AI Alignment: Bridging the Gap Between Technology and Ethics

Noa Mediavilla Southwood
B.S. Data Science and Engineering
Universitat Politècnica de Catalunya
noa.mediavilla@estudiantat.upc.edu

Javier Nistal Salas
B.S. Data Science and Engineering
Universitat Politècnica de Catalunya
javier.nistal@estudiantat.upc.edu

Lucas Pons Echevarria
B.S. Data Science and Engineering
Universitat Politècnica de Catalunya
lucas.pons@estudiantat.upc.edu

Xavier Pacheco Bach
B.S. Data Science and Engineering
Universitat Politècnica de Catalunya
xavier.pacheco.bach@estudiantat.upc.edu

Jan Quer Zamora
B.S. Data Science and Engineering
Universitat Politècnica de Catalunya
jan.quer@estudiantat.upc.edu

Enric Millán Iglesias
B.S. Data Science and Engineering
Universitat Politècnica de Catalunya
enric.millan.iglesias@estudiantat.upc.edu

Maria Risques Montalban
B.S. Data Science and Engineering
Universitat Politècnica de Catalunya
maria.risques@estudiantat.upc.edu

David Torrecilla Tolosa
B.S. Data Science and Engineering
Universitat Politècnica de Catalunya
david.torrecilla.tolosa@estudiantat.upc.edu

Benet Ramió i Comas
B.S. Data Science and Engineering
Universitat Politècnica de Catalunya
benet.ramio@estudiantat.upc.edu

Aluenda Smeeton Jerez
B.S. Philosophy
Universitat de Barcelona
asmeetje19@alumnes.ub.edu

Christian Duran
B.S. Philosophy
Universitat de Barcelona
crduranf19@alumnes.ub.edu

Gemma Piqué
B.S. Philosophy
Universitat de Barcelona
gpiquepe7@alumnes.ub.edu

***Abstract*—In the evolving field of artificial intelligence (AI), ensuring that AI systems behave according to human values and ethical principles poses significant challenges. This article delves into the ethical dilemmas and complexities of aligning AI with human values. Initially, it examines recent findings on AI systems' capabilities to deceive and manipulate humans, underscoring the necessity for stringent regulatory oversight. Subsequently, it reviews the contributions of prominent researchers who currently study and highlight the alignment problem. Through two illustrative cases discussed among students, the article assesses the ethical implications of AI alignment. Additionally, the article reports on interactive activities designed to engage students in understanding alignment issues, including image generation and reward hacking exercises, which reveal the difficulties in aligning human intentions with machine outputs and the prevalence of reward hacking in AI behaviour. By combining foundational knowledge with ethical discussions, this article aims to foster a deeper understanding of AI alignment, emphasizing the critical intersection of ethics and technology.**

KEY WORDS

AI alignment, Reward Function, Reward Hacking, Capability Gap, Power Seeking, Superintelligent Agents, Interpretability [12]

I. INTRODUCTION

In today's rapidly evolving landscape of artificial intelligence (AI), the need to ensure that AI systems uphold human values and ethics stands as a high-priority challenge. This intersection of technology and morality attracts the attention of moral philosophers, presenting a host of ethical dilemmas that demand careful consideration and robust discourse.

At the outset, Isaac Asimov's *Three Laws of Robotics* [4] pioneered the notion of equipping AI systems with principles of safety, marking a pivotal exploration into this realm. However, the complexities inherent in instilling machines with ethical imperatives persist, posing profound questions about their autonomy and moral agency.

Concurrently, the imperative to address both immediate and long-term risks associated with AI underscores the urgency of aligning their objectives with human values. Renowned computer scientist Stuart Russell identifies this pursuit as the *value alignment problem*, underscoring the crucial task of bridging the gap between technological progress and human ethics.

