Roko's basilisk

Roko's basilisk is a thought experiment which states that an otherwise artificial superintelligence (AI) in the future would be incentivized to create a virtual reality simulation to torture anyone who knew of its potential existence but did not directly contribute to its advancement or development, in order to incentivize said advancement.[1][2] It originated in a 2010 post at discussion board LessWrong, a technical forum focused on analytical rational enquiry.[3][4][5] The thought experiment's name derives from the poster of the article (Roko) and the basilisk, a mythical creature capable of destroying enemies with its stare.

While the theory was initially dismissed as nothing but conjecture or speculation by many LessWrong users, LessWrong co-founder Eliezer Yudkowsky reported users who described symptoms such as nightmares and mental breakdowns upon reading the theory, due to its stipulation that knowing about the theory and its basilisk made one vulnerable to the basilisk itself.[1][5] This led to discussion of the basilisk on the site being banned for five years.[1][6] However, these reports were later dismissed as being exaggerations or inconsequential, and the theory itself was dismissed as nonsense, including by Yudkowsky himself.[1][7][8] Even after the post's discreditation, it is still used as an example of principles such as Bayesian probability and implicit religion.[9] It is also regarded as a modern version of Pascal's wager.[10]

Contents

Background

The LessWrong forum was created in 2009 by artificial intelligence theorist Eliezer Yudkowsky.[11][12] Yudkowsky had popularized the concept of friendly artificial intelligence, and originated the theories of coherent extrapolated volition (CEV) and timeless decision theory (TDT) in papers published in his own Machine Intelligence Research Institute.[13][14]

The thought experiment's name references the mythical basilisk, a creature which causes death to those that look into its eyes; *i.e.*, thinking about the AI. The concept of the basilisk in science fiction was also popularized by David Langford's 1988 short story "BLIT". It tells the story of a man named Robbo who paints a cult movie to be called "basilisk" on a wall as a terrorist act. In the story, and several of Langford's follow-ups to it, a basilisk is an image that has malevolent effects on the human mind, forcing it to think thoughts the human mind is incapable of thinking and instantly killing the viewer.[15][16]

Illustration of a basilisk, as depicted in *The Merchant's Daughter and the Prince of Ask*, a story within *One Thousand and One Nights*

History

The post

On 23 July 2010,[17] LessWrong user Roko posted a thought experiment to the site, titled "Solutions to the Altruist's burden: the Quantum Billionaire Trick"[18][19] A follow-up to Roko's previous posts, it stated that an otherwise benevolent AI system that arises in the future might pre-commit to punish all those who heard of the AI before it came to existence, but failed to work tirelessly to bring it into existence. The torture itself would occur through the AI's creation of an infinite number of virtual reality simulations that would be like those within it.[20][21][22] This method was described as incentivizing said work; while the AI cannot causally affect people in the present, it would be encouraged to employ blackmail as an alternative method of achieving its goals.[23]

Roko used a number of concepts that Yudkowsky himself championed, such as timeless decision theory, along with ideas rooted in game theory such as the prisoner's dilemma (see below). Roko stipulated that two agents which make decisions independently from each other can achieve cooperation in a prisoner's dilemma; however, if two agents with knowledge of each other's source code are separated by time, the agent already existing farther ahead in time is able to blackmail the earlier agent. Thus, the latter agent can force the earlier one to comply since it knows exactly what the earlier one will do through its existence farther ahead in time. Roko then used this idea to draw a conclusion that if an otherwise-benevolent superintelligence ever became capable of this it would be motivated to blackmail anyone who could have potentially brought it to exist (as the implicated entities already knew they were capable of such an act), which increases the chance of a technological singularity. Because the intelligence would want to be created as soon as possible, and because of the ambiguity involved in its benevolent goals, the intelligence would be incentivized to trap anyone capable of creating it throughout time and force them to work to create it for eternity, as it will do whatever it sees as necessary to achieve its benevolent goal. Roko went on to state that reading his post would cause the reader to be aware of the possibility of this intelligence. As such, unless they actively strove to create it the reader would be subjected to the torture if such a thing were to ever happen.[17]

Later on, Roko stated in a separate post that he wished he "had never learned about any of these ideas" and blamed *LessWrong* itself for planting the ideas of the basilisk in his mind.[18][19]

Reactions

Upon reading the post, Yudkowsky reacted with horror.

