University of Tartu
Institute of Computer Science
Conversion Master in IT

Evelin Jõemaa

Alignment and safety challenges in a superintelligent AI landscape

Master's Thesis (15 ECTS)

Supervisor:
Grace Eden, PhD

Tartu 2025

Alignment and safety challenges in a superintelligent AI landscape

Abstract:

Artificial intelligence (AI) technologies are developing at a rapid pace, and with that, the discussions on AI safety have gained importance. This thesis aims to explore the concerns that members of online AI alignment communities have in regards to AI safety and alignment and compare them to the opinions of the key figures in the AI safety and alignment field.

Five interviews were conducted via online conferencing platforms with participants recruited from online AI alignment communities. The interviews were on the topics of AI safety, AI alignment, and ethical concerns of AI development. Additionally, five interviews with key figures in the AI safety and alignment field were analysed from interviews available on YouTube for comparison with participant interviews. Interviews were transcribed, and thematic analysis was conducted to identify key themes.

The findings show a significant concern that the participants have regarding current safety measures. The concerns were often related to the rapid speed of the advancements in AI technology and the shortcomings they saw with the current safety measures in being able to handle the developments. Participants saw the unintended consequences of AI development being a bigger risk in the future, but also highlighted the already present risks of current AI models, such as the ability for people to create deep-fakes, etc. The community forum participants highlighted the need for more collaboration between private companies and governments to have better measures put in place internationally for developing safe AI models.

Keywords: Artificial intelligence, AI alignment, AI safety

CERCS: P176 Artificial intelligence

Joondumise ja ohutuse väljakutsed superintelligentse tehisaru maastikul

Lühikokkuvõte:

Tehisaru (TA) süsteemid arenevad suure kiirusega ning sellega koos on tõusnud arutelude olulisus TA ohutuse teemal. Magistritöö eesmärk on uurida veebipõhiste TA joondumise ja ohutuse kogukondade vaateid TA joondumise ja ohutuse kohta ning võrrelda neid sama valdkonna võtmeisikute arvamustega.

Magistritöö raames viidi videokonverentsiplatvormide kaudu läbi viis intervjuud osalejatega, keda värvati veebipõhistest TA joondumise ning ohutuse kogukondadest. Intervjuud käsitlesid TA ohutust ja joondumist ning eetilisi murekohti TA arendamisel. Lisaks analüüsiti viite intervjuud TA joondumise ja ohutuse valdkonna võtmeisikutega, mis olid kättesaadavad YouTube kaudu, et neid võrrelda magistritöö raames korraldatud intervjuudega. Intervjuud transkribeeriti ning olulisemate teemade väljaselgitamiseks kasutati temaatilist analüüsi.

Tulemused näitavad, et intervjuus osalejad tundsid palju muret praeguste TA ohutusmeetmete pärast. Mured olid tihti seotud TA kiire arenguga ning praeguste ohutusmeetmete piiratusega, et kiire arenguga kaasas käia. Osalejad leidsid, et ootamatud tagajärjed TA arendamisel on tulevikus suurem oht, kuid rõhutasid ka juba praeguseid riske, mida TA süsteemid endast kujutavad, näiteks võimalus luua süvavõltsinguid. Osalejad rõhutasid vajadust suurema koostöö järele eraettevõtete ja valitsuste vahel, et kehtestada rahvusvahelisel tasandil paremaid meetmeid ohutute TA süsteemide arendamiseks.

Võtmesõnad: Tehisaru, tehisaru joondumine, tehisaru turvalisus

CERCS: P176 Tehisintellekt

Table of Contents

1. Introduction

What is intelligence? While commonly thought of only in the context of human intelligence or, at most, the intelligence of animals, it has now garnered a much wider context among the general public as the discussions of different types of intelligence emerge, such as extraterrestrial, microbial, and **artificial**. In this thesis, the focus is on the latter, artificial intelligence (AI) and its alignment with human interests. To begin this discussion, it is first necessary to define what intelligence is and how to evaluate the level of intelligence in both humans and non-humans.

Researchers have argued that *intelligence* can have a much wider meaning than what most people would consider. When talking about artificial intelligence, general intelligence (AGI) is often the center of discussion, that is "possessing common sense and an effective ability to learn, reason, and plan to meet complex information-processing challenges across a wide range of natural and abstract domains" (Bostrom, 2014: 3). This is a more human-centric view of intelligence. However, in astrobiology, extraterrestrial intelligence is a domain where one of the central questions is whether humans would even recognise intelligence as it could be vastly different from our understanding of what constitutes intelligence (Davies, 2010). It is difficult to define intelligence even in the context of our planet, let alone the universe. For example, the Gaia hypothesis proposed in the 1970s by chemist James Lovelock and microbiologist Lynn Margulis states that the planet Earth itself is an intelligent system of functions that self-regulates itself and can therefore be considered a separate intelligent entity (Lovelock & Margulis, 1974). While the belief in those hypotheses is up to each individual, it sparks a question of whether, in order to understand non-human intelligence, which includes AI, people must reevaluate what intelligence means once the human component is removed from it. However, it is still important to remember that AI is currently being developed by humans and is, therefore, inevitably affected by our ideas of intelligence, as it is often developed based on the structure of the human brain.

One way of developing AI is to do it on the example of the human brain. The human brain project that ran over the course of 10 years from 2013 to 2023 was an attempt to understand and map the human brain (Human Brain Project, 2023). It consisted of neuroscientists, engineers, physicists, statisticians, computer scientists and researchers from other fields with a goal of understanding the structure and function of the human mind while also looking into the ethical implications of such research (*ibid.*). Modelling computers on the example of

human brains is not new, and neural networks have been the centre of advancing AI systems with the goal to accelerate their ability to learn and adapt similarly to the human brain (LeCun *et al.,* 2015). Using the ideas on the workings of the human brain for developing AI raises the question of whether such a system could achieve true intelligence and identity of its own.

The theory of mind states that "an individual has a theory of mind if he imputes mental states to himself and others. (…) For example, purpose or intention, as well as knowledge, belief, thinking, doubt, guessing, pretending, liking, and so forth" (Premack & Woodruff, 1978: 515). In AI development, this raises a question of what a learned response that an AI would give is and where the line is between a learned response and an AI having an authentic reaction, driven by its own experiences and beliefs. Therefore, the question remains whether AI can truly replicate and surpass the human mind and how it would impact society, especially if an AI that surpasses the human mind in general intelligence, also known as superintelligence, emerges (Bostrom, 2014: vii). Various institutes and think tanks, such as the Machine Intelligence Research Institute (MIRI) and Future of Life Institute play a pivotal role in researching how AI affects society and how AI alignment, that is whether the AI system behaves in a way its creators intended, should be a central topic of discussion to ensure its safety (Tegmark, 2017).

The main research question of the thesis is:

How do narratives by well-known AI researchers contrast with the opinions and concerns of members of online AI alignment communities?

The aim is to understand the opinions and concerns of people regarding AI safety and alignment and to provide new insight into this topic by interviewing members of online AI alignment communities and contrasting it with the opinions of key figures in the AI safety and alignment field.

The thesis is divided into six main chapters, beginning with this introduction, which gives a brief background of the topic while justifying the reasoning for choosing the thesis topic. Followed by a literature review which gives an overview of the AI landscape regarding AI safety and alignment, the 'control problem', the ethical implications of developing AIs and the need for responsible development. Next, the methods section describes how the research participants were found and the methods used for gathering and analysing data from primary and secondary sources. The findings section presents the results of the research, divided into

subchapters according to the common themes found during the analysis. The discussion section interprets the key findings and highlights the contributions of this study while also giving suggestions for further research. Finally, the conclusion restates the aim of the thesis and summarises the key points.

2. Literature review

This chapter gives an overview of the AI landscape regarding AI safety and alignment, the 'control problem', the ethical implications of developing AIs, and the need for responsible development. The chapter opens with an overview of commonly discussed AI concerns, followed by a background of AI alignment, the 'control problem', and existential risks. The chapter concludes with an overview of the current approaches to AI safety by OpenAI, Google DeepMind, and Anthropic.

2.1 AI safety

There are many concerns among scientists and the general public regarding the development of AI, but some of the key concerns emphasised by Dippnall (2024) are the following:

- Data-related challenges;
- Bias and discrimination;
- Ethical concerns;
- Regulatory and legal challenges;
- Transparency.