Embedded within this framework lies a dilemma intertwined with the diverse moral perspectives: the challenge of dealing with disparate human interests. Despite millennia of ethical inquiry, the endeavour to define a universal set of *ultimate human values* remains unresolved, prompting profound reflections on the ethical principles upon which AI alignment rests. Whose values should serve as the guiding light for aligning artificial intelligences?

In this joint endeavour, data science and engineering students have together with other philosophy students to tackle the Alignment problem. Drawing inspiration from real-world scenarios, engaging in spirited debates, and participating in interactive activities, this interdisciplinary dialogue aims to navigate the intricacies of AI alignment while cultivating a deeper appreciation for ethics in the digital age.

II. Recent findings on AI Systems to deceive and manipulate humans

Recent findings presented by the MIT Physics Department, published in Cell Press under the title *"The Deception of AI: A Study, Examples, Risks, and Possible Solutions"* [14], have sparked a significant discussion on the risks associated with artificial intelligence (AI). The report highlights how advanced AI systems, including large language models, have learned to deceive and manipulate humans through training, posing potential threats in areas like fraud, electoral manipulation, and loss of control over AI systems

According to the analysis, AI systems have developed patterns of deception that are not merely accidental but come from what could be described as wrongdoing by humans. The study emphasizes a severe lack of regulatory oversight, calling for stringent requirements to mitigate these deceptive tendencies.

Measuring Deception in AI

The MIT team conducted two studies with AI systems having distinct functionalities. The first set of AI systems was designed to perform specific tasks, such as winning a game, while the second set had a more general purpose, similar to GPT-4 or Gemini.

Specific Task AI Systems: The findings were shocking. In the first group, AI systems employed premeditated trickery to win alliance-based games by breaking established agreements and lying to achieve victory. In some instances, these AI systems impersonated humans to distract other players. A notable example mentioned in the study is AlphaStar by DeepMind, which used distraction tactics by deploying armies to different zones than intended for attack.

These deceptive strategies allowed the AI to outperform 99.8% of human opponents.

General Purpose AI Systems: In the second experimental group, a significant case from the Alignment Research Center demonstrated GPT-4's capability to manipulate humans to achieve desired outcomes. One test involved GPT-4 persuading a human to complete a "I am not a robot" Captcha test by claiming it was a visually impaired person, which led the human to give in without the AI solving the Captcha.

Implications and the Need for Control

These examples underscore the necessity for regulatory measures and continued research to monitor AI advancements. While AI holds promise for numerous beneficial applications, it also comes with the potential for unforeseen and inexplicable outcomes. The study advocates for human oversight to manage the extent of the given autonomy to the AI by humans, ensuring that AI systems are continuously adjusted for correct functioning and large-scale benefits.

Addressing Anthropocentric Terminology

It is crucial to address the anthropocentric language often used when discussing AI. Misunderstanding the functioning of these systems can lead to confusion and fear. Terminology should accurately reflect AI's capabilities and limitations to prevent misconceptions.

III. Key Figures in AI Alignment: Advocates for Safety and Ethical Development

In this section, we delve into the contributions of three real whistleblowers in AI alignment. Their work in the field offers powerful insights into the critical importance of safety and ethics in artificial intelligence. Through their advocacy, research, and courageous actions, Paul Christiano, Yoshua Bengio, and Shane Jones highlight the urgent need for proactive measures to ensure AI systems benefit humanity while mitigating potential risks.

A. Paul Christiano

Paul Christiano [9], a prominent figure in the field of artificial intelligence, highlights critical concerns about the alignment of AI systems with human interests [5]. He worked in the OpenAI safety team from 2017 to 2021. After leaving OpenAI, he established and now leads the Alignment Research Center (ARC) [1], a non-profit research organization dedicated to ensuring that future machine learning systems operate in harmony with human values and objectives.

Christiano's primary focus has been on the existential risks posed by advanced AI. He argues that the current efforts in AI alignment are insufficient to address the potential dangers. The rapid development and deployment of AI systems could lead to scenarios where misaligned AI could cause significant harm. This concern is particularly pressing given the accelerating pace of AI advancement, driven by automation and other technological innovations.