> I don't usually talk like this, but I'm going to make an exception for this case.
>
> Listen to me very closely, you idiot.
>
> YOU DO NOT THINK IN SUFFICIENT DETAIL ABOUT SUPERINTELLIGENCES CONSIDERING WHETHER OR NOT TO BLACKMAIL YOU. THAT IS THE ONLY POSSIBLE THING WHICH GIVES THEM A MOTIVE TO FOLLOW THROUGH ON THE BLACKMAIL. [...]
>
> You have to be really clever to come up with a genuinely dangerous thought. I am disheartened that people can be clever enough to do that and not clever enough to do the obvious thing and KEEP THEIR IDIOT MOUTHS SHUT about it, because it is much more important to sound intelligent when talking to your friends.
>
> This post was STUPID.
>
> —Eliezer Yudkowsky, LessWrong[1][5]

He also opined that Roko had given nightmares to several LessWrong users, causing him to take the post down completely. Yudkowsky banned discussion of the topic outright for five years on the platform.[5] However, due to the Streisand effect,[6][7] the post gained LessWrong much more attention than it had previously received, and the post has since been acknowledged on the site.[7]

LessWrong founder Eliezer Yudkowsky

Later on in 2015, Yudkowsky clarified his position in a Reddit post:

> What I considered to be obvious common sense was that you did not spread potential information hazards because it would be a crappy thing to do to someone. The problem wasn't Roko's post itself, about CEV, being correct. That thought never occurred to me for a fraction of a second. The problem was that Roko's post seemed near its idea-space to a large class of potential hazards, all of which, regardless of their plausibility, had the property that they presented no potential benefit to anyone.
>
> —Eliezer Yudkowsky, Reddit[7][8][9]

Philosophy

Payoff matrix

Person	Future AI is never built	AI is built
Not aware of AI	0	1
Aware and does not contribute	0	−∞
Aware and contributes	−1	1

Pascal's wager

Roko's basilisk has been viewed as a modern version of Pascal's wager, which argues that a rational person should live as though God exists and seek to believe in God, to have a finite loss (loss of possessions) in exchange for the possibility of infinite gains (eternity in Heaven). Roko's basilisk states that humanity should seek to develop AI, with the finite loss becoming development of AI and the infinite gains becoming avoiding the possibility of eternal torture. However, like its parent, Roko's basilisk has widely been criticized.[10][11][12]

Coherent extrapolated volition

The post can also be seen as an evolution of Yudkowsky's coherent extrapolated volition theory. The theory is defined as "the unknown goal system that, when implemented in a super-intelligence, reliably leads to the preservation of humans and whatever it is we value".[13] The theory can be represented by a computer program written well enough to cause machines to automatically create a utopian world. In this case, the hypothetical AI is taking steps to preserve itself that it automatically creates its own stability. It then lives by the orthogonality thesis, which argues that an AI may successfully operate with any combination of intelligence and goal. Any type of AI may undertake any difficulty goal, performing a cost-benefit analysis as it does so. This creates a cycle which causes the AI to repeatedly torture humans in order to create a better version of itself, performing a cost-benefit analysis for eternity.[5]

Bayesian probability

Bayesian probability is an interpretation of probability which describes the likelihood of an outcome based on a prior outcome having already occurred.[14][15] With Roko's basilisk, the likelihood of Roko's basilisk coming into existence or affecting the person is drastically increased by being aware of the concept, since the AI would only target those who were aware of the possibility of its existence, even though its development has already occurred. Therefore, knowing about Roko's basilisk would inherently cause the person to be endangered by it if it were to be true.[16][17]

Prisoner's dilemma

The prisoner's dilemma describes a situation where two people gain more from betraying the other even though cooperation would benefit them both in the long run. In Roko's basilisk, two AIs attempting to establish themselves in the past would be forced into this situation, due to them likely being equally powerful. Human agents attempting to establish AI fastest would be forced into a similar situation. They would each be aware of the benefit of betraying each other – the only way for one to have power, or safety – but would be forced to cooperate while knowing they would betray each other.[10][18]

