a minuscule part of it, at least if we use traditional probability estimates.

Being the only survivor is also a type of s-risk (a possible outcome of misaligned AI where it instead of killing humans, causing them infinite sufferings).

Humans will survive because misaligned AI will never happen

The more likely outcome where I survive is one where there is no AI Doom. Here there are two variants:

1) AI will be aligned. Most civilizations in our situation will survive because alignment is easy. “Alignment easiness” is a universal mathematical property, similar to the digits of π.
2) AI will not be created. Most civilizations in our situation would experience AI doom because alignment is hard or impossible and you need extreme quantum influenced luck to survive. In that case, the survivors will survive because of some random unique reasons, like nuclear war, which stops AI development.

Logical probability and can QI be applied to it?

We use two different types of probability. When we speak about random events preventing AI creation, we use normal frequentist probability, which corresponds to the share of real worlds with some outcome.

However, for alignment easiness, we need to use logical probability (Garrabrant et al., 2016). Logical probability deals with questions like what are the chances that the nth digit of π is 9. It is not obvious that we can use logical probability for quantum immortality, because in the case of a wrong answer, there is no physical share of worlds.

QI can be applied to guessing the digits of π (and thus escaping logical probability by replacing it with frequentist one) if we assume that the same experiment is happening in many different worlds and different digits of π are under question.

The problem with the original thought experiment is that the "me" that guesses "7" could be considered the same as the me that guesses "8". But this is solved with a more complicated thought experiment where I apply to 5 colleges, have a friend collect the decision letters, and then instead of guessing π, I guess the colleges I got into. Since which colleges I get into will greatly influence my future life, it isn't the same me that correctly guessed getting into all as getting into none.

Another way to make logical probability look like frequentist probability is to suggest that the alignment problem is not unique, and there are several similar abstract global risks, like "complexity always self-destructs", "coordination is impossible", "biological weapons are easy". We survive only in the worlds where a favorable combination of abstract risks is present. But in such different worlds "we" will not be the same.

Applying functional identity theory may help: I regard as me anyone who is in the same epistemic situation and has my current line of thoughts and ignore all personal details. In that case, an alien reptile which is thinking about anthropic effects will also be "me" (Yudkowsky & Soares, 2017). But from human self-preservation view such outcome is unsatisfactory.

A third way to treat alignment chances as frequentist probability is to suggest there is something specific about our alignment; for example, human language is especially good for developing LLMs and next-token predictors, and some aliens with different thinking mechanisms would not benefit from LLMs in alignment.

Or we can bite the bullet and say that we are more likely to observe even high logical probability favoring our survival. This is basically the idea behind the Presumptuous Philosopher thought experiment: if there are two variants of the theory of universe and under one of them there are 1000 times more observers, the fact that I exist at all is evidence -- under Self-Indication-Assumption - that the second theory is more probable. This works if modal realism is false (not all observers exist) and there is a pool of non-born observers from which born observers are "pulled". Only in that case the fact that I was born is evidence that there will be maximum number of observers. However, as I discussed in the post this idea is self-defeating as it gives the biggest probability mass to the hypothesis that all possible observers exist -- which is needed and assumed by our applying quantum immortality.

In other words, SIA proves quantum immortality. But we can apply it again and say that not only will there be an infinite number of observers, but also the highest concentration of observers, which requires logical probability of easy alignment to be high. Turchin applied similar reasoning to prove that panspermia is likely (Turchin, 2020).

Quantum immortality and identity

Another aspect of quantum immortality relates to what "I survive" means. There are two interpretations: a set of worlds where me-10-years-older exists, or a share of worlds to

which there is a continuous path from my current state where I survive. The difference is that a person with my memory may not be my continuation. For example, in the multiverse there can be a different world, with a slightly different political configuration, where my copy exists and will eventually survive all global risks because of that different political configuration.

QI depends on identity theory. The difference is between mind-state-based identity and continuity-based theory of identity. Deciding which identity theory is correct is out of scope of this paper, but mind-state-based identity theory seems preferable to the authors as continuity is unmeasurable and fragile; eventually both theories merge as continuity in MWI can pass through all possible minds (more in Turchin's forthcoming "Immortality and identity").