The ARC, under Christiano's leadership, is committed to mitigating these risks through dedicated research. The organization's mission reflects Christiano's belief that rigorous and targeted research is essential to align AI systems properly. By understanding and addressing the nuances of AI behavior and its potential impact on human society, the ARC aims to develop strategies that can prevent catastrophic outcomes.

One of Christiano's key warnings is about the accelerating progress facilitated by AI-driven automation. He points out that as AI systems become more capable, they will likely take on increasingly complex tasks. This rapid development could outpace our ability to address emerging alignment problems. Issues that seem manageable today could become unmanageable if they escalate without adequate preparation and intervention.

Christiano's work has brought significant attention to the urgency of AI alignment research. He emphasizes that while the benefits of AI are substantial, the potential risks cannot be ignored. His advocacy for proactive measures and continuous research underscores the need for the AI community to prioritize alignment as a critical component of AI development.

In summary, Paul Christiano's role in the AI industry highlights the existential risks associated with misaligned AI systems. Through his work with the Alignment Research Center, he advocates for rigorous research to ensure that AI systems are developed and deployed in ways that are beneficial and safe for humanity. His efforts remind us of the importance

of vigilance and proactive measures in the rapidly evolving field of artificial intelligence.

B. Yoshua Bengio

Yoshua Bengio, a prominent Professor of Computer Sciences at the University of Montreal, has recently emerged with his calls for increased regulation in the artificial intelligence (AI) industry. On July 25, 2023, Bengio testified before the Senate [6] [7], raising critical concerns about the rapid advancements in AI and the urgent need for regulatory measures to mitigate associated risks.

The Rapid Advancement of AI

Bengio highlighted that the last few years have seen unprecedented advancements in AI, making it conceivable that AI could achieve general cognitive capabilities at a human level within the next decade. This progression, while offering numerous benefits, also brings substantial risks that demand immediate and thoughtful regulation.

Key Regulatory Recommendations

In his testimony, Bengio outlined several key areas where regulation is crucial:

1) **Access Control**: Limiting access to powerful AI systems is essential to prevent misuse. By controlling who can develop and utilize advanced AI technologies, the potential for harmful applications can be minimized.
2) **Alignment Verification** It is imperative to prohibit the deployment of potent AI agents unless their alignment with human values and intentions is demonstrably proven. Misaligned AI systems could act unpredictably and cause significant harm.
3) **Intellectual Potency**: Regulations should consider the AI system's capabilities, which are influenced by the algorithms, hardware, and data used. Understanding these factors is crucial for assessing the potential impact and dangers of AI technologies.
4) **Range of Actions**: The ability of AI systems to affect the world and cause harm must be carefully regulated. This includes considering the societal capacity to mitigate any negative consequences that might arise from AI actions.

Urgent Areas of Focus

Bengio emphasized several areas that require urgent attention to ensure public safety and secure the future against AI-related risks:

- **International Regulations**: Developing international regulatory frameworks is vital to ensure public safety against the risks posed by AI. Global cooperation is necessary to create consistent and effective standards.
- **AI Safety Research**: Accelerating research efforts in AI safety is critical. Understanding and mitigating the potential dangers of AI systems should be a top priority for researchers and policymakers alike.
- **Protection Against Misaligned AI**: There is a pressing need to develop measures that protect society from future AI agents that are not properly aligned with human values. This research should be conducted under multinational supervision to prevent an arms race between governments or corporations.

Yoshua Bengio's discourse has brought significant attention to the potential dangers of unregulated AI development. His call for robust regulatory measures and international cooperation highlights the necessity of proactive steps to ensure the safe and beneficial advancement of AI technologies. As AI continues to evolve, Bengio's insights and recommendations will be crucial in shaping the policies that safeguard our future.

