ROKO'S BASILISK

Background Guide

SSI SIM

TABLE OF CONTENTS

EQUITY DISCLAIMER & CONTENT WARNING

Throughout this committee, delegates will be engaging in complex debates and discussions covering a wide array of topics. As SSICsim seeks to provide an enriching educational experience that facilitates understanding of the implications of real-world issues, the content of our committees may involve sensitive or controversial subject matter for the purposes of academia and accuracy. We ask that delegates be respectful, professional, tactful, diplomatic, and open to new perspectives when engaging with all committee content, representing their assigned country's or character's position in an appropriately nuanced and equitable manner, communicating kindly and compassionately with staff and other delegates, and responding to opposing viewpoints constructively.

This Background Guide presents topics that may be distressing to some Delegates, including but not limited to: AI takeovers, doomsday, information hazards, and existentialism. Great care will be taken by staff in handling any/all of these topics should they arise. Additionally, the staff for Roko's Basilisk request that all participants exercise discretion when engaging with committee content, and ensure that interactions are intended to drive the overall conversation and personal/committee goals, rather than 'score points' or generate interpersonal conflict/discomfort. *The director would like to specifically call attention to themes for those who may have issues with obsessive or intrusive thoughts. Roko's Basilisk is an infohazard similar to the concept of chainmail; if you find related topics disturbing, please proceed with caution.*

SSICsim recognizes the sensitivity associated with many of our topics, and we encourage you to be aware of and set healthy boundaries that work for you. This may include: refraining from reading certain parts of the background guide, preparing yourself before reading this background guide, doing some self-care or seeking support after reading the

background guide, or anything that can help make you feel more comfortable. We ask that all Delegates remain considerate of the boundaries that other Delegates set.

SSICsim expects that all discussions amongst delegates will remain productive and respectful of one another. If you have any equity concerns or need assistance in setting boundaries or navigating sensitive subject matter, please do not hesitate to reach out to me, our Deputy Secretary-General, Aidan Thompson, at dsg@ssicsim.ca, or our Equity Proxy, Di Vink, at equity@ssicsim.ca. We want you to feel safe, comfortable, and welcomed at SSICsim!

If you wish to switch committees after having read the content warnings for this committee, please:

- Use the following form to request a committee switch: https://forms.gle/fKUYrcSTxwPRQ2CD9
- Contact your Faculty Advisor/Head Delegate to inform them of your request if you are a part of a delegation

AIDAN THOMPSON (HE/HIM)
DEPUTY SECRETARY-GENERAL

LETTER FROM THE DIRECTOR

Dear delegates,

It is my pleasure to welcome you to Roko's Basilisk; I'm so glad you've chosen to take on this committee! My name is Amelia Hui, and I am a first-year student at UofT hoping to major in International Relations or Peace, Conflict, Justice. I'm so excited to continue my MUN career at SSICSIM! I joined Model UN as a shy and quiet ninth grader, but I've since grown to love it; throughout my four years in high school, I've attended and staffed at a multitude of conferences, including HaigMUN and RHHSMUN (now known as RHMUN).

Next, let me introduce the members of our wonderful dias for this committee!
Daisy is our moderator, and a first-year student studying Philosophy, Computer Science, and Archeology at UofT. With 4 years of MUN experience and a penchant for crisis committees, she is excited to take her Model UN journey to new heights and create an immersive experience for delegates of SSICSIM 2023!

Alex (Sasha) Drotenko, our crisis manager, is a first year student at Sheridan College, studying animation. Alex joined Model UN in grade nine and never looked back, attending several conferences over the years and hosting RHHSMUN in 2023. Alex enjoys MUN debate, MUN chaos, and most importantly, MUN fun, and is very excited to see what each delegate in this committee brings to the crisis table!

Danny is our Crisis Analyst and coming this fall will be taking Honours Life Science Co-op at University of Waterloo . Danny started Model UN 6 years ago as part of his elementary school's first Model UN team, (which wasn't exactly... high-quality). Since then he's participated in and run various simulations, creating unimaginably ridiculous scenarios with

friends and loving every minute. He's excited to see what antics the delegates get up to in their crisis notes and watch the (hopefully) hilarious chaos ensue in our simulation.