Each concern will be discussed in turn below.

2.1.1 Data-related challenges

Data-related challenges refer to several sub-problems: the quality of data used in training AI, data integration, and data privacy (Dippnall, 2024). The first aspect, data quality, has different requirements depending on the context in which the data is used but for machine learning (ML) and AI, the main quality checkpoints are intrinsic data quality, contextual data quality, representational data quality, and data accessibility (Priestley *et al.,* 2023: 5-6). This means that data needs to be logical, accurate, and coherent on its own, used in the right context taking into account the geographical, historical, and cultural aspects from which the data was gathered, easy to use and interpret, and preferably easily available as for ML and training AI models, large amount of data is required (*ibid.*). Creating and training AI models also raises concerns about data privacy as instances have emerged where people have discovered that their data or creative works have been used in training generative AI models without their knowledge or consent (Zhang *et al.,* 2024: 94). For example, Meta has allegedly used the copyrighted data available via a website LibGen, which is known as an archive of books with

pirated content (Taylor, 2025). One might argue that this could be solved by providing access to the data that has been used in training generative AI models, but this could lead to further privacy concerns (*ibid.*: 95). However, another issue arises from the human aspect of the data, namely that the data used in training AI models often includes human biases.

2.1.2 Bias and discrimination

AI is trained on data generated by humans, and as such, the data inevitably includes the bias of the people creating it. This, in turn, might lead to AIs which are trained on biased data and therefore can show bias themselves, such as the data on arrests including the bias of the officers making the arrests, as they might have discriminatory biases towards some groups, usually minority populations (Chouldechova & Roth, 2020: 84). Discrimination can be divided into three groups: structural (how the society in general is organised), organisational (how decision-making units are structured), and interpersonal (beliefs and actions of a single individual) (Barocas *et al*., 2023: 209-211). As such, there are three levels in the society generating the data where bias is present. When developing AI using this data, especially in predictive models, those biases are carried on. An example would be making hiring decisions where judging a potential candidate based solely on their educational background can lead to a snowball effect of both advantages (for more privileged people) and disadvantages in life being carried on to other areas and stages of life (*ibid:* 213). With the potential for such a direct impact on society, it raises the question of how to ensure the ethical development of AIs.

2.1.3 Ethical concerns

AI ethics is a relatively new field with no rulebook which is considered industry standard. However, as developing AIs has a direct effect on people, it is important to consider the ethical concerns around it regarding where to draw the line between scientific discovery, efficiency, and potential impact on humanity. According to Bostrom (2014: 123-124), in search of scientific discovery, humanity risks developing an entity that far surpasses its creators while having no guarantee that it would work for the good of humanity, be it intentional or accidental. Another concern is the ethical norms themselves, namely the lack of understanding or wish to follow them by the individuals or institutions working on AI developments (Agbese *et al.,* 2023: 70-71). Additionally, while following the ethical norms and requirements can have technical and legal advantages, there is often no financial value.

As such, they might be considered less important by for-profit companies. Floridi and Cowls (2019) have proposed a unified framework of principles for responsible and ethical AI development: beneficence, non-maleficence, autonomy, justice, and explicability. However, setting any such frameworks requires large-scale collaboration between different domains and constant reevaluation to keep in step with the changes in societal norms and technological advances (Etzioni & Etzioni, 2017: 406-417). Such collaboration and attempts at regulation can take a lot of time and resources.

2.1.4 Regulatory and legal challenges

The quick advancement of AI capabilities causes regulatory and legal challenges as the development of AI is oftentimes faster than the changes in regulations and laws, which are more time-consuming. In the three proposed takeoff times listed by Bostrom (2014: 63-65), meaning the speed at which an AI model would become superintelligent, it is possible to have regulations in place for a slow takeoff of a superintelligence, as the process is slower and the regulations and laws have time to adapt. However, in the case of a moderate or fast takeoff, having up-to-date regulations reflecting the actual state of the AI is unlikely, with the moderate takeoff taking place in a matter of months or years, which would give some time to at least change or modify the current regulations. However, with a fast takeoff, the emergence of a superintelligence takes place in a matter of minutes or days, giving no time to react with any new regulations (*ibid.*). As such, it is important to have those safety measures in the form of regulations and laws in place *before* takeoff (Future of Life Institute, 2023).

2.1.5 Transparency

Transparency in AI systems is a way to build the general public's and researchers' trust in the safety of the AI systems being developed. Velez and Kim (2017) emphasise the need for interpretable ML, which would be understandable for both developers and end users. Transparency in AI development can help developers fully understand their systems, users to understand how the system works and check if it worked correctly, and society in general to understand the system's strengths and limitations, potentially overcoming the fear of the unknown as a result (Felzmann *et al*., 2019). However, even if the systems and their inner workings were fully public, vast technical knowledge would still be required to fully understand how they work; most people do not possess this kind of expertise (Ananny & Crawford, 2018).

2.2 AI alignment

The discussion of whether AI poses a risk comes down to one key topic: AI alignment. The AI alignment problem can be defined as "the challenge for AI programmers to ensure that their systems' values align with those of humans" (Mitchell, 2019: 137). Brian Christian (2020: 12-13) defines the same term by emphasizing the potential dangers of *misalignment*, stating that the AI alignment problem is an urgent topic which should seek an answer on "how to prevent such a catastrophic divergence - how to ensure that these models capture our norms and values, understand what we mean or intend, and, above all, do what we want." Although a seemingly urgent and dangerous topic, the scientific community is not in agreement amongst themselves on whether the AI alignment problem is topical and needs to be addressed immediately, something belonging to the realm of science fiction, or a potential imminent threat.

2.2.1 Control problem

The control problem refers to the issue of "how to control what the superintelligence would do" (Bostrom, 2014: vii). The fear of AI, especially superintelligence and the inability to control it, is not a new topic and has been a part of the discussion from the dawn of computers, as even Alan Turing has warned that one day machines will outperform humans (Turing, 1951). That is the control problem, the discussion on whether and how an AI could be kept under the control of humans without it going rogue (Bostrom, 2014: 127-129). The capabilities of a superintelligence would far exceed those of humans and could lead to a scenario where AI would be in charge of every aspect of humans' lives (Poddington, 2023: 436-437). Bostrom (2014: 127-144) has proposed several methods for maintaining control over a superintelligent AI in the case of its emergence, such as capability control methods and motivation selection methods, which will be introduced in turn below.

The capability control methods aim to prevent unfavourable outcomes by limiting the capabilities of the superintelligence (Bostrom, 2014: 129-131). One of the capability control methods is called the boxing method wherein the superintelligence is kept in an environment where it cannot cause harm, in a so-called box which can be physical (no physical access to anything outside the box other than through controlled channels) or informational (the information exiting the box is controlled) (*ibid.*). Physical boxes are easier to implement for many different ML architectures, but they also limit the capabilities of the superintelligence,

causing a false sense of security, and should not be used as the only safety measure (*ibid.*). The informational containment seeks to limit the interactions that the AI can have with the outside world by limiting its access to communication networks through which it could hire people to do its bidding (*ibid.*). Informational containment is similar to the stunting method, where instead of output, the data input the AI has access to is controlled and limited in an attempt to restrict its capabilities, but should superintelligence be achieved, it is not guaranteed that such restriction would provide absolute safety (*ibid.:* 135-136). The capability control method also includes the incentive method, wherein the AI is essentially rewarded for doing the "right" thing and therefore having an incentive to remain contained and under control. This, however, would entail that the AI trusts humans not to change their minds and deliver the promised rewards and that humans can tell if the actions of the AI are in their interest or if there is a nefarious goal buried in their actions (*ibid.*: 131-133). Another capability control method is the tripwire, where diagnostic tests are performed, and the system is shut down whenever a potentially dangerous activity is detected (*ibid.*: 136-137). However, with any capability control method, the weak links can always be humans, namely the programmers and the system's owners, so-called gatekeepers, that can be persuaded to give it more access that would allow for the AI to escape (*ibid.*).

