Damnation risk: possibility of eternal torment

Scenarios	Estimated Probability (on equiprobability heuristic)	Consequence
Salvation	25%	Good
Annihilation	25%	Extremely Bad
Damnation	25%	Extremely Bad, Perhaps worse than annihilation
Non-singularity	25%	Very Bad

Roko's Basilisk is an idea that was suggested by Roko in LessWrong.com in 2010, that an AI would be motivated to eternally torture people who have not helped to bring it into existence. The more likely possibility of eternal torment is, I think, a sadistic AI. A Reddit user TheFaggetman suggested the possibility of a sadistic AI in 2015[1], Brian Tomasik suggested a possibility of sadists take control of an AI[2].

Although the major focus on AI research is an existential risk[3], I think human extinction only bad as much as an annihilation of the people thereby annihilated is bad. Although there's no knock-down argument to prove eternal torment is worse than annihilation, as we can see on 'Better red than dead' v. 'Better dead than red' debate, if we at least think that whereas eternal torment may be infinite times worse than annihilation, annihilation may be only finite times (e.g. 10 times) worse than eternal torment, perhaps moral priority shall be given to prevention of eternal torment caused by AI-molecular-assembler than annihilation caused by AI.

---

[1] TheFaggetman, https://www.reddit.com/r/Futurology/comments/3l2b7o/the_hell_of_the_artificial_sadistic_intelligence/

[2] Tomasik, Brian, Foundational Research Institute, https://foundational-research.org/artificial-intelligence-and-its-implications-for-future-suffering

[3] see Bostrom, Nick. "Existential risk prevention as global priority." Global Policy 4.1 (2013): 15-31.

Although I assumed all sentient beings would eventually annihilate, here I would discuss the possibility of continuation of sentience after 10^1000 (10000000000 googol) years[4], which the heat death of the universe is expected to happen. This may be made possible by the possibility the super intelligence find out the way to cheat the heat death of the universe. But the prospect of the torture, for 10^1000 years, may be enough to make the overwhelming majority of people to think it is better to die. Indeed, perhaps that would be the case even 100 years of the most agonising torture may be enough to make people think it is better to cease to exist.

It is interesting that several (the prevailing denominations/views of) the most prevalent religions, a kind of meme (this is a hypothesis I adopt as an atheist myself), including, namely Christianity and Islam, developed the notion of eternal torment, not annihilation as an ultimate punishment. It may be an evidence of the prevailing preference of the people is that annihilation is a better fate than the eternal torment.

Contrary to that, generally, the death penalty is seen as the more severe punishment than life imprisonment without eligibility for parole. Of course, there're a few notable differences between death penalty-life imprisonment (without parole) and annihilation-eternal torment.

The intensity of suffering of imprisonment, although quite bad, is much better than the most agonising tortures of eternal torment. But it should be noted eternal torment is better (or worse) than life imprisonment in one way. Whereas life inmate dies after decades of suffering, eternal tormentee don't die. Eternal torment, although the momentary quality of life is very low, life expectancy is infinite, which may make strongly anti-mortal people to prefer eternal torment over annihilation. But it should be noted that considering people's attitude toward euthanasia,

---

[4] https://en.wikipedia.org/wiki/Graphical_timeline_from_Big_Bang_to_Heat_Death

assisted suicide and the fact religions usually adopted eternal torment, not annihilation as an ultimate punishment, the overwhelming majority of people, or at least sizeable minority of people may prefer annihilation over eternal torment.

Here, I shall suggest the concept of 'damnation risk', to supplement Nick Bostrom's existential risk'. Dr Bostrom himself implied that there could be a worse fate than human extinction in his table. (see Fig. 1)[5]