Correctly guessing a digit of π the thought experiment above requires mind-state-based theory of identity, where we look at the shares of future worlds where minds with my memory exist, without caring whether they are continuously connected to me.

War in Taiwan and the "future anthropic shadow"

A possible war in Taiwan as a potential way to stop AI Doom is an example of reasons for survival because of luck.

It was suggested (by gwern) that a possible war in Taiwan would cause hardware shortages that would pause AI development globally, and that US sanctions on China's AI tech increase the chances of such a war. Moreover, commentators suggested that our presence in a timeline where such a war is likely and could stop AI development might be explained by quantum immortality (note that this argument became weaker in 2025 when China developed powerful DeepSeek AI and started producing domestic AI chips).

They incorrectly used the term 'anthropic shadow,' which was originally used by Bostrom to denote something akin to survivorship bias (Ćirković et al., 2010) – that is, the underestimation of past risks, not the change of future probabilities caused by quantum immortality; let's call this modified idea the 'future anthropic shadow.'

> *'This is also related to the concept of an anthropic shadow: if artificial intelligence was to cause human extinction but required a lot of computing power, you would be more likely to find yourself in world lines in which the necessary conditions for cheap computing are not met. In such world lines, crypto miners causing a GPU shortage, supply chain disruptions due to a pandemic, and a war between the United States and China over Taiwan in which important chip fabrication plants are destroyed are more likely to occur in world lines that are not wiped out. An anthropic shadow hides evidence in favour of catastrophic and existential risks by making observations more likely in worlds where such risks did not materialize, causing an underestimation of actual risk (https://twitter.com/XiXiDu/status/1582440301716992000).*

We can define 'future anthropic shadow' as finding evidence now that you will survive via quantum immortality an impending catastrophe in the future. Future anthropic shadow will become a normal anthropic shadow after the event: if we survive, we will see many

harbingers of our survival in the past, which can be explained only by observation selection effects. If Donald Trump takes an improbably action that saves us from AI doom, his barely escaping assassination will likely be seen as a normal anthropic shadow.

Future anthropic shadow is similar to our thought experiment with π-guessing, as you have more credence that currently observed events will cause future survival than you should have without applying quantum immortality.

In some sense, future anthropic shadow is a reverse version of the Doomsday Argument: instead of 'I live in a world which will end soon,' we get 'I live in the world most suitable for survival.'

An example of observable “miracles” based on anthropics is Adam and Eve thought experiment by Bostrom where Adam and Eve can confidently predict that they will not conceive based on expected shortness of their timeline (Bostrom, 2001).

Are we getting new information from the future via future anthropic shadow?

Future anthropic shadow may have two interpretations:

1) Betting: Only a bet on long survival will be paid. Here there is no new information, but we bet as if we have it, because if there is short survival, there will be no payments on the bet. Thus, we admit that betting strategy differs from objective truth. Or if we think that betting reveals real truth, then we get new information. Similarly, a quantum Russian roulette player can (and should) bet that the gun will not fire. And likewise, we bet on correctly guessing the pi-digit.
2) Truth level: There is a natural selection effect that favors longer timelines. This selection principle differs from classical SIA or SSA. It claims that I am randomly selected from all timelines proportional to their length (and this is not compensated by later learning my actual position in time). Here the timeline's length plays the role of its weight or measure.
3) The classical test ground for anthropic thought, Sleeping Beauty, can illustrate this idea:
4) Betting version: After awakening but before learning the day of the week, she can correctly bet that she is in the longer timeline—typically, tails. After learning that today is Monday, she will not update her bet because she knows that in the longer timeline she will continue to bet on tails (on Tuesday), effectively doubling her bet on tails. Or, if she is killed at the end of the experiment, she should bet on tails if she can use her Monday bet on Tuesday.
5) Truth version: Learning the current date does not exactly compensate for the update in the longer timeline. For example, in a very long timeline (million days in tails), there will be days when she is only dreaming of being on Monday before real awakening, and even if it is a rare dream, this possibility distorts the update in the direction of the longer timeline.