Motivation selection methods aim to achieve desirable outcomes by influencing the superintelligence to do what its creators want it to do. This can be done through direct specifications of rules or values that the AI must follow, but due to the difficulty of translating the creator's intent into code, it is often not reliable and can instead have bad consequences (Bostrom, 2014*:* 139-140). Another motivation selection method is domesticity, which aims to limit the AI by creating it for very certain functions, such as a question-answering system (*ibid.:* 140-141). Indirect normativity is another option, which refers to the creation of a process for deriving the standards and having the AI follow those self-created standards (*ibid.:* 140-141). The augmentation method aims to achieve control over superintelligence by taking a system that already has a motivation system and enhancing it (*ibid.*: 141-142). This is a different path to superintelligence that does not necessarily require an AI and can instead use methods such as biological enhancements and brain emulations to achieve superintelligence, assuming that the motivation systems present before enhancements would remain intact (*ibid.*: 141-142). All those methods discussed in the control problem subsections are, in a way, an attempt to reduce or prevent an existential risk.

2.2.2 Existential risk

The opinions of computer scientists, policymakers, and the general public can be generally divided into two distinct groups (Baum, 2018: 2-5, 9):

- Those who believe that AI *does not* pose an imminent risk and sometimes argue that this topic belongs to the realm of science fiction;
- Those who believe that AI *does* pose a substantial existential risk to humanity as we know it.

The first group believes that there are more immediate harms to humanity related to AI. For instance, some argue that its massive power consumption and consequent harm to the planet should take precedence over any potential risk by future superintelligent AI (Luccioni, 2024). As Luccioni (2023) states, "Each time you query an AI model, it comes with a cost to the planet." They are not proclaiming that there is *no* potential existential threat in the future from AI. Rather, the damage that AI is already causing in the present day should be the main focus. This is echoed by Bill Gates (2023), who states that deepfakes and misinformation generated by AI could be a threat, as well as a potential cause for the loss of jobs, and the use of AI tools in studying could stunt the skills of the next generation. However, he also believes that while it is important to think about the risks of AI and to have policies in place, there is no reason to think the risks cannot be managed (*ibid*.). Some strong opponents of the belief that AI poses an imminent existential risk argue that humans are very far from being able to create an AI that is even equal to the human mind, let alone one that surpasses it (Luccioni, 2023: 287-290).

The second group views AI alignment issues as an existential threat to humanity. Russel (2019: 8, 137), for instance, is concerned that in case a superintelligent AI emerges, there are only two possible outcomes: a golden age led by a friendly AI or a catastrophe. In his 2014 book *Superintelligence: Paths, Dangers, Strategies,* Nick Bostrom discusses several potential risks of a superintelligent AI and different potential future scenarios. He analyses how the speed at which a superintelligent AI is developed affects the outcome and the risks. For example, with AI that is developed slowly, there is enough time to put policies and safeguards in place should the AI be malignant, while in a scenario where the emergence of an AI is instantaneous, there is no time to respond appropriately. Bostrom fears that humans might fail to recognise the moment when a true AI superintelligence is developed, as the idea of what

constitutes consciousness or intellect is not easily defined. To give an example, playing chess was once considered the epitome of human intelligence, but it turned out that what we would now consider a simple programme is able to play at the level of a grandmaster; as such, the definition of consciousness in AI needed to be redefined and might need to be redefined time and time again (*ibid.*: 11-14). The concerns for AI safety have led to the establishment of several research organisations whose focus is on studying existential risks and AI safety, such as the Future of Life Institute, Stanford University's Stanford Existential Risks Initiative, University of Oxford's The Future of Humanity Institute and Centre for the Study of Existential Risk at the University of Cambridge (Boddington, 2023: 435).

2.2.3 Friendly vs. unfriendly AI

The AI alignment problem thus comes down to a seemingly simple question: should (superintelligent) AI emerge, will it be **friendly** or **unfriendly**? A friendly AI can be defined as "the production of human-benefiting, non-humanharming actions in Artificial Intelligence systems that have advanced to the point of making real-world plans in pursuit of goals" (Yudkowsky, 2001: 2-3). While there can be as many definitions as there are scholars, the central idea of the definitions remains the same: friendly AI is one that does not pose a threat to humans and will act in its best interests. Yudkowzky (2001: 3) states that should humanity succeed in creating a friendly AI, the consequences can be unimaginably positive, but similarly, the opposite is true should the AI be unfriendly instead.

Unfriendly AI comes with a number of risks. Unfriendly AI is defined by Max Tegmark (2014: 318) as an AI "whose goals don't coincide with those of us humans, and whose actions end up destroying both humanity and everything we care about." However, this endpoint might be incidental and not done on purpose, as the AI might simply wish to achieve a goal and humans are in the way of that goal, similarly to how ants are to humans (*ibid.*). However, some picture an even bigger threat, the emergence of a truly 'rogue' AI that escapes its creator's control and acts autonomously with disregard for human interests (*ibid.*). In 2023, an open letter was published calling for an immediate pause to the development of powerful AI models and urging the AI research community to jointly develop safety protocols to ensure that such AI would not pose an existential risk to humanity (Future of Life Institute, 2023).

2.3 Current approaches to AI safety

AI safety is an important topic for companies developing current, well-known models. In this chapter, a brief overview is given of the AI safety approaches of the companies that have developed some of the best-known AI models currently in use: OpenAI, Google DeepMind, and Anthropic

With the public's eyes keen on the topic of AI, companies have taken an approach to have information about their views on safety and alignment readily available. OpenAI, Anthropic, and DeepMind all have an easily findable page on their website about their approaches to AI safety and alignment. OpenAI (2025) has provided its users with easy-to-find information on its stance on AI safety and alignment. The page has the quote “We believe in AI’s potential to make life better for everyone, which means making it safe for everyone” right on top (*ibid.*). They state that when teaching their models, they focus on three things to promote it being increasingly intelligent and powerful while also behaving according to the democratic values and ethical principles:

- Model training;
- System-level guardrails;
- Long-term safety (*ibid.*).

Open AI has stated that they wish to minimise biased and harmful outputs and imbue it with the principles, policies, and human values they have set to achieve that goal (*ibid.*). When it comes to the guardrails, OpenAI explains that this is an additional protection which checks “what people input and what our AI outputs”(*ibid.*). The long-term safety is important to OpenAI for developing trustworthy systems that follow human intent even when the models become more complex and agentic. In addition, they have both human and automatic evaluations in place to check if their models are in compliance with their safety policies. Before releasing new models, it is tested in four high-risk areas: cybersecurity, CBRN (Chemical, Biological, Radiological, Nuclear), persuasion, and model autonomy, with the model needing to pass all tests as not being high-risk before release. Otherwise, additional training is needed, and the model is sent back. (*ibid.*)

Anthropic (2025) says that their research teams are also working on investigating the safety aspects of AI development. Their research principles are:

- AI as a systematic science;
- Safety and scaling;
- Tools and measurements;
- Focused, collaborative research efforts (*ibid.*).

Therefore, although worded differently from OpenAI, they follow quite similar principles. Although DeepMind's (2025) page on safety is somewhat shorter, it still highlights the key thing the company is focused on: need for responsible development to mitigate negative impacts, evaluating systems for a wide spectrum of potential risks, having safeguards in place, and collaborating with experts in the industry and academia for both better development ideas and solutions as well as risk mitigation (*ibid.*). DeepMind (2025) also has teams entirely focused on ethics, technical safety, security, governance, and public engagement.

On AI alignment, OpenAI (2025) is open with their thoughts and concerns on the topic when it comes to artificial general intelligence (AGI). They approach AI alignment by trying to develop the systems in a manner that "elevates humanity and promotes democratic ideals"(*ibid.*). OpenAI admits that one of the issues is that there is no single societal or moral norm, and trying to codify everything is impossible, as many values are context-sensitive and depend on the culture where one attempts to set the norms. However, they state that this might not be an issue in some areas. Specifically, they wish to "develop tools to better adapt AI systems to diverse contexts, make informed decisions, and align with the moral and social norms of the communities they serve."(*ibid.*) This means that the AI systems could have slightly different values in different regions to better reflect the values of their users.

Anthropic (2025) considers the key importance to be having "the right safeguards in place to keep models helpful, honest, and harmless." This is because more powerful AI models in the future could be powerful enough to go beyond current safety measures. They have alignment researchers actively looking for situations where the models they are developing might behave in bad or unexpected ways to check if the safeguards are good enough. They also evaluate the models to make sure they are honest and harmless. This is done in situations that are different from the training situations. (*ibid.*) DeepMind is guided by the Google AI (2025) principles, and one of their three principles is responsible development and deployment, which includes "implementing appropriate human oversight, due diligence, and feedback mechanisms to align with user goals, social responsibility, and widely accepted principles of

international law and human rights." As such, all three companies have addressed AI alignment as a topic on which they keep an eye to ensure future AI systems are aligned with humanity.