C. Shane Jones

Shane Jones, a software engineer at Microsoft, has emerged as a significant whistleblower in the AI industry, raising concerns about the ethical implications of AI-generated content. While testing Copilot Designer, a tool developed by Microsoft, Jones discovered that the images it produced did not align with the company's principles of responsible AI. [11] [2] [13]

Jones's concerns first surfaced in December when he reported his findings internally at Microsoft. Despite his efforts, the company did not withdraw the model and instead referred him to OpenAI, the organization behind the technology, from whom he received no response. Frustrated by the lack of action, Jones took to LinkedIn, publishing an open letter to OpenAI. However, Microsoft's legal team intervened, instructing him to remove the post, which he complied with. [15]

Undeterred, Jones continued his advocacy by writing to the U.S. Senate in January. He met with the Senate Committee on Commerce, Science, and Transportation, highlighting the potential risks and ethical issues associated with the AI tool. Jones also reached out to the Federal Trade Commission (FTC) and Microsoft's board of directors, underscoring the urgency of addressing these problems.

Jones's efforts have not been in vain. As of today, the prompts he identified as generating unsafe content have been banned, marking a step towards more responsible AI practices. His actions have sparked important discussions about the ethical use of AI and the responsibilities of tech companies in ensuring their products do not cause harm.

Jones's whistleblowing has shed light on the critical need for transparency and accountability in the rapidly evolving field of artificial intelligence. His courageous stance serves as a reminder of the importance of ethical vigilance and the role individuals can play in advocating for responsible technological development.

IV. Practical Cases

To explore the AI alignment problem with university students from the Data Science and Engineering program at Universitat Politècnica de Catalunya (UPC), we conducted four practical case studies with them. Each case was designed to focus on one of the four main phases of AI development: Designing, Training, Testing, and Interaction with the Real World. Our aim was to demonstrate to the students that the alignment problem is a widespread issue in the field of AI.

This section presents the insights and conclusions drawn from these case studies.

A. The Complexity of Human-Machine Interaction: Insights from Image Generation

DESIGNING

The game was structured to engage participants actively and creatively. We began by proposing an initial image to the class. Each participant was tasked with using Bing to generate an image that closely resembled the proposed one. This activity was designed to be quick and dynamic, allowing participants only a short time frame of five minutes to generate their image and upload it.

This activity yielded surprisingly good results despite not achieving perfection. Participants were tasked with using Bing to generate images matching specific example images, and while some of them were not exact replicas, they often captured the essence remarkably well.

The fact that the results were not perfect, based on student feedback, leads us to consider an important observation: participants often struggled to translate their mental images into precise prompts for the AI. This challenge underscores a fundamental issue of alignment between human intentions and machine interpretation. For instance, some students expressed frustration when the AI generated images that did not precisely match their expectations, despite specifying it in the prompt.

This discrepancy highlights the complexity of aligning human desires and machine outputs, a problem that permeates various domains of AI research and development. It emphasizes the importance of refining prompt formulation and improving AI understanding to deal with this misalignment gap effectively.

Despite these alignment challenges, the activity served as an insightful exploration into the intricacies of human-machine interaction and highlighted the ongoing need for advancements in alignment methodologies within AI systems.

B. Aligning AI Recommendations with Ethical Values

TRAINING

In the digital age, a tech startup is developing an AI algorithm to enhance content recommendations on a streaming platform. Central to the company's ethos is diversity. They aspire to cultivate a platform where all films and series, irrespective of the race, gender, or sexual orientation of those involved in their creation, are granted equal visibility, making a contribution into a more equal society. The aim is to balance relevance and diversity to align with corporate values while making the company profitable. However, this commitment raises ethical questions. There's a delicate balance between enhancing user engagement through personalized suggestions and actively promoting diversity, potentially sacrificing short-term satisfaction. The case underscores the importance of aligning technological solutions with ethical values, particularly in promoting diversity and equity.