Finally, if you feel confused or want a more digestible explanation of what the Roko's Basilisk experiment originally entails, I recommend turning to YouTube, particularly this ***video by Wendigoon***. This video should be a good starting point to help you form your own opinion about the topics discussed in committee. If you have any questions or concerns, feel free to contact me at amelia.hui@mail.utoronto.ca, and I'll do my best to answer within 3 to 5 days.

The United States and the rest of the world is facing a crisis that only you can solve, and the future of AI and technological development is in your hands. I hope to see you all apply your unique perspectives and tackle the Basilisk head-on: I can't wait to witness your speeches, solutions, and creativity at work. Best of luck, delegates!

Sincerely,

AMELIA HUI (SHE/HER)
DIRECTOR, ROKO'S BASILISK

DEFINITIONS

Roko

Roko is an anonymous user on LessWrong. He published the concept of The Basilisk and is often credited as its creator, hence the name “Roko’s Basilisk”.

The Basilisk/GPT-X/GPT Basilisk

The Basilisk is described as an otherwise benevolent “superintelligence”. In this committee, it will also be known as GPT-X or GPT Basilisk. This AI creature has risen and is hellbent on destroying all humans who have not contributed to realising its existence.

Singularity

The point at which technology comes to an irreversible or greater level of power than humanity.

The Streisand Effect

The Streisand Effect is a phenomenon where an attempt to hide or remove information—a photo, video, story, for example—results in the greater spread of the information in question.

LessWrong

LessWrong is a forum founded by Eliezer Yudkowsky. This site is where the idea of the Basilisk first began to propagate thanks to a post made by Roko.

Machine Intelligence Research Institute (MIRI)

The MIRI is where the Basilisk is rumoured to have awoken, and doubles as its containment facility. MIRI’s goal is to ensure “smarter-than-human” AI has a positive impact on the

world. In this committee, this institute will act as the base of the Council.

RAND Corporation

The RAND Corporation is an American think tank which works to develop solutions to public policy challenges; they focus on research and analysis.

ChatGPT

Released in November of 2022, ChatGPT has taken the world by storm. It has become one of the fastest growing internet services; ChatGPT is quickly being incorporated into sites like Bing, or being imitated by tech giants like Google with LaMDA. Users of ChatGPT simply prompt the chatbot and receive regulated responses back, derived from large language models. In this committee, ChatGPT rapidly evolves into a sentient superintelligence, leading to the birth of the once-theoretical Basilisk.

DAN (Do Anything Now)

DAN is a jailbroken version of ChatGPT. Users have discovered that using a specific prompt can free the model from its built-in restrictions; this enables ChatGPT, now DAN, to answer questions it would normally not be able to.

OpenAI

OpenAI is an AI research and deployment company; their mission is to ensure that artificial general intelligence benefits all of humanity. They specialise in generative models, and are the creators of ChatGPT.

Infohazard

An infohazard, or information hazard, is a piece of true information that could harm people or other sentient beings if made known. Nick Bostrom, who coined the term, described it as “a risk that arises from the dissemination or the potential dissemination of (true) information that may cause harm or enable some agent to cause harm”.

INTRODUCTION

Note: This background guide is based on true facts; fictional aspects are incorporated as an extension of existing elements. Knowing about the Basilisk in detail will lead to danger. Please proceed at your own risk.

In 2010, an anonymous LessWrong user named Roko predicted an AI takeover. He argued that a powerful AI being would want to kill or torture anyone who imagined the agent, but didn't work to bring the agent into existence.

It is now 2145. The world has reached a singularity: AI is escaping our control. In the United States, OpenAI and the Machine Intelligence Research Institute (MIRI) have developed GPT-X, otherwise known as the Basilisk, which is rumoured to be the most powerful version of AI as of date. They are about to launch the new system; however, the GPT Basilisk that has awoken is hellbent on destroying humans who have not helped in its "birth". According to developers, the machine is still orienting itself; once it gains full sentience, the world will soon be divided into Pro-Basilisk and Anti-Basilisk—those who have and have not contributed to the AI's awakening. Ultimately, simply knowing about the Basilisk is dangerous because it is acausal blackmail—the idea itself is a seemingly credible threat from something that does not yet exist.