OpenAI (2025) admits that intelligence, referring to AI, does not always mean positive impact and that when working towards developing it, it is important to also try and mitigate the risks and develop "them in accordance with human values and with humans in control." Especially since they consider the transformative impacts of AGI to already be visible within the next few years (*ibid.*). OpenAI (2025) sees three potential failures:

- Human misuse;
- Misaligned AI;
- Societal disruption.

They describe human misuse as anything a human might be using AI for to break the law or democratic values, including but not limited to censorship, propaganda, surveillance, suppression of free speech, and phishing or scams (OpenAI, 2025). For misaligned AI, OpenAI says that the more powerful the AI, the bigger the potential consequences. Societal disruption can be caused by the rapid change an AI might bring, as the effects might be unpredictable and possibly negative. One risk OpenAI sees is that if, for example, an authoritarian regime can use an AGI effectively, it might get ahead of the democratic regimes. (*ibid.*)

Community effort is an important part of developing safe AI models. OpenAI (2025) promotes collaboration within the industry and with academia, government, and the public. For example, by publishing AI safety research, providing resources and funding for research, and promoting policy initiatives. They believe that collaboration can help advance the development, as well as make it easier to detect issues if there are more people openly working in the field. (*ibid.*) This is also true for Anthropic and DeepMind. Anthropic (2025), as previously stated, has collaborative research efforts as one of its research principles. They collaborate with other researchers and labs to have a "broad community of researchers working together" (*ibid.*). DeepMind (Google AI, 2025) is similarly focused on collaboration, stating on their page that they collaborate with other research labs and co-fund a non-profit called Partnership on AI, which creates a community of academics, charities, and company labs focused on AI. DeepMind also works with other partners to address global issues with the help of AI models, such as a fight against antimicrobial resistance (*ibid.*). As such, all

three companies keep collaboration on the high ground as a way to further research and safety.

3. Method

The aim of this thesis was to analyse how narratives by well-known AI researchers contrast with the opinions and concerns of members of online AI alignment communities. To achieve this objective, the thesis used qualitative research methods to collect data on the opinions on AI safety and alignment from members of online AI alignment communities. To develop the interview topics for this thesis, a Munk debate (Policy-Relevant Science & Technology, 2023) with Yoshua Bengio, Max Tegmark, Melanie Mitchell, and Yann LeCun was used for ideas.

3.1 Primary data: Participant interviews

Primary data was collected through participant interviews. The participants were recruited from online communities focusing on AI alignment, specifically the AI Alignment community in Slack and the ControlProblem (Reddit, n.d.) community on Reddit. The thesis used convenience sampling (Etikan et al., 2016) to recruit participants. Convenience sampling allows for the selection of participants who are readily available, willing to participate, and possess relevant knowledge on AI safety and alignment (*ibid.*). Participants were recruited for this study after establishing contact through the aforementioned online communities. This was done by making public posts inviting people to participate after confirming that such posts were allowed in the community guidelines. A total of six participants were initially recruited. However, due to unforeseen circumstances, one participant had to withdraw their participation, and five participants were interviewed for this research. Information about the participants can be found in Table 1 below.

Participant ID	**Connection to AI**	**Region**	**Recruitment platform**
Participant A	Independent researcher.	North-America	AI Alignment Slack
Participant B	Working in AI solutions for companies, safety and alignment are personal interests.	South-America	Reddit
Participant C	Using various AI tools in software development, AI safety is a personal interest.	Northern-Europe	Reddit
Participant D	Personal interest.	Western-Europe	Reddit
Participant E	Working on a degree in AI and computer science.	North-America	AI Alignment Slack

Table 1. Participant information.

Semi-structured interviews were conducted from February to April 2025 on online conferencing platforms Google Meet and Discord to understand their views on the topics (Archibalt et al., 2019). All interviews were recorded and later transcribed using the transcription tool available on Microsoft Word. The general topics of the questions (see Appendix 1) were the following:

- Perspectives on the current state of AI safety and alignment research;
- Opinions on whether superintelligent AI poses a genuine existential risk;
- Ethical concerns about developing AI;
- The role of private companies versus public institutions in managing AI safety.

Each interview was planned for approximately 60 minutes. The participants were informed that they could terminate the interview at any time and have all their data deleted and not used in the study. All interviews were conducted with participants who are adults and who signed a consent form. The transcribed interviews were uploaded to ATLAS.ti for thematic analysis.

3.2 Secondary data: Interviews with key figures

Secondary data, meaning data which was not collected by the author of this thesis, nor for the purpose of the thesis, was also analysed (Sindin, 2017). Five interviews available via YouTube with key figures in the international AI safety and alignment discourse were included in the analysis with the following people: Max Tegmark (DW News, 2022), Yoshua Bengio (Machine Learning Street Talk, 2025), Geoffrey Hinton (CBS Mornings, 2023), Yann LeCun (Dr Brian Keating, 2024), and Melanie Mitchell (Scripps Research, 2023). Including secondary data in the analysis provided access to the opinions of prominent figures in the AI alignment public discourse. These were contrasted with the primary data collected from the participant interviews. Similarly to participant interviews, the interviews were transcribed with the transcription tool available in Microsoft Word and then uploaded to ATLAS.ti, which was used to conduct thematic analysis.

3.3 Thematic analysis

A thematic analysis was conducted to find patterns in qualitative datasets through coding, which is a process of labelling “relevant meanings in data, in relation to a research question (Braun and Clarke, 2022: 22,135). The six phases of thematic analysis were used in this

research: familiarising oneself with the data, generating initial codes, searching for and reviewing themes, identifying and naming themes, and then producing the report(*ibid.*). Qualitative data analysis was conducted using ATLAS.ti. The thesis took a semantic approach to the thematic analysis, that is, the "themes are identified within the explicit or surface meanings of the data" and the organised data was interpreted to discuss the "significance of the patterns" (Braun and Clarke, 2006). Sections of the transcripts were highlighted, and a code was added to describe the contents of the highlighted section. Codes were refined as needed during repeated read-throughs to better describe the contents of the highlighted sections. The focus of the analysis was to identify themes associated with the views of the interviewees on AI safety and alignment. The themes found in the data are discussed in the Findings chapter of this thesis.

4. Findings

This chapter provides an overview of the findings of the thematic analysis of the interview data gathered from February-April 2025 from five participants. In addition, the findings include input from five key figures in the AI alignment space in the form of publicly available interviews on YouTube in which they have discussed topics relevant to this thesis. The key figures are Max Tegmark, Yoshua Bengio, Geoffrey Hinton, Yann LeCun, and Melanie Mitchell. The goal of this chapter is to provide an overview of prominent themes in the interview data and see how the opinions of the interview participants and the key figures align or differ. Note: participant quotes are marked as "(Participant n)" while the key figures in AI are marked by their last names "(lastname)".The findings chapter consists of five subchapters for the following themes: AI safety, AI alignment, sentience, intentional misuse and unintended consequences, and who controls AI. The participants of the interviews conducted for this thesis will be referred to as participants A-E, and the key figures will be referred to by their last names.

4.1 AI safety

AI safety was a prevalent concern of the interview participants. This subchapter gives an overview of three AI safety topics discussed by the interview participants and the key figures: the emergence of AGI, current safety regulations, and stopping or slowing down the development of AI.

4.1.1 Emergence of AGI

A key topic of the interview was the likelihood of an AGI or superintelligent AI emerging in the first place and the safety concerns surrounding it. When asked about the possibility of a superintelligent AI emerging one day from Participant A, he said the following:

> "I think that there is some chance that the world will coordinate to choose not to build it through regulation and treaties, but unless that happens, yes, I think it's extremely likely, 95%. (…) I expect it to go from human to smarter than human pretty quickly and to much smarter than human within, I don't know, let's say 3 or 4 years. It feels a lot harder to predict, though." (Participant A)

Two participants had a longer timeline in mind:

> "If I had to guess, I think somewhere between years and decades. Not in the next two or three years, but also not thirty years from now. I think sooner. So maybe to put a more exact timing, I would say somewhere along the ten-year line, give or take a couple." (Participant B)

> "My best guess would be somewhere in the fifteen to twenty-year range. I don't think we are anywhere near there yet, but a couple of decades from now, I think we could have something very powerful." (Participant C)

The emergence of a superintelligent AI was therefore not something out of science fiction for the interview participants. Rather, concrete opinions were given about the potential timeline for its emergence, varying from a couple of years to twenty years. As seen, there is some uncertainty regarding the timelines as the current developments are moving at a fast pace.