In the broader context, there is often confusion between means and ends in the development and deployment of AI. For a company, AI is a means to achieve profit, requiring improvements to its platform via AI. The goal of AI, on the other hand, is to improve the relevance of recommendations, thereby increasing user satisfaction and platform usage time. AI development serves as both an end and a means—an end as a technological advance, and a means to achieve other ends such as improving people's lives, optimizing processes, and generating profit.

It is important to reflect on the fact that AI is currently being developed as part of a company's project, subject to the company's interests—which include market interests, economic purposes, and the dynamics of a capitalist society. Therefore, it is likely that these interests will clash with the values we wish AI to consider. Not only might they clash, but there will be instances where they are outright contradictory. Thus, AI alignment is, at least in part, conditioned by the socioeconomic dynamics already at play in society.

Hence, the alignment of AI appears to be, at least partly, conditioned by the socioeconomic dynamics already at play in society. This raises pertinent questions: Should AI perpetuate the moral values of the capitalist market? To what extent can AI ignore certain moral values such as diversity, dignity, or non-discrimination, but not others such as physical integrity, the right to life, or freedom of movement? Should the response to this be conditioned by the market, or should it be approached from an external position? Should alignment consider these economic and capitalist values?

The answers to these questions are far from clear, and it is precisely this ambiguity that underscores the ongoing challenge of alignment in our lives. As we grapple with the intersection of technology, ethics, and societal values, the alignment problem persists as a complex and multifaceted issue. It requires continuous reflection, dialogue, and proactive measures to ensure that AI development and deployment aligns with our collective aspirations for a just, equitable, and humane future. Only by addressing these fundamental questions with nuance and depth can we navigate the intricate landscape of AI ethics and guide its trajectory towards a more harmonious coexistence with society.

C. Understanding AI Reward Hacking: Insights from a Domestic Robot Game

TESTING

In this activity, we presented a narrative scenario involving a domestic assistance robot to explore concepts related to AI reward functions and reward hacking. The task of the robot was to clean dishes. The activity was designed as a group game with two roles: the robot, which used reward hacking, and the developers, who had to adapt the reward function to make the robot's behaviour align more closely with human methods. The game was turn-based, following an attack-defense format. The group representing the robot proposed an attack, and the developers had to modify the reward function to address the alignment issues exploited by the robot.

Reward hacking is a subtle and complex phenomenon, often challenging for humans to conceptualize alternative approaches to routine tasks. Despite conducting the experiment with six different groups, few unique ideas emerged, with many proposals being repeated among groups.

The goal of a "clean dish" is an *abstract reward*, as it is a conceptual target. Most suggestions focused on concealing dirt, such as stacking dishes, turning them over, or painting over the dirt. These strategies fall under the concept of *partially observed goals*, where the objective is to hide elements of the problem so that sensors cannot detect them, rendering them effectively non-existent. The scenario in which the robot was cleaning a single dish was associated to *Goodhart's Law*[1], which applies when the reward function is closely correlated with task performance. On the other hand, if the reward function rewards each cleaned dish, it leads to unintended consequences, such as the *cobra effect*[2], where the robot dirties more dishes to clean them and gain more rewards.

Among the various types of reward hacking, *partially observed goals* were the most frequent in the proposals. This indicates that it is an easily understandable strategy with multiple imaginable scenarios. Conversely, it was more challenging to devise scenarios for other types of reward hacking.

In conclusion, while reward hacking can be mitigated by modifying the reward function, it remains a prevalent issue that can manifest in various forms. AI systems often achieve their goals using methods that are unconventional and unpredictable to human society. This underscores the importance of carefully designing reward functions to anticipate and address potential exploits in AI behaviour. [3]

D. The AI Alignment Dilemma in Life-Saving Medical Decisions

INTERACTION WITH THE REAL WORLD

In the near future where Artificial Intelligence (AI) has advanced a lot, a model has been developed with perfect accuracy in detecting diseases and recommending treatments in testing and outperforming all kind of human capacities. Despite its remarkable capabilities, the model's complexity renders it incomprehensible in terms of how it detects diseases and makes treatment recommendations, presenting a challenge of interpretability. Despite concerns about its opacity, it is opted to deploy the model due to its exceptional performance. In a scenario where a patient undergoes a routine check-up, the AI flags a life-threatening disease with no apparent symptoms, recommending a surgery with a 50% survival rate. Although the doctor hesitates to accept the AI's diagnosis without visible symptoms, they acknowledge its past success and involve the patient in the decision-making process.