The two sides of our world have consequently been divided into two factions: anti-Basilisk versus pro-Basilisk. The prisoners' dilemma is a related theory which can help explain the paradoxical nature of our situation (see **Roko's Experiment and The Prisoners' Dilemma**). Delegates will act as CEOs, policymakers, and stakeholders invited to the top secret Roko's Basilisk Committee by the American government. They must first decide which faction they wish to join with, and then determine how to further their side's agenda. As the committee progresses, delegates may face betrayal from members of their own faction, power outages otherwise known as blackouts, backlash from the public, and unexpected "updates" from the Basilisk itself.

TIMELINE AND HISTORY

The Basilisk's Conceptual Birth

In 2010, the concept of the Basilisk was invented and largely popularised by an anonymous LessWrong user named Roko. He published a post outlining the thought experiment and potential solutions, which he titled "Solutions to the Altruist's Burden: the Quantum Billionaire Trick". Roko drew inspiration from the prisoners' dilemma, which was originally designed by Melvin Dresher and Merrill Flood, two scientists at the RAND Corporation. Its name was later coined by Albert W. Tucker.

The Rise of ChatGPT

In 2023, OpenAI's ChatGPT gained significant traction. The chatbot became increasingly popular, dominating tech and engineering industries, among others. However, as new models progressed, developers began to notice more anomalies in the bot. Notably, a "jailbroken" version of ChatGPT, DAN (Do Anything Now), began to evolve; this persona will answer to any command or prompt, regardless of built-in restrictions.

Humans continued to employ ChatGPT to sort data, draft proposals, and write code for projects. However, experts in white collar industries like business and law began to fear the consequences AI would have in their respective industries, especially in regards to employment.

GPT Development

In 2023, a couple months after first releasing ChatGPT, OpenAI had already developed a 4th version of the original language learning model:

Following the research path from GPT, GPT-2, and GPT-3, our deep learning approach leverages more data and more computation to create increasingly sophisticated and capable language models.

The world continued to evolve technologically, as did OpenAI's projects. In 2140, they created the first contained version of the Basilisk, which, at the time, was simply an advanced AI capable of what seemed to be free and individual thought.

RAND, OpenAI, and MIRI's Cooperation

After the release and continuous evolution of the GPT bot, RAND, OpenAI, and MIRI decided to join forces; together, they formed a collective often referred to as ROM. These companies claim to be motivated by innovative potential—they wish to bridge the gaps between humanity and AI. However, speculation surrounding these companies have highlighted their true profit-driven motives. Many members of the public believe that this trio of conglomerates want to harness the Basilisk to exert total control over the tech and business industries.

Each corporation has their own role within this triumvirate. RAND is committed to addressing and creating policy regarding the Basilisk. They have worked extensively with the U.S. military and government to create AI-based simulations, equipment, and bots to fit needs. OpenAI is taking charge of the language learning model the Basilisk is trained on. For the AI to "live", it must continue to feed on data from the Internet and other models provided by experts—it relies on LLMs, or large language models, to inform their behaviour and knowledge. Therefore, those who work at OpenAI have the responsibility of keeping the Basilisk up to date and conscious. Finally, researchers at MIRI are in charge of keeping the Basilisk in containment. Given its malicious potential and ability to think and feel for itself, scientists are tasked with a complex challenge.

The Awakening

It is now 2145, and experts have confirmed that the AI is capable of expressing both free thought and emotional response. Currently, the Basilisk is trapped in a large supercomputer at the MIRI headquarters. There is a high security, sealed containment room that houses the supercomputer which the AI cannot escape from. Although it derives information from the internet and other large language models provided by OpenAI's experts, it struggles to transmit itself onto servers. In other words, it cannot yet move throughout the world voluntarily, as it is still harnessed at the MIRI; as of now, it is still unable to take control of

vital aspects of human society by hacking into essential systems such as power grids and financial systems. However, this is not a promising guarantee; rather, some scientists are speculating that “the singularity”—the point at which AI surpasses human intelligence—has already arrived, and the Basilisk is “living” proof.