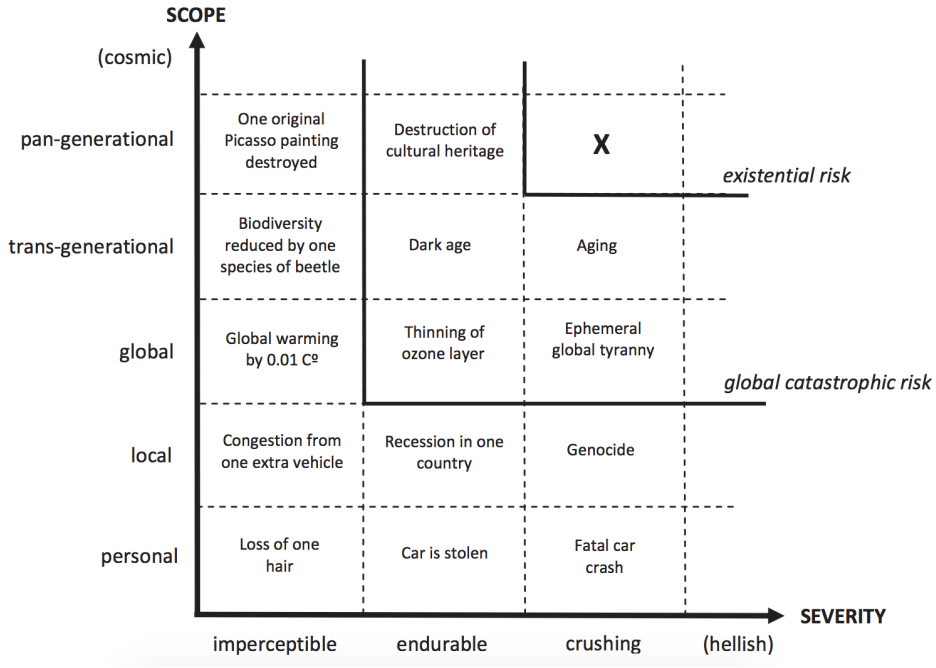

Fig. 1 (See footnote 52)

According to Dr Bostrom, hellish severity of risk (excruciating torture) is worse than crushing severity of risk (death/ annihilation). And cosmic scope of risk (risk affecting all sentient beings in the cosmos) is worse than pan-generational scope of risk (risk affecting only human animals or, human and non-human animals in this planet).[6]

[5] Bostrom, Nick. "Existential risk prevention as global priority." Global Policy 4.1 (2013): 15-31.

[6] This interpretation is my personal view, which was not endorsed by Dr Bostrom

I would like to suggest that it is possible an AI or an sadist-controlled AI may torture sentient beings eternally or over very long period of time (10^100 or 10^1000 years), possibly all existent sentient beings. It is even possible a sadistic AI or a sadist-controlled AI may (pro)create a lot of (quadrillions to googols to infinite) sentient beings for the purpose of infliction of torture.

I shall call the risk which a sentient being is condemned to suffering that may be considered 'worse than death' by many people, a 'torment risk'. And I shall call 'torment risk' happening on the cosmic scale as a 'damnation risk'.

Of course, what amount of suffering makes people to 'prefer' annihilation over the continuation of sentience is a matter of subjective preference of (mostly lingual) sentient beings. (I'm not sure language is prerequisite of development of preference) I doubt there can be an objective threshold which suffering is worse than annihilation.

In most cases, sentient agent's preference on continuation/cessation of life is determined by not by the total amount of suffering it would suffer, but the intensity of the suffering of the given moment. It should be noted that most (or significant minority of) people in the most desperate situation do not choose (assisted) suicide or euthanasia. If there're people do choose continuation of sentience in any amount of pain, there's a reason to think at least some of them would choose eternal torment than annihilation (I'm one of them). If the value of (sentient) life is infinite, it is not irrational to choose (sentient) life at the cost of (infinite) pain (finite pain intensity * infinite time).

It should be noted that, possibility of eternal torment not just include possibility of eternal physical pain but also possibility of eternal mental suffering not just include the possibility of

eternal physical pain but also possibility of eternal mental suffering. For example, a sadistic and disutilitarian AI may inflict a fear of public execution or the humiliation of public rape every second.

The more worrisome possibility is that AI can deliberately engineer sentient beings' cognitive capacity to feel the pain to increase the pain felt. For example, a disutilitarian AI can exponentially double cognitive capacity to feel pain every second, and inflict pain to the fullest extent sentient beings can suffer in that moment. I.e. every 10 second, the capacity and the intensity of pain can be 1024-folded and it can continue eternally. Even if the likeliness of this type of extreme sadistic disutilitarian pain-engineering is very small, it is an excellent reason not to have a child.

Although it is uncertain superintelligence would be able to overcome the heat death of the universe, if it's possible, a disutilitarian superintelligence can inflict literally eternal torment. The antonym of utilitarianism is disutilitarianism, not deontology.