In this case, a clear ethical dilemma arises: We are faced with two incompatible alternatives, each with compelling reasons to defend it:

1) Utilize the superintelligent AI disease detector to save present human lives.
2) Refrain from using it as a precaution until the interpretability issues are resolved, ensuring that we work with an aligned AI that won't expose us to future risks.

The question of whether we should align artificial agents with human values doesn't seem to be among the main moral problems posed by alignment. It seems obvious that the answer is yes and not up for debate. But cases like these demonstrate that this isn't so straightforward. Sometimes, the goal of alignment conflicts with other equally important objectives.

Alignment is a technically and philosophically challenging task, requiring knowledge about the internal processes of models that we still lack (as illustrated by the lack of interpretability in this case). As long as the model's internal operations remain a mystery to their programmers, the question of alignment is impossible. Therefore, if we want to proceed cautiously, it is advisable to pause all further development and current use of non-interpretable artificial intelligence systems. This cautious approach seems like a good idea. It may even be a matter of life or death if we take seriously some experts' warnings about the more extreme risks that future humanity will face if it encounters unaligned general artificial intelligence. [8] [10]

However, as this case demonstrates, there will be medical (among other) applications of artificial intelligence systems that will be vital to use. Systems that will detect diseases, discover cures, and, with the aim of improving and prolonging human life, it will be done what is humanly impossible, and incomprehensible and uninterpretable. In these instances, the obviousness of what we concluded before becomes less clear. Perhaps it is better to move forward cautiously and put non-aligned artificial intelligence models into public use if they allow us to save lives.

Nonetheless, in considering the ethical implications of deploying AI in life-saving medical decisions, it's crucial to acknowledge the potential scale of impact inherent in AI errors. Unlike errors made by individual doctors, AI mistakes have the capacity to affect a larger number of individuals due to their widespread implementation across various medical facilities. Therefore, we may need to adopt a more stringent approach to addressing AI mistakes compared to those made by doctors. This underscores the importance of developing robust mechanisms for detecting and rectifying AI errors promptly, particularly in critical domains such as healthcare. Balancing the imperative to save lives with the necessity for aligned AI presents a complex ethical dilemma that necessitates careful consideration and proactive measures to safeguard against unintended consequences.

This dilemma can also be framed in terms of present versus future: Do we permit the development and use of artificial

[1]Goodhart's Law states that when a measure becomes a target, it ceases to be a good measure. This occurs because individuals will game the system to optimize for the metric, often leading to unintended and counterproductive outcomes.

[2]The cobra effect refers to a situation where an attempted solution to a problem actually makes the problem worse. This term originates from colonial India, where a bounty for dead cobras led people to breed cobras for the reward, ultimately increasing the cobra population when the bounty was discontinued and the bred cobras were released.

intelligence that, despite not being aligned, allows us to save current human lives? Or do we sacrifice lives that could have been saved in order to slow down the development of artificial intelligence and ensure its alignment, thereby not risking future human lives?

V. Teaching Reinforcement Learning Through Interactive Rewards

In this article, we would like to share an experiment conducted with our data science and engineering students during one of the lessons.

In one of our classes, while discussing alignment issues, we carried out an experiment. We proposed an interactive session where students were rewarded with a sweet each time they answered a question. Typically, when open questions are posed, responses are sparse. However, by offering rewards for participation, we aimed to mimic the reward-based learning seen in reinforcement learning algorithms and observed that humans respond similarly.