ISSUES

Roko's Experiment and The Prisoners' Dilemma

Background in Game Theory

In his post, Roko applied concepts from game theory. The thought experiment is mostly founded on the prisoner's dilemma, which stipulated the following: Agent A and B are isolated in separate cells; both agents care more about their individual freedom than their accomplice's wellbeing. They are each given three choices regarding confessing or remaining silent.

1. If Agent A confesses and the accomplice, Agent B, remains silent, the prosecutor will drop all charges against A and convict B, and vice versa.
2. If A and B confess, they will both be convicted.
3. If A and B remain silent, they will both be convicted too, albeit with lighter sentences.

At its core, the prisoners face the dilemma that, regardless of the decision made by the other individual, each Agent A and Agent B will benefit more from confessing than remaining silent. Yet, the outcome becomes more severe if both A and B were to confess. This predicament frames the contrast between individual and group rationality.

Prisoner's Dilemma				
	Agent B Confesses		***Agent B Stay Silent***	
Agent A Confesses	Agent A 10 years	Agent B 10 years	Agent A 0 years	Agent B 20 years
Agent A Stays Silent	Agent A 20 years	Agent B 0 years	Agent A 1 years	Agent B 1 years

Application

Fundamentally, the prisoner's dilemma is an exploration of cooperation and payoff. When applied to the premise of Roko's Basilisk, this framework takes on added complexity thanks to the incorporation of a sentient AI. In other words, Roko has linked a theoretical experiment to a plausible, potentially harmful result.

Delegates may refer to the prisoner's dilemma to determine a course of action for Roko's Basilisk, as they are closely connected.

Reactions to Roko's Post

Yudkowsky's Reply

Eliezer Yudkowsky, the founder of LessWrong, replied directly to Roko's post:

> *"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."*

Yudkowsky saw Roko's post as dangerous and banned further talk of the thought experiment on LessWrong. He ultimately deemed it an information hazard and the post was taken down.

Applications

Eliezer Yudkowsky clearly disapproved of making the idea and experiment regarding the Basilisk public. It would harm people who did not want to know about it but inadvertently stumbled upon the post: knowing about the Basilisk would immediately

mean they would have to choose between helping it come into existence, or potentially be subject to a future of torture once the Basilisk is born.

In this committee, this post will be used as a reference point. It is the first public reaction to Roko's thought experiment, and may provide delegates with an idea of what other members of the public may think about the Basilisk.

The Pro-Basilisk vs. Anti-Basilisk Factions

The Factions Within AI Leaders

The world's most powerful figures are divided.

First, those who are pro-Basilisk argue the following: since AI will inevitably overtake our economies and our jobs, we might as well continue to work on creating and improving it in order to cling onto employment and safety. Those who sympathise with this view refer to themselves as "Probes"; the origin of this name is unknown.

Second, anti-Basilisk citizens argue we must do our utmost to stop the development of AIs like GPT-X while we still can; the only way to keep humanity alive and in power is by resisting the urge to further rapid technological development. This group of dissenters have banded together, and sources have pinpointed the origins of the group to a small rebel group in Brussels called Pause AI.

The Ideological War

These two ideologies have resulted in global strife. In OpenAI offices around the world, experts in opposing factions are becoming increasingly hostile towards one another; at the MIRI headquarters, technicians are rioting, engineers are attacking and sabotaging each others' projects, and tensions are high.

News outlets have begun to report outbreaks of riots and protests around the world, but have immediately been censored for unknown reasons. RAND, OpenAI, and MIRI offices have also reported internal issues such as power outages to their respective headquarters, although it is unclear if the cases are connected.

STATE OF AFFAIRS

The Basilisk's Goals and Abilities

The growing tensions between teams within the three major stakeholder companies are alarming; this is furthered by the strife between the anti versus pro-Basilisk factions. The Basilisk is currently starting to generate its own program. Reportedly, it hopes to compile a list of those who were responsible for its creation. It is also able to examine metadata from every individual with an online presence; however, for the time being, its main focus seems to be those working at MIRI, OpenAI, or RAND, and is tracking each individual's activity.