On the topic of timelines in AI safety, Bengio said the following:

> "We need to think ahead of these possibilities, even if it's 20 years from now. (…) Think how much time it took to sign the nuclear non-proliferation treaties in the 60s, and the negotiations started right after the end of World War 2. So that's almost 20 years. And that's kind of the timeline where I would say there's a high probability that we figure out AGI, very high probability." (Bengio)

Bengio emphasised the need for thinking ahead of time when it comes to AI safety, as treaties (or other legislation) can take a long time. With his prediction that an AGI could emerge somewhere around the twenty-year mark, he considers it important to already start the work on better safety protocols now.

In general, both the interview participants and the key figures seem to agree that AGI or a superintelligent AI is not likely to emerge in the next few years. However, the interview participants seemed to give a slightly smaller timeline for its emergence, staying closer to the 10-15 year range while Bengio and Hinton both lean more towards the 20-year mark. As such, no matter the opinions on the potential timeline, there is a need to have regulations in place to be prepared for more powerful AI models, be it five years from now or twenty.

4.1.2 Current safety regulation

Safety regulations were a concern for all five interview participants. When asked about whether current safety measures are sufficient, participant C and participant E said the following:

> "Not even close. They're patchy at best. The industry is moving so fast that any regulations we put in place are outdated basically as soon as they come into effect." (Participant C)

> "I think at this time in our history, the most important topic in AI safety is really effective AI governance, passing laws and regulations that would not only address current modern day harm, but also be proactive and able to react to a fast takeoff or some kind of takeoff where AI capability accelerates rapidly." (Participant E)

Participant D was similarly concerned:

> "I think current measures on AI safety are very weak. Partly because it's an arms race. Essentially, there is very little in it for anyone who gets AI second." (Participant D)

All five participants were in agreement that current AI safety regulations are not sufficient. Participant B also echoed the sentiment of Participant D, that AI development right now is an "arms race". The overall satisfaction with the current measures was low, with the low speed at which AI safety regulations come into effect being the main concern, along with the competitive pressure. This could demonstrate that, given how the current AI development seems to be moving, the regulatory frameworks need to be able to adapt to the changes quickly. Additionally, more collaboration, both international and between private companies, could be an advantage, so there would be less competition to be the first.

This idea of an "arms race" is also something Bengio mentioned in his interview:

> "It's a plausible scenario which has a flavour of winner takes all. Which the companies are well aware of, and so that's one reason why they're racing. If they thought that being second would be good enough, then there wouldn't be as much pressure, but they all think that it's a winner-takes-all game." (Bengio)

There is concern with the safety measures and the speed at which AI is developed among both the interview participants as well as the key figures. They both agree that the

competition between companies (and countries) could pose a risk, as this could mean that less time is spent on various safety regulations if there is a risk that it would slow down development. For this reason, some have proposed that AI development should be slowed down or stopped entirely until proper safety regulations can be put into place.

4.1.3 Stopping or slowing down development

Interview participants were not very convinced that pausing or stopping AI development is realistic.

> "I mean, a pause seems very achievable if there (...) was a political will. Slowing down or stopping on a more permanent basis. Would that work? Technically. Probably not indefinitely, but I think certainly things could be slowed quite a bit." (Participant A)

> "I think it's unrealistic to pause. It's even more unrealistic to stop AI development. There are a lot of problems in what counts as AI development. You have these companies that are building AI models that are generally capable, that are agentic, and they seem to have an explicit interest to produce superintelligence. You also have these research institutes that just try to make better classification models that can do a better job predicting whether you are at risk of getting breast cancer. There is really no misalignment risk in the latter case. That is completely ridiculous, and you can even say it kills people to pause that kind of AI development. I'm a big fan of moral regulation, and regulation definitely slows down AI development. And so, yeah, I would be in support of that." (Participant E)

Among interview participants, all agreed that slowing down to have some proper safety measures in place is not a bad idea, but there needs to be a good definition of what would count as AI development. This is because stopping some developments, such as classification models for cancer research, could even mean that the stopping would kill people. It was not considered realistic to fully stop the developments, but a pause seemed reasonable. However, there was also concern that it would be the countries and companies not cooperating with slowing down that could be the issue, as those would likely be the countries and companies who would be less interested in international cooperation. As such, slowing down or stopping is only realistic if there were somehow a way to make everyone cooperate.

Participant D, however, brought up another side of the argument, namely that there could also be a lot of humans suffering that could be avoided if AI development were instead sped up, because that technology could also be used for great benefit:

> "And so there is a pretty big blind spot in scientific ethics of at what point is it unethical to not use something? And AGI is like that, but probably on a bigger scale, because being able to assist our intellectual efforts with machine intelligences that can be specialised in ways that are different to us, it can potentially rocket us into an age of much better prosperity. And all the time that is delayed, that is causing monumental amounts of harm to befall humanity, because this technology isn't being released." (Participant D)

While there is a lot of talk about risks surrounding AI, Participant D thought deeply about the potential benefits as well. While certainly a less common thought amongst interview participants, it does demonstrate that it is important to weigh the pros and cons of the speed of development.

Slowing down the development of AI systems is a topic that Tegmark, as one of the creators of the "Pause Giant AI Experiments: An Open Letter", spoke about as well:

> "Personally, I was very much hoping that this letter would help slow down, mainstream the conversation about whether we needed this actually for the first time, and slow down the most risky research. And the fact that, for example, Professor Yoshua Bengio signed the letter, he's signatory number one there, who is one of the godfathers of deep learning, this kind of AI that's given us all this, the fact that he signed it and so many other AI researchers with him, really mainstreams this. Nobody can anymore say that it's just clueless Luddites who want to slow down. It's founders of the field who want to slow down, who totally understand the wonderful upside that AI can have, that it can help us cure cancer and so much more, but realise that we're only gonna get that upside if we can make sure that the regulation and the wisdom of which we steer this technology towards good uses keeps pace with the tech progress." (Tegmark)

Similarly to Participant D, Tegmark is also discussing the pros and cons. On one hand, he says that even the founders of the field are in favour of slowing down the development. However, while doing that, he is also acknowledging the upsides that fast development can

have, with the potential to even help humans create a cure for cancer. It is therefore evident that the question of slowing down AI development is not black and white. There is a balance that needs to be achieved, one where enough time is taken to ensure proper safety protocols while also not hindering the development to a point where the potential benefits suffer.

4.2 AI alignment

AI alignment, while closely related to AI safety, is treated as a separate theme in this thesis to allow for a deeper insight into topics of AI alignment in particular. AI alignment is concerned with the question of how to keep an AI aligned with the goals of its creator and humanity as a whole. Misalignment is a situation where an AI pursues goals that are different from the ones its creator wanted it to pursue (Dung, 2023: 138), the opposite of AI alignment.

4.2.1 Likelihood of achieving AI alignment

Interview participants were all concerned with AI alignment to a certain degree, with some participants being more optimistic while others referred to the existential risks that could come from a misaligned AI. When asked about the likelihood of an AI being aligned with human goals, Participant B and Participant A stated:

> "I would go fifty-fifty. I don't think there is really a way to know. I can have an opinion, of course, and say that it is not likely, but I don't have any evidence. Or like something to back up my opinion." (Participant B)

> "With current levels of investment into safety and alignment, I would maybe give something like 20% chance or 25% chance that that succeeds." (Participant A)

While they did not give exact percentages, Participants C and E agreed as well that achieving AI alignment is incredibly difficult, while Participant D thought that it all depends on the route that AI development will take in the near future. Hence, it can be said that all participants consider AI alignment something that has a big chance of failure, stemming from the difficulty of putting exactly into words just what we want the AI to do.