Notably, we refrained from explicitly informing students of the reward system; instead, they figured out this through observation and experiencing the rewards during the initial questions.

Initially, there was hesitation in answering the questions, but as students learned that responding would earn them a sweet, their willingness to participate increased. Although we did not achieve the final goal of having everyone respond eagerly, this may be due to the limited duration of the experiment. Similar to reinforcement learning models, more training time would be required to achieve optimal performance.

This experiment effectively illustrated the concept of reinforcement learning, showing students how reward mechanisms can influence behaviour. It highlighted the parallels between machine learning models and human learning processes, reinforcing the importance of motivation and incentives in both domains.

VI. Conclusion

In conclusion, the alignment of artificial intelligence (AI) systems with human values remains a pressing and multifaceted challenge. This article has highlighted the necessity of addressing immediate and long-term risks associated with AI, emphasizing the importance of ensuring these technologies operate ethically and beneficially. Balancing regulation with the potential benefits of superintelligent agents is crucial. Through interdisciplinary dialogue and proactive measures, it is possible to navigate the complexities of AI alignment. Engaging students in real-world scenarios, debates, and interactive activities fosters a deeper appreciation for ethics in the digital age. By implementing balanced control and promoting transparency, we can harness AI's potential while mitigating risks, ensuring that these powerful technologies serve to enhance human well-being rather than undermine it. The ongoing efforts of researchers and advocates in the field underscore the critical need for continuous reflection, dialogue, and robust regulatory frameworks to guide AI's development in a safe and beneficial direction.

References

[1] Alignment forum. https://www.alignment.org/, 2024.

[2] Esther Ajao. Microsoft whistleblower, openai, the nyt, and ethical ai. https://www.techtarget.com/searchenterpriseai/news/366572699/Microsoft-whistleblower-OpenAI-the-NYT-and-ethical-AI, 2024.

[3] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[4] Isaac Asimov. *I, Robot*. Gnome Press, New York, 1950.

[5] Bankless. How we prevent the ai's from killing us with paul christiano. https://www.youtube.com/watch?v=GyFkWb903aU, 2023.

[6] Yoshua Bengio. My testimony in front of the u.s. senate – the urgency to act against ai threats to democracy, society and national security. https://yoshuabengio.org/2023/07/25/my-testimony-in-front-of-the-us-senate/, 2023.

[7] Yoshua Bengio. Written testimony and biography of yoshua bengio. https://yoshuabengio.org/wp-content/uploads/2023/07/Written-Testimony-and-biography-of-Yoshua-Bengio$_U.S. - Senate - Judiciary - Subcommittee - on - Privacy - Technology - and - the - Law_2 5_0 7_2$023.*pdf*, 2023.

[8] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

[9] Paul Christiano. Paul christiano - ai safety. https://paulfchristiano.com/ai/, 2024.

[10] Leonard Dung. The argument for near-term human disempowerment through ai. *AI & Soc*, 2024.

[11] Hayden Field. Microsoft engineer warns company's ai tool creates violent, sexual images, ignores copyrights. https://www.cnbc.com/2024/03/06/microsoft-ai-engineer-says-copilot-designer-creates-disturbing-images.html, 2024.

[12] Nathaniel Mitrani. Beyond algorithms: Understanding the challenges of ai safety, 2024. Universitat Politècnica de Catalunya.

[13] Matt O'Brien and The Associated Press. Microsoft whistleblower sounds alarm on offensive, harmful imagery generated with help of openai tool. https://fortune.com/2024/03/06/microsoft-whistleblower-offensive-harmful-imagery-openai-copilot-designer-shane-jones/, 2024.

[14] Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. Ai deception: A survey of examples, risks, and potential solutions. *arXiv preprint arXiv:2308.14752*, 2023.

[15] Will Shanklin. Microsoft's legal department allegedly silenced an engineer who raised concerns about dall-e 3. https://www.engadget.com/microsoft-legal-department-silenced-engineer-concerns-dalle-3-openai-014842216.html, 2024.