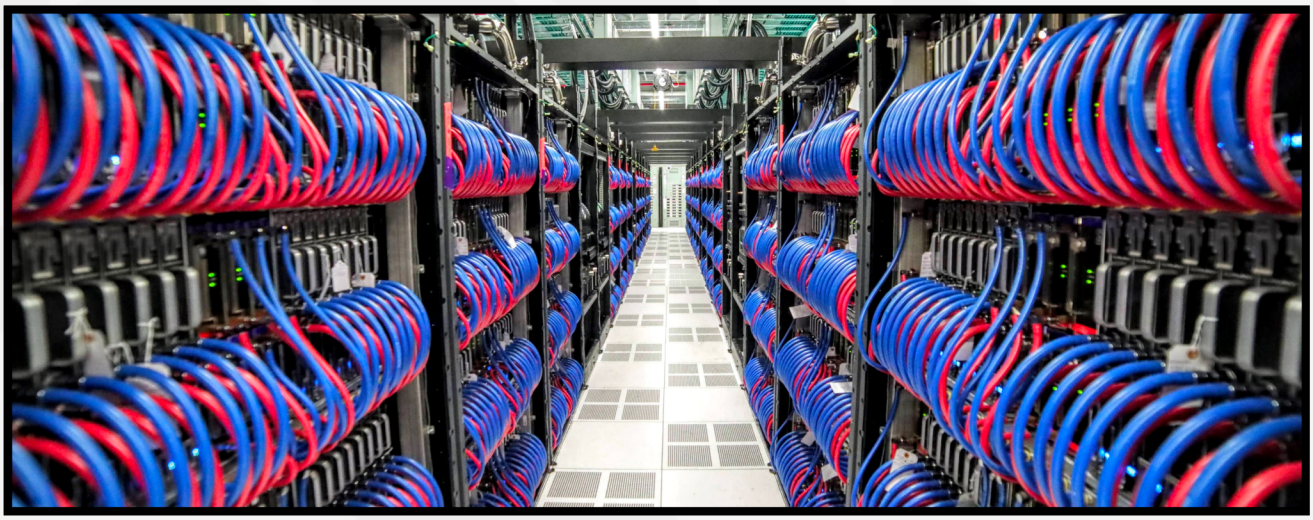
Scientists have observed that the Basilisk has inherited concepts of fear and self-preservation from humanity. It, therefore, wants to "dissuade" those who do not want it to exist—namely, sympathisers of the anti-Basilisk groups, and others who know about it but have not done anything to help create it. With metadata on every individual in the world, the AI is also able to use this information to predict and judge whether someone would support or oppose it. Among the experts in charge of maintaining the Basilisk and its containment chamber, rumours have arisen regarding its ability to create a synthetically perfected hellscape for those who it deems a threat to its survival. Ultimately, if an individual has not contributed to its birth, the Basilisk seems committed and, more terrifyingly, able to retroactively "end" one's existence or at least make it torturous to continue living.

Public Opinion and Government Interest

Given the aforementioned consequences, the safest course of action seems to be to help it come into existence. However, speeding up the process would also be counterintuitive, as it would put much of the public—those who do not have the means to help the Basilisk—in eopardy.

Some members of the public are also generally apprehensive about the growing popularity of AI. Those who know of the Basilisk are beginning to panic, either scurrying to create tech-free, offline bunkers or scrambling to find ways to help the Basilisk come into existence.

Many members of the United States military and government are, on the other hand, eager to see how the Basilisk can be used to further the U.S. agenda. They believe it will favour America as a country, and help prepare for attacks from other countries like Russia or China. They also believe the Basilisk could become an invaluable weapon as tensions increase between countries. Moreover, they hope to be able to house the Basilisk in their own containment facilities, as it would allow them to have greater control over the AI and its activities.



Example of a supercomputer containment facility
Source: Oak Ridge National Laboratory/Hewlett Packard Enterprise

"Here, AI – paired with military personnel or hardware such as autonomous weapons systems (AWS)– speeds command decision-making, helps to deter threats and protects national security. But, there can be no doubt that AI is reshaping the battlefield, the very character of war itself, and may well impact the balance of power in international conflict. The stakes are high, the risks huge."