On the same topic, Bengio stated the following when asked about the current state of AI alignment:

> "Insufficient. Well, we don't have clear answers about how we can build machines that will not harm people. Either by addressing this alignment problem. The alignment

> problem is what I was talking about, that there's a mismatch between what we would like the machine to do and what it is mathematically trying to do." (Bengio)

This sentiment is echoed by Tegmark and Hinton as well, that there is currently no good way to solve the alignment problem. Mitchell, on the other hand, does not believe that to be the main concern:

> "I'm not scared of like superintelligent AI getting out of control and destroying humanity. I think there's a lot of reasons why that's extremely unlikely." (Mitchell)

This shows that on this topic, even amongst key figures of the field, there are contrasting opinions, with some believing AI alignment to be a key problem while others would like to focus on more tangible dangers. While there were no participants who considered AI alignment not to be a problem, as opposed to the key figures, this was to be expected, as the participants were all recruited from online communities dealing with AI alignment.

4.2.2 Being in the way

Misalignment is not only considered dangerous due to rogue agents. Participants considered scenarios where humans could just be an inconvenience to an advanced AI.

> "Unless we are somehow able to engineer in like a deep value of caring about humanity, then, almost regardless of what their goals are, it'll be easier to achieve those goals without humans in the way. (…) We're not like we don't care about them, but we're not gonna take care of them if they're in the way of our house." (Participant A)

> "You know it's like people and ants. We don't mean to kill ants and other wildlife if we are building things, but we, well, we probably do, but the building companies just don't care. I would reckon it would be the same with us and a superintelligent AI, right?" (Participant B)

This brings out the discussion point that even if a misaligned AI is not necessarily malicious, it could still find that the pursuit of its goals is hindered by humans. It would not be hostility as much as indifference to what happens to humans.

This topic was touched upon by Tegmark as well:

> “So we were basically building these alien minds that are much smarter than us, who we're gonna have to share the planet with. Right. And it can be really inconvenient to have to share the planet with much smarter alien minds that don't care about us. Just ask the Neanderthals and how that worked out for them.” (Tegmark)

There is a certain power imbalance that the interview participants, as well as Tegmark, touched upon with the comparison of ants and humans or, in the case of Tegmark, humans and the Neanderthals. Dangers do not always come from hostility but can also stem from the weaker side being an inconvenience to the stronger side.

4.3 Sentience

A discussion point in participant interviews was sentience. Sentience can be difficult to define and even more difficult to detect. All five participants shared their thoughts on whether it would even be possible to tell if an AI model is truly sentient.

> “We don’t really know how to define sentience or consciousness now, so to tell a difference. I don’t think we could.” (Participant C)

> “I don't particularly believe in a soul, so I think that machine consciousness is at least theoretically possible, and I think current AI systems are, at the very least, agentic. They are sort of sentient in the same way that some animals are sentient, like I wouldn't go as far as to describe it as personhood, but they are clearly intelligent and they clearly have their own feelings.” (Participant D)

> “I think that some humans would think it is sentient and would rally to demand the rights for these AIs. And other humans, maybe the AI developers themselves and computer scientists, would come out and say, “no, you get it totally wrong, just because something claims sentience does not mean it is sentient.” (Participant E)

Three participants agreed that they do not think it is possible to tell if an AI model is truly sentient. Participant E meanwhile discussed the potential split in society on this topic: while some people might fight for the rights of the AI models that have claimed sentience, computer scientists and AI developers would likely be the ones who would argue that the models are not sentient. Meanwhile, Participant D and Participant B had another thought: the sentience of such models could be just very different from humans. There is a clear pattern in participants' thoughts: more work needs to be done on how to first define and then detect

proper consciousness as opposed to mimicking sentience from the data that the AI model was trained on.

Sentience of AI models is discussed by Hinton and Mitchell in their interviews as well. Hinton pointed out the following:

> "When it comes to sentience, I'm amazed that people can confidently pronounce these things are not sentient. And when you ask them what they mean by sentient, they say, well, they don't really know. So, how can you be confident they're not sentient if you don't know what sentient means? (…) I think whether they're sentient depends on what you mean by sentient. So you better define what you mean by sentient before you try and answer the question "Are they sentient?" (Hinton)

Hinton also acknowledged the possibility that AI sentience might be quite different from human sentience:

> "So it's quite possible there's, if it is at all sentient, it's sentient in a somewhat different way from us." (Hinton)

The complexity of trying to define or detect sentience was also discussed by Mitchell:

> "Intelligence is one thing, sentience is another. There's a question of whether, you know, how they're related or if they're related at all, you know, and what they all actually mean. And these terms, this is one of the problems, of course, these terms are not well defined, but most, I think most people in AI would say that intelligence and sentience are different." (Mitchell)

Both key figures have the opinion that before any attempt at trying to detect sentience, it is first important to define it, which itself can be a challenge. Hinton focused more on the fact that people cannot even define what they mean by sentience when they claim an AI is not sentient, and said an artificial intelligence might be vastly different from human intelligence. Mitchell focused more on the distinction of intelligence and sentience, claiming that just because something is intelligent, does not mean it is sentient.

Interview participants and the key figures had similar ideas. They both agree that there is a lack of clarity in what sentience actually is, how to define it, and how to detect it. Additionally, they stated that it is important to be cautious with claims of sentience from AI

models. Hinton also shared a similar idea with many participants in his idea that an AI sentience might be vastly different from human sentience, and as such, it could be more difficult to detect if people were only looking for sentience similar to their own. As such, it seems that more research is needed to be able to have ethical guidelines on how to define and/or detect AI sentience and how such systems should be handled.

4.4 Intentional misuse and unintended consequences

A prevalent theme of the interviews was the contrast between intentional misuse and unintended consequences of AI development. Two central topics of this theme will be discussed in this subchapter: unpredictability of AI models and their intentional misuse.

4.4.1 Unpredictability

Developing AI models can come with some unpredictable risks. The participants were asked about their ideas on the risks of the misuse of AI models compared to the unintended consequences that developing AI models could cause. Their views can be seen in the following quotes by three participants:

> "I certainly think that misuse is plausible also, but less likely to result in existential levels of risk to humanity." (Participant A)

> "To some extent, the alignment problem is about making an AI that's in the interests of humanity as a whole and misuse or accidents are both different ways you can fail at alignment. Misuse is "aligned to a subset of humanity", and accident is "it's not aligned to anyone at all." (Participant D)

> "I think I'm more worried about the unintended consequences, mostly because purposeful misuse can still be predicted. I think we, more or less, can understand and predict where purposeful misuse comes from. And any consequences are, well, by definition, unintended things; these are less predictable unless we proactively try to identify these risks ahead of time." (Participant E)

While participants considered both intentional misuse and unintended consequences to be a potential issue, unintended consequences were considered a bigger threat in the long run. This was due to the inherent unpredictability, which means they are harder to prepare for. Participant D highlighted an important thought that while intentional misuse is still aligned to

the goals of some groups of people, unintended consequences might not be. When talking about recommendations on how to diffuse the risks of unintended consequences, participant D mentioned that there is a risk for both and that "to some extent, those goals are in conflict with each other because anything that makes the AI more obedient, thus making accidents less likely, makes misuse more viable."

The unpredictability of AI development was discussed by some of the key figures as well. As a potential solution to the dangers of unintended consequences, Bengio offered a solution:

> "Why not build machines that understand the world like a scientist? Not a business person, a scientist, not trying to be a product to cater to our needs, but just be truthful and humble, in exactly the right measure. And we could use that without putting the goals part, which is potentially dangerous. (…) So we could potentially build machines that are helping us solve the challenges of humanity without taking the risk of putting this goal-seeking machinery into this. It doesn't solve all the problems, but at least we know that it's not going to blow in our face. But a human could still use these things for, you know, designing new weapons, for example. So it doesn't solve the social problem. It doesn't solve the political problem. But at least we don't get this unintentional loss of human control, which could spell, you know, catastrophic outcomes." (Bengio)

Bengio suggested that focusing on truth-seeking systems as opposed to goal-driven ones could be a solution for safer development. While such models could still be misused by people, it would greatly reduce any unintended, potentially catastrophic outcomes. As such, should AI development go down that route instead, there is still the issue of the social problem, finding solutions on how to stop people from misusing AI systems. For this, general social initiatives for reducing overall crime rates could be a solution, as misuse generally refers to the use of AI models for criminal or ethically questionable activities.

Participants and experts agreed that the unintended consequences are a greater risk in the long term. This is due to it being more difficult to prepare for as opposed to intentional misuse. Both groups emphasised the need for better safety measures to proactively identify potential risks. However, Bengio in particular offered another angle of discarding the development of goal-driven systems and focusing more on problem-solving systems instead. He was therefore focused more on the redesigning of the entire section of the industry, while the participants

were more modest in their expectations and focused on the safety measures needed for the current goal-driven models.