6) A very strong truth version: The measure is timelessly redistributed along the timeline. This idea is not as obvious as the previous ones. First, we must mention that in the continuity-based identity model, different copies can have different probabilities, depending on how they were created: if a copy of me were created and after that this copy was copied again, I get chances to be any of them 1:2, 1:4, 1:4. Inside this process, it will look like the pathway leading to the 1:2-copy gets more chances. If we assume that at the end of some timelines there is a mind with the highest possible measure, then the timelines that lead to it get a similar probability boost. Thus, timelines leading to very long survival or other ways of measure boost will get a probability update, and it can be observed as the future anthropic shadow.

Future anthropic shadow for alignment easiness

Just as we can be sure that we correctly guessed the digit of π and that we are observing signs of a future war that will stop AI, we can claim that we currently observe signs that alignment will be easy. There are two such signs:

- LLMs learn human values easily
- LLMs may have double diminishing returns from scaling: their intelligence grows only logarithmically with scale, and the efficiency of that intelligence in the world also grows logarithmically with intelligence. As a result, AI impact may be self-limiting
- Having both shadows simultaneously is not contradictory.

The fate of the observer

The idea of quantum immortality is that the fate of an observer from their own point of view differs from the fate of a random mind: the observer will survive. What we described here adds to this idea: the observer will not only survive but will be more likely to exist in a world where AI Doom will not happen. For the same reasons, other catastrophes will also not happen in most of my surviving futures.

Also, long-term survival is more likely in a world where life extension technologies are developed. Current growth of life extension technologies can be seen as a future anthropic shadow of that. It is surprising to both authors that they were born at a time in which everyone accepted they would die but now think they have a meaningful chance of being alive in a million years.

No AI Doom also means higher chances of eventually developing friendly AI which will ensure my long-term survival.

These factors combined seem like a 'free lunch' which allows me to live a much longer and better life than I could expect. Moreover, a logical continuation of this idea is that I will eventually reach a god-like status, probably via merger with superintelligent AI. We call this free lunch 'transcendental advantage.' In other words, I will evolve in the direction of a being with the highest measure, that is, with the highest share in the multiverse (this claim is rather tautological, as the state with the highest measure is the one in which all other timelines end).

But as we know, all free lunches are not actually free. There can be dangerous caveats like s-risks. The question arises: how can we steer QI outcomes in the direction of the best possible futures? This requires additional research.

Conclusion

Quantum immortality guarantees our survival, but we can influence how it will happen. In the case of AI risk, we need to perform actions that increase the chances of the best outcome—that AI alignment works—compared to bad outcomes: nuclear war as the end of AI risk, myself as the sole survivor of an AI catastrophe, and s-risks.

This means that we need to continue working on AI alignment, perhaps using future anthropic shadow to guess the most promising directions: for example, if anthropic shadow implies that LLM alignment is easy, we should pursue that direction.

References

Bostrom, N. (2001). The Doomsday Argument Adam & Eve, UN++, and Quantum Joe. *Synthese*, *127*(3), 359–387.

Ćirković, M. M., Sandberg, A., & Bostrom, N. (2010). Anthropic shadow: Observation selection effects and human extinction risks. *Risk Analysis, Vol. 30, No. 10, 2010*.

Garrabrant, S., Benson-Tilsen, T., Critch, A., Soares, N., & Taylor, J. (2016). Logical induction. *arXiv Preprint arXiv:1609.03543*.

Randall, A. F. (2004). Quantum miracles and immortality. *Quantum*, *5*, 8.

Turchin, A. (2018). Forever and Again: Necessary Conditions for the “Quantum Immortality” and its Practical Implications.

Turchin, A. (2020). Presumptious philosopher proves panspermia.

Yudkowsky, E., & Soares, N. (2017). Functional Decision Theory: A New Theory of Instrumental Rationality. *arXiv:1710.05060 [Cs]*. http://arxiv.org/abs/1710.05060