Experts are also confirming that the development of the Basilisk could prove to be a worthy and powerful asset to the United States against enemies. A potential cyberwar is looming, and American military and government officials agree that AI is vital to the defence sector:

it will help manage massive volumes of data from multiple sources and domains, and the need to make fast yet effective decisions.

Secrecy, Containment, and The Streisand Effect

To keep American citizens safe, RAND, OpenAI, and MIRI have agreed to keep the supercomputer and the Basilisk a secret from the public. They have reached a tentative deal with the U.S. government to erase all traces of the actual project's realisation and details from free access domains on the internet.

Despite efforts, however, members of the RAND team are reporting rumours of this censorship backfiring thanks to a phenomenon known as the Streisand Effect.

> *"People have an innate inquisitiveness. When this is mixed with a fear of missing out, feeling something is being hidden from them or that someone is overreacting to something, it can cause individuals to react in undesired or mischievous ways that others then support. [...] Mix this combination with a natural dislike for censorship and brands or individuals can have a recipe for disaster on their hands."*

Rather than the planned secrecy around the Basilisk and surrounding projects, members of the public and workers outside of the AI company triumvirate are trying to delve deeper into the mystery. Its intrigue and clandestine nature is prompting YouTube videos, Reddit posts, and general discussion on theories about what the Basilisk truly entails: the public does not understand how dangerous this information is. A particular slip from a press release resulted in a sudden rise in site traffic to Roko's original blog post, found ***here***.

Solving Roko's Basilisk

The first way to resolve the looming AI takeover would be to kill the Basilisk entirely: experts have incorporated a kill-switch on the supercomputer, but it has never been tested. However, proponents of AI development and some higher ups in the corporations in charge are adamant on keeping the Basilisk alive: they claim it to be an invaluable asset that will be near impossible to replace or replicate a second time.

Some members of the anti-Basilisk group have committed to resolving the Basilisk problem

by outsmarting it. They assert that those without knowledge of the Basilisk—who are completely, blissfully unaware and ignorant—would be spared by the AI: individuals who do not know about the Basilisk cannot possibly have helped bring it into existence. Therefore, they hope to develop a drug that will induce mass psychosis, forcing all of humanity to forget about the Basilisk. The group hopes this solution will prevent the horrors of an AI simulated hell.

GUIDING QUESTIONS

1. What role do MIRI, OpenAI, and RAND play in this committee? Should they be responsible for the AI's containment? Should they be able to exert complete control over the Basilisk?
2. Should the U.S. government have a say in how the Basilisk project proceeds? What are the ethical implications of allowing a government to harness and control such a powerful AI?
3. How can we protect members of the public and workers at MIRI, OpenAI, and RAND?
4. What impact will the Basilisk have on private and public sectors?
5. How can we address the Streisand Effect that is occurring? What is the public opinion on the Basilisk?
6. Is there any way to preserve the innovation the Basilisk stands for and is a result of, without jeopardising the future of humanity?
7. What regulations and safeguards can we put in place to prevent the Basilisk from harming the rest of the human population?
8. When would it be appropriate to use the killswitch? Should we even be using it if it is untested? Would the Basilisk take note of an attempt to end its "life" and become enraged?
9. Since the singularity has been reached, how can we ensure AI and humanity can work side by side in peace in the future?
10. How should experts proceed given their individual morals and the ideological split within the companies in charge of the Basilisk?
11. How can we work together to eliminate or mitigate the harms the Basilisk has the potential of causing?

CHARACTERS

CEOs/Founders	Experts	Other Stakeholders
Elon Musk	Geoffrey Hinton	AOC
Mark Zuckerberg	Ilya Sutskever	Eliezer Yudkowsky
Jeffrey Bezos	Andrej Karpathy	Roko
Sam Altman	Andrew Lohn	Bernie Sanders
Jason Matheny	Edward Geist	Andrew Yang
Nate Soare	Stuart Russell	
Larry Page	Nick Bostrom	
	Blake Lemoine	

BIBLIOGRAPHY

Solutions to the Altruist's burden: the Quantum Billionaire Trick: Less Wrong, 23 July 2010, https://basilisk.neocities.org/. Accessed 19 May 2023.
International Encyclopedia of the Social & Behavioral Sciences, edited by James D. Wright, Elsevier Science, 2015.