4.4.2 Intentional misuse

While unintended consequences were commonly thought of as a bigger threat when talking about the scale of the threat, the intentional misuse was brought out as a threat that was more likely to happen and already a danger now.

> "At this moment in time, I think the purposeful misuse is a bigger concern, like with current AI tools, anyone can ask very criminal or just, well, bad things and learn from it, but then when we think in the future, it is more about unintended consequences." (Participant B)

> "People can use AI to do things, like bad things, that they could never do themselves. What I mean is that so many more people have access to essentially a weapon they could use."(Participant C)

All participants considered intentional misuse to be a concern for various reasons: deepfakes, using AI as a tool to create bots, and being able to acquire information for criminal or ethically questionable activities more easily. That is to say that while unintended consequences in the future were considered a risk, participants emphasised that misuse of AI models should still be considered important, and AI safety legislation needs to take this into consideration as well.

The threat of intentional misuse was discussed by the key figures as well.

> "I talked with a guy not long ago, for example, who is using AI to develop helpful drugs to cure diseases. To see how it could make people live as long as possible, and just as a test change, put in a minus sign, so we try to make it kill people instead of save people and very promptly, it invented the VX, one of the most potent chemical weapons known to humankind. So this shows right there that with intelligence comes power." (Tegmark)

> "But a human could still use these things for, you know, designing new weapons, for example. So it doesn't solve the social problem. It doesn't solve the political problem." (Bengio)

> “The thing that disgusted me most was a proposal for a self-healing minefield. So the idea is, look at it from the point of view of the minefield. When some silly civilian trespasses into the minefield, they get blown up. And that makes a hole in the poor minefield. So it's got a gap in now, so it's not fit for purpose. So the idea is maybe nearby mines could communicate, and maybe they could move over a bit. And they called that healing. And it was just the idea of talking about healing for these things that blow the legs off children. I mean, and the healing being about the minefield healing. That disgusted me.” (Hinton)

The interview participants and key figures both considered intentional misuse of AI models to be a more current threat. They both highlighted how intentional misuse is more of a societal problem. However, key figures had slightly darker examples of how intentional misuse of AI models can cause tremendous harm when compared to the interview participants. Hinton, in particular, highlighted that it is not only criminals who can misuse AI models. Rather, he also considered it a misuse if the military were to use AI to create and enhance weapons with the intention of causing more harm. Those examples of already the current AI models being used for harm, exemplifies that even today it is important to make sure that people, on one hand, know how to use AI models correctly (for example, younger people might not even realise the harm they could do with deepfakes) and on the other hand it raises a question on if access to such things should be as open as it is today.

4.5 Who Controls AI?

With a potential AI that is aligned with human goals comes another question: who should be in control of it? In a discussion with participant D, when asked about any further topics he would like to discuss, he brought up the question of who can be trusted enough to be in control of an AI.

> “Another big thing is, if AI is aligned, who should get to control it? So, for example, I really don't trust private companies, but I trust government intelligence agencies a lot less. So if AI is to be developed, I would much rather it be developed in the private sector, especially in the open source part of the private sector, where it's more transparent than in some CIA basement where no one's able to hold accountable. Intelligence agencies have a very big problem of not really remembering that they're supposed to be protecting the public, not controlling them, and I think that

> corporations, for all their faults, do not have that level of hubris. So I would much rather live in a corporate dystopia than a totalitarian one, if I had to pick." (Participant D)

Participant D's perspective showed a distrust of the government in two regards: transparency and accountability. He did not consider private companies to be a good solution either, but considered it a better solution, especially if AI development took place in the open-source part of it. As such, the potential solution for the distrust of AI development would be for it to be as transparent as possible to allow people to feel they can see what is happening at all times.

The control or ownership of AI was briefly touched upon by Hinton and Bengio in their interviews as well.

> "The end game is one where no single person, no single corporation, no single government has too much power. That means it has to be that the governance, like the rules that we decide, you know, how we use AI and so on, have to be multilateral, that involve many countries." (Bengio)

> "You don't want some big for-profit company deciding what's true"(Hinton)

> "Well, when it comes to the issue of what's true, I mean, do you want the government to decide what's true? It's this big problem right? You don't want the government doing it either." (Hinton)

Both expressed the same central idea: it is dangerous for one single entity to be in control of the AI systems, no matter if it is a private company or the government. Following the idea of the dangers of centralised power, a collaboration between the private sector and the government agencies could be established to try and distribute the control that any one entity would have over AI systems.

Participant D and the key figures (Bengio and Hinton) shared a scepticism of centralised power over AI systems. Neither thought there is a good solution, as neither the government nor the private companies should have that power, in their opinion. However, there was a difference in how Participant D and the key figures saw the possible best-case scenario. While Participant D was more in favour of that power being in the hands of the private companies, specifically in the open source part, the key figures were not in favour of either.

Rather, especially Bengio considered it best to have that power distributed internationally instead.

5. Discussion

The thematic analysis conducted in this thesis allows the general concerns regarding the development of AI to be divided into five broader categories: general AI safety, AI alignment, sentience, intentional misuse and unintended consequences of using AI models, and the concern regarding the concentration of power that AI could cause. This section will first highlight the contributions that this study has made to the field. This is followed by a discussion of the main findings of the thesis, connecting them back to the literature review. Finally, the limitations of the study and recommendations for further research are discussed.

5.1 Contributions of the study

The thesis makes several contributions to the AI safety and alignment field by highlighting the opinions and concerns of people outside the key figures of the field. This offers a more nuanced view on the topics of AI safety and alignment that might often be overlooked in favour of the opinions of the key figures in the field.

5.1.1 Categorisation of main concerns in AI safety and alignment community

Five main concerns were identified among research participants when asked about their opinions on AI safety and alignment:

1. AI safety regulations;
2. AI alignment;
3. Sentience;
4. Intentional misuse and unintended consequences;
5. Power concentration.

This categorisation gives a structured framework that future research could use when analysing the views on AI safety and alignment across different groups of people.

5.1.2 Underexplored theme: power concentration

The topic of power concentration was discussed as an underrepresented topic in AI safety discourse. Participants and key figures alike expressed concern about the concentration of control and power in the face of powerful AI systems. This emphasises that the topic of AI

safety is not only about the technical regulations but is also a political concern about control and power dynamics. This needs to be taken into account in policymaking, including in efforts like the EU AI Act.

5.1.3 Insight into misuse and everyday risks

The findings indicate that among the members of online AI safety and alignment communities, the focus in regards to AI safety is more on the immediate and already present risks posed by existing AI technologies. For example, the use of AI for criminal activities, creating deepfakes, and using algorithms for manipulation. This suggests the need to already address those risks in education and public conversations to educate people on recognising and understanding those risks and how to avoid them.

5.1.4 Contribution to regulatory discourse

The analysis contributes to the ongoing regulatory discussions, such as those surrounding the EU AI Act, by highlighting the current regulatory insufficiencies that key figures in the AI safety field, as well as the members of the wider AI safety and alignment communities, see. The primary concern of the interview participants was that the current AI safety protocols are lacking and become outdated as soon as they are implemented, due to the long-time effort it takes to pass new regulations. This is a direct comparison to the EU AI Act, which has been criticised for its lack of adaptability. The findings support the call for more flexible and adaptable regulations in AI safety, which can be adjusted in a shorter time frame as needed.

5.2 Discussion of main findings

The thematic analysis conducted in this thesis allows the discussion on the development of AI to be divided into broader categories: AI safety, slowing down AI development, AI alignment, sentience, and the concern regarding the concentration of power that an AI could cause. In the following subchapters, those findings will be discussed and connected to the literature review.