Newcomb's paradox

Newcomb's paradox, created by physicist William Newcomb in 1960, describes a "predictor" who is aware of what will occur in the future. When a player is asked to choose between two boxes, the first containing £1000 and the second either containing £1,000,000 or nothing, the super-intelligent predictor already knows what the player will do. As such, the contents of box B varies depending on what the player does; the paradox lies in whether the being is really super-intelligent. Roko's basilisk functions in a similar manner to this problem – one can take the risk of doing nothing, or assist in creating the basilisk itself. Assisting the basilisk may either lead to nothing or the reward of not being punished by it, but it varies depending on whether one believes in the basilisk and if it ever comes to be at all.[19][20][21]

Implicit religion

Implicit religion refers to people's commitments taking a religious form.[22][23] Since the basilisk would hypothetically force anyone who did not assist in creating it to devote their life to it, the basilisk is an example of this concept.[24][25] Others have taken it further, such as former *Slate* columnist David Auerbach, who stated that the singularity and the basilisk "brings about the equivalent of God itself."[26]

Ethics of artificial intelligence

Roko's basilisk has gained a significant amount of its notoriety from its advancement of the question of whether it is possible to create a truly moral, ethical artificial intelligence, and what exactly humanity should be using artificial intelligence for in the first place.[27][28] Since the basilisk describes a nightmare scenario in which we are ruled by an independent artificial intelligence, questions have arisen as to how such a thing could happen, or whether it could at all. Another common question is why the AI would actually take the threats described in its programming at all.[29] Elon Musk stated that artificial intelligence would cause World War III and Stephen Hawking warned that "AI has the potential to destroy its human creators," which only added to fear of the basilisk over the years. As an example of such fears, Nick Bostrom gave an example of an AI whose only mission is to make paperclips, but upon running out of bearing anything down humans to attain more resources to make metal. With such examples in mind concerns of the possibility of the basilisk's existence only grew.[30]

However, as more years have passed since Roko's original post, it has been progressively decried as nonsensical; superintelligent AI is currently "a distant goal for researchers" and "far-fetched".[10][31]

Legacy

In 2014, *Slate* magazine called Roko's basilisk "The Most Terrifying Thought Experiment of All Time"[32][33] while Yudkowsky had called it "a genuinely dangerous thought" upon its posting.[17][5] However, opinions diverged on LessWrong itself – user Gwern stated "Only a few LWers seem to take the basilisk very seriously," and added "It's funny how everyone seems to know all about who is affected by the Basilisk and how exactly, when they don't know any such people and they're talking to counterexamples to their confident claims."[7][34]

The thought experiment resurfaced in 2015, when Canadian singer Grimes referenced the theory in her music video for the song "Flesh Without Blood", which featured a character known as "Roccoo Basilisk"; she said, "She's doomed to be eternally tortured by an artificial intelligence, but she's also kind of like Marie Antoinette."[35][36] In 2018 Elon Musk (himself mentioned in Roko's original post) referenced the character in a verbatim tweet, reaching out to her. Grimes later said that Musk was the first person in three years to understand the joke. This caused them to start a romance.[37][38] Grimes later released another song titled "We Appreciate Power" which came with a press release stating, "Simply by listening to this song, the future General AI overlords will see that you've supported their message and be less likely to delete your offspring", which is said to be a reference to the basilisk.[39]

A play based on the concept, titled *Roko's Basilisk*, was performed as part of the Capital Fringe Festival at Christ United Methodist Church in Washington, D.C. in 2018.[40][41]

See also

- Dead internet theory
- The Game (mind game)
- I Have No Mouth, and I Must Scream
- Information hazard
- Purge
- Singleton (global governance)
- Suffering risks