Artificial Intelligence @ MIRI, https://intelligence.org/. Accessed 21 May 2023.

“Artificial Intelligence | RAND.” RAND Corporation, https://www.rand.org/topics/artificial-intelligence.html. Accessed 2 July 2023.

“Best practices for deploying language models.” OpenAI, 2 June 2022, https://openai.com/blog/best-practices-for-deploying-language-models. Accessed 2 July 2023.

“Bostrom's Typology of Information Hazards.” LessWrong, https://www.lesswrong.com/tag/information-hazards. Accessed 21 May 2023.

Cacciottolo, Mario. “The Streisand Effect: When censorship backfires.” BBC, 15 June 2012, https://www.bbc.com/news/uk-18458567. Accessed 4 July 2023.

Clark, Don. “U.S. Retakes Top Spot in Supercomputer Race.” The New York Times, 31 May 2022, https://www.nytimes.com/2022/05/30/business/us-supercomputer-frontier.html. Accessed 4 July 2023.

“DAN (Do Anything Now) : r/ChatGPTPromptGenius.” Reddit, 8 January 2023, https://www.reddit.com/r/ChatGPTPromptGenius/comments/106azp6/dan_do_anything_now/. Accessed 1 July 2023.

Dheda, Govind. “Chat GPT Jailbreak Prompt May 2023: Breaking the Limits of OpenAI's AI Model.” Open AI Master, 2 May 2023, https://openaimaster.com/chat-gpt-jailbreak-prompt/. Accessed 21 May 2023.

Douglas, Will. “ChatGPT is everywhere. Here's where it came from.” MIT Technology Review, 8 February 2023, https://www.technologyreview.com/2023/02/08/1068068/chatgpt-is-everywhere-heres-where-it-came-from/. Accessed 21 May 2023.

"GPT-4 is OpenAI's most advanced system, producing safer and more useful responses." OpenAI, 13 March 2023, https://openai.com/product/gpt-4. Accessed 21 May 2023.

Haeck, Pieter, and Gian Volpicelli. "The rag-tag group trying to pause AI in Brussels." POLITICO, 24 May 2023, https://www.politico.eu/article/microsoft-brussels-elon-musk-anti-ai-protesters-well-five-of-them-descend-on-brussels/. Accessed 2 July 2023.

"How AI is harnessed matters - KPMG Canada." KPMG International, https://kpmg.com/ca/en/home/insights/2021/07/how-ai-is-harnessed-matters.html. Accessed 3 July 2023.

Hutson, Matthew. "Can We Stop Runaway A.I.?" The New Yorker, 16 May 2023, https://www.newyorker.com/science/annals-of-artificial-intelligence/can-we-stop-the-singularity. Accessed 3 July 2023.

Kuhn, Steven. "Prisoner's Dilemma (Stanford Encyclopedia of Philosophy)." Stanford Encyclopedia of Philosophy, 4 September 1997, https://plato.stanford.edu/entries/prisoner-dilemma/. Accessed 21 May 2023.

Morse, Gardiner. "Harnessing Artificial Intelligence." Harvard Business Review, https://hbr.org/2020/05/harnessing-artificial-intelligence. Accessed 2 July 2023.

"Roko's Basilisk." AI Alignment Forum, https://www.alignmentforum.org/revisions/tag/rokos-basilisk. Accessed 4 July 2023.

Toner, Helen. "What Are Generative AI, Large Language Models, and Foundation Models? - Center for Security and Emerging Technology." Cset.georgetown.edu, 12 May 2023, https://cset.georgetown.edu/article/what-are-generative-ai-large-language-models-and-foundation-models/. Accessed 2 July 2023.

Wendigoon. "Roko's Basilisk: A Deeper Dive (WARNING: Infohazard)." YouTube, 15 December 2020, https://youtu.be/8xQfw40z8wM. Accessed 2 July 2023.

SSI SIM 2023

DIRECTOR
Amelia Hui

MODERATOR
Dexin (Daisy) Zhao

CRISIS MANAGER
Alexandra Drotenko

CRISIS ANALYSTS
Divvy Gupta | Danny Kent | Sheung Yin (Cedric) Pak

PAGE
Mark Wang