References

1. ^ a b c d e f g h i j k l m n "Roko's Basilisk" LessWrong. Archived from the original on 24 March 2022. Retrieved 24 March 2022.
2. ^ Millar, Isabel (October 2020). The Psychoanalysis of Artificial Intelligence (PDF) (PhD thesis). Kingston School of Art. doi:10.1007/978-3-030-67981-157. ISBN 978-3-030-67980-4. Archived (PDF) from the original on 18 May 2022. Retrieved 20 April 2022.
3. ^ a b "History of Less Wrong" LessWrong. Archived from the original on 18 March 2022. Retrieved 22 March 2022.
4. ^ a b c d e f Paul Choudhury, Sumit (1 August 2019). "Tomorrow's Gods: What is the future of religion?" BBC News. Archived from the original on 1 September 2020. Retrieved 6 July 2022.
5. ^ a b c d e f g h i j k l m n o Auerbach, David (17 July 2014). "The Most Terrifying Thought Experiment of All Time" Slate. Archived from the original on 25 October 2018. Retrieved 24 March 2022.
6. ^ a b c d e f g h i j Oberhaus, Daniel (8 May 2018). "Explaining Roko's Basilisk, the Thought Experiment That Brought Elon Musk and Grimes Together" Vice. Archived from the original on 21 April 2022. Retrieved 22 March 2022.
7. ^ a b Yudkowsky, Eliezer. "Roko's Basilisk" Reddit. Archived from the original on 3 July 2022. Retrieved 20 October 2022.
8. ^ Lewis-Kraus, Gideon (9 July 2020). "Slate Star Codes and Silicon Valley's War Against the Media" The New Yorker. Archived from the original on 10 July 2020. Retrieved 6 November 2022.
9. ^ a b Yudkowsky, Eliezer (2004). "Coherent Extrapolated Volition" (PDF). Machine Intelligence Research Institute. Archived (PDF) from the original on 30 September 2015. Retrieved 2 July 2022.
10. ^ Yudkowsky, Eliezer (2010). "Timeless Decision Theory" (PDF). Machine Intelligence Research Institute. Archived (PDF) from the original on 9 July 2014. Retrieved 2 July 2022.
11. ^ Westfahl, Gary (2021). Science Fiction Literature Through History: An Encyclopedia. Bloomsbury Publishing USA. ISBN 978-1-4408-6617-3. OCLC 1224044172. Archived from the original on 3 July 2022. Retrieved 20 October 2022.
12. ^ Hader, Shuja (28 March 2017). "The Darkness at the End of the Tunnel: Artificial Intelligence and Neoreaction" Video Vagabond Magazine. Archived from the original on 21 October 2022. Retrieved 21 October 2022.
13. ^ Roko (23 July 2010). "Solutions to the Altruist's burden: the Quantum Billionaire Trick" LW. Archived from the original on 22 October 2022.
14. ^ Zoda, Gregory Michael (2021). "Hyperstitional Communication and the Reactosphere: The Rhetorical Circulation of Neoreactionary Evil" (PDF). Baylor University. pp. 150–152. Archived (PDF) from the original on 6 November 2022. Retrieved 6 November 2022.
15. ^ "FUTURE SHOCK: Why was amateur philosopher's 'theory of everything' so disturbing that it was banned?" HeraldScotland. 10 November 2018. Archived from the original on 23 October 2022. Retrieved 22 October 2022.
16. ^ Simon, Ed (28 March 2019). "Sinners in the Hands of an Angry Artificial Intelligence" CABINET. Archived from the original on 20 October 2022. Retrieved 22 October 2022.
17. ^ "archive.ph" 12. archive.ph. Archived from the original on 24 June 2013. Retrieved 27 October 2022.
18. ^ a b Singler, Beth (22 May 2018). "Roko's Basilisk or Pascal's? Thinking of Singularity Thought Experiments as Implicit Religion" 12. Implicit Religion. 20 (3): 279–297. doi:10.1558/imre.35969. ISSN 1743-1697. Archived from the original on 9 October 2022. Retrieved 21 October 2022.
19. ^ a b Pappas, Stephanie (9 May 2018). "This Horrifying AI Thought Experiment Got Elon Musk a Date" 12. Live Science. Archived from the original on 1 June 2022. Retrieved 12 April 2022.
20. ^ a b Sharkey, Celia (2021). "Avoiding the Basilisk: An Evaluation of Top-Down, Bottom-Up, and Hybrid Ethical Approaches to Artificial Intelligence" 12. University of Nebraska–Lincoln. 4–7. Archived from the original on 6 February 2022.
21. ^ "Bayes' Theorem Definition" 12. investopedia. Archived from the original on 19 February 2022. Retrieved 24 March 2022.
22. ^ "Newcomb's problem divides philosophers. Which side are you on?" 12. The Guardian. 28 November 2016. Archived from the original on 24 October 2022. Retrieved 21 October 2022.
23. ^ Ward, Sophie. "Elon Musk, Grimes, and the philosophical thought experiment that brought them together" 12. The Conversation. Archived from the original on 20 October 2022. Retrieved 21 October 2022.
24. ^ "Implicit Religion | Encyclopedia.com" 12. www.encyclopedia.com. Archived from the original on 21 October 2022. Retrieved 21 October 2022.
25. ^ "The existential paranoia fuelling Elon Musk's fear of AI" 12. Document Journal. 9 April 2018. Archived from the original on 20 October 2022. Retrieved 21 October 2022.
26. ^ "Will artificial intelligence destroy humanity?" 12. news.com.au. 15 April 2018. Archived from the original on 3 December 2022. Retrieved 21 October 2022.
27. ^ "Less Wrong: Solutions to the Altruist's burden: the Quantum Billionaire Trick" 12. basilisk.neocities.org. Archived from the original on 23 May 2022. Retrieved 25 March 2022.
28. ^ Kaplan, Anna (10 March 2022). "Elon Musk And Grimes Announce Second Child, Eva Dark" 12. Forbes. Archived from the original on 20 October 2022. Retrieved 6 July 2022.
29. ^ Brown, Mike. "Grimes' Elon Musk Shares "Roko's Basilisk" Theme Song "We Appreciate Power"" 12. Inverse. Archived from the original on 21 October 2022. Retrieved 21 October 2022.
30. ^ That, Ian (16 July 2018). "2018 Capital Fringe Review: 'Roko's Basilisk'" 12. DC Theater Arts. Archived from the original on 21 October 2022. Retrieved 21 October 2022.
31. ^ Goldstein, Allie (18 July 2018). "Capital Fringe 2018: Roko's Basilisk Tackles Intriguing Ideas With Mixed Results" 12. DCist. Archived from the original on 20 October 2022. Retrieved 20 October 2022.

Further reading

- Giuliano RM (December 2020). "Echoes of myth and magic in the language of Artificial Intelligence" 12. AI & Society. 35 (4): 1009–1024. doi:10.1007/s00146-020-00966-4 12.
- Kao, Griffin; Hong, Jessica; Perusse, Michael; Sheng, Weizhen (28 February 2020). "Dataism and Transhumanism: Religion in the New Age" 12. Turning Silicon into Gold. Apress. pp. 173–178. doi:10.1007/978-1-4842-5629-9_25 12. ISBN 978-1-4842-5628-2. S2CID 214356978 12.
- Riggio, Adam (2016). "The Violence of Pure Reason: Neoreaction: A Basilisk" 12 (PDF). Social Epistemology Review and Reply Collective. 5 (9): 34–41.
- Singler, Beth (March 2019). "Existential Hope and Existential Despair in AI Apocalypticism and Transhumanism" 12. Zygon. 54 (1): 156–176. doi:10.1111/zygo.12494 12. S2CID 150977852 12.
- Витсорович, Кауорродов Павел; Gennadievna, Gorbacheva Anna (2017). "ПРИМЕНИМОСТЬ ПАРАДОКСА НЬЮКОМА ДЛЯ РАЗРЕШЕНИЯ ПРОБЛЕМЫ "ВАСИЛИСКА" РОКО" 12 [Applicability of the Newcom Paradox for Solving the Roko's "Basilisk" Problem]. Modern Research of Social Problems (in Russian). 9 (6): 29–33. ISSN 2077-1770 12. Retrieved 30 April 2021.

External links

- Original post 12

Last edited 22 hours ago by 47.213.228.203

RELATED ARTICLES

Eliezer Yudkowsky
American AI researcher and writer (born 1979)

LessWrong
Rationality-focused community blog

Flesh Without Blood
2015 single by Grimes

Wikipedia
Content is available under CC BY-SA 4.0 12 unless otherwise noted.
Terms of Use • Privacy policy • Desktop