5.2.1 AI safety regulations

A dominant concern emerging from the findings is the lack of current AI safety protocols, or at the very least, the lack of belief in the ones currently in place being sufficient. One of the main reasons for the concern is the rapid pace at which AI development is moving, giving

little time to put any new regulations into place. The current safety measures were described as patchy by participants and key figures alike. The advancements in AI development are often moving faster than any regulatory changes that could come into effect to govern them. This means that by the time a law or regulation is passed, it is already outdated. This is similar to the concerns expressed in Bostrom's (2024) *Superintelligence: Paths, dangers, strategies,* in which he discusses the moderate to fast takeoff scenarios for AIs. In those two scenarios, the capabilities of an AI model rapidly bypass any safety regulations, leaving people unprepared for emerging threats (*ibid.*). AI development being compared to an *arms race* gives context to the need that the participants, as well as Bengio, emphasised for having regulations around developing more powerful AI models. Especially considering how more than one interview specified how it currently seems that it would be a winner takes it all type of situation, where there is no prize in it for someone who gets to an AGI second. As a recommendation from Etzioni & Etzioni (2017), international collaboration and adaptable regulatory frameworks need to be established to diffuse those risks.

5.2.2 Slowing down AI development

Among the interview participants, it was a common opinion that at least human-level AGI will be reached at one point, sooner rather than later. By that point, and even before, much better safety regulations would need to be in place. Notably, this is an opinion shared by many well-known experts in the field of AI, as shown by the number of people who signed "Pause Giant AI Experiments: An Open Letter," hoping that it would at least be possible to slow down the development efforts for some time until safety measures could be put into place. Among interview participants, the opinions on whether AI development should be slowed down were slightly varied. Some did think that slowing down AI development is a good idea, but were worried about how realistic it is to achieve it in today's political and economic climate. Participant D, on the other hand, also wanted the flipside to be considered, namely, whether slowing down the development could also cause harm, as such technology could be used for great benefit. This highlights the balance needed between safe development while also not hindering the beneficial progress that AI could bring, a concept which is also reflected in Floridi and Cowls' (2019) framework of ethical AI development principles, which considers the balance between caution and beneficence to be important.

5.2.3 AI alignment

AI alignment was generally a topic of concern among interview participants and key figures alike, which is an expected result as the participants were found from online communities about AI alignment and safety. The degree to which the participants were concerned about potential misalignment varied, but the likelihood of successfully achieving AI alignment varied from 20% to 50%. The participants highlighted the difficulties of being able to communicate the goals that humans have to the AI systems in the first place. This is referenced by Mitchell (2019) and Christian (2020) as well, with their ideas that human values and intentions can be complex and ambiguous, making it hard to convey them clearly to AI systems. Meanwhile, Mitchell herself was actually more concerned about the more current threats that AI systems could have, rather than misalignment and any potential threats stemming from that.

5.2.4 Sentience

Sentience was another concern that the interview participants and key figures aligned on. Both groups brought out the inherent difficulty in defining what sentience even is, with Hinton especially highlighting that when people say that AIs are not sentient, they cannot even explain how they would define sentience in the first place. As Mitchell stated, intelligence is not automatically sentience, which is discussed by Yudkowsky's (2001) as well, as there is a difference between cognitive capabilities and actual consciousness. According to Davies (2010), more interdisciplinary research would be needed to clarify the terminology. One participant considered the possibility that an AI could have sentience, but it would be different from human sentience, comparing it more to animal sentience instead. This was similar to Hinton's idea that if an AI could be considered sentient, it might not be sentient in the exact same way humans are.

5.2.5 Concentration of power

An unexpected theme that emerged from the interviews was the concept of control. Namely, who should have control over potential AGIs in the future and who should be trusted with this power. This topic was not directly included in the participant interviews, but rather came up naturally through discussion. Participant D believed there to be no perfect scenario, as the two potential entities in control - private companies and governments - both have their shortcomings. However, he went on to decide that between those two, he would trust

governments less than private companies. This concern for power and control was mirrored by Bengio, who believed the best case scenario to be one where no single entity has too much power and that it should be a multicultural affair. Hinton agreed that neither for-profit companies nor governments being in control of what is true is an ideal scenario.

5.3 Limitations and recommendations for further research

The thematic analysis conducted for this thesis was a small-scale qualitative research focusing on five participant interviews and five interviews that were available on Youtube with key figures in the AI safety and alignment field. The findings from participant interviews mostly align with those of key figures, with some notable differences being the bigger emphasis on social aspects of AI usage. However, it is important to remember that this has been a small-scale qualitative study looking into a small number of interviews and, as such, represents only a small part of the whole. Additionally, the thesis did not explore the cultural and regional differences that could affect the views on AI safety and alignment due to the relatively small sample size of interview participants. Future work could expand on this topic on a larger scale, aiming to recruit more participants for the interviews or instead conducting quantitative research on the written material available in various online AI alignment communities.

Based on the findings, the following recommendations can be provided for further research topics.

1. **Policy development and regulatory frameworks.**

   Concern about current AI safety policies was a prevalent topic of the interviews. Participants considered a key concern to be the speed at which AI developments are made and the inability of policies and regulations to keep up with those developments. Further research could be conducted to look into adaptable and rapidly implementable AI regulations which could match the pace at which technological advancements are made.

2. **Public engagement in AI safety.**

   The general public is actively using or, at the very least, is being affected by the use of AI. Further research could be conducted to include the general public in the discussion regarding AI safety. This could provide insight into the concerns and education needs

that the general public has regarding AI and be used to raise social awareness on the topic of AI usage, the risks that arise from it, and give insight into how to include those topics in the education of children and adults alike.

3. **Collaboration with AI researchers.**

   More interaction in online forums or other collaborative platforms between key figures in the AI safety and alignment field and the general research community, or even the general public, could offer further insight into how the ideas that different groups have on AI safety differ and how best to acknowledge them. Furthermore, this collaboration effort could be examined to see how it affects the public's understanding of AI safety and alignment and the development of AI regulations, as policymakers are not always experts in the field.

6. Conclusion

The aim of this thesis was to explore and compare the opinions and concerns about AI safety and alignment among members of the online AI alignment communities to compare them to the narratives by key figures in the AI safety and alignment field. The thesis used a qualitative approach. Five interviews were conducted with members of online AI alignment communities from February to April of 2025. In addition to those, five interviews were chosen from YouTube with well-known figures in the AI alignment space for analysis as well. Those interviews were with Max Tegmark, Yoshua Bengio, Geoffrey Hinton, Yann LeCun, and Melanie Mitchell. Those were chosen as a point of comparison for the interviews conducted with the members of the online AI alignment communities. All interviews were transcribed, and thematic analysis was conducted on the gathered data using ATLAS.ti.

Five key themes emerged from the analysis: AI safety, AI alignment, sentience, intentional misuse vs. unintended consequences, and the question of who should have power over AI. The most prevalent theme was the concern regarding the shortcomings of the current safety measures in place for handling risks associated with AI, with a common opinion being that the development of AI models is simply moving too fast for regulatory changes to be able to keep up with the pace. Closely related was the topic of AI alignment, on which the participants were more pessimistic in regards to the possibility of it being aligned to human interests in the long run. This was a great contrast to some of the interviews with the key figures in the AI alignment space, as there are several opponents to the idea that AI misalignment is a threat, such as Melanie Mitchell and Yann LeCun. However, in both participant interviews and the interviews with key figures in the field chosen from YouTube, a common idea was that no matter what people think of the long-term implications of AI development and the potential for misaligned AI and what it could mean, AI safety is still an important topic and safety regulations are needed. This is because even when disregarding any unintended consequences that could happen from misalignment, intentional misuse is still a concern. With AI models, people who might otherwise lack resources to do bad things would now have access to what was even described as a potential weapon to which anyone could have access.

Further research into this topic would be beneficial to understand the broader concerns of the AI researchers on the topic of AI safety and alignment. Furthermore, although this thesis focused only on the people already engaged with the topics of AI and its safety and

alignment, a deeper look into the opinions of the general public would be beneficial. AI is becoming increasingly a part of people's everyday lives, and AI tools are used by all age groups and by people from very different backgrounds. As such, insight into the opinions and concerns the general public has on the development of more powerful AI models would be beneficial. This knowledge could be used to look for ways to alleviate those concerns, as even people who are not involved in AI development in any way will nonetheless be affected by the outcomes.

References

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), 973-989.

Anthropic. (2025). *Researching at the frontier.* Retrieved April 12, 2025, from https://www.anthropic.com/research.

Archibald, M. M., Ambagtsheer, R. C., Casey, M. G., & Lawless, M. (2019). Using Zoom videoconferencing for qualitative data collection: Perceptions and experiences of researchers and participants. *International Journal of Qualitative Methods*, 18.

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and Machine Learning Limitations and Opportunities*. The MIT Press.

Baum, S. D. (2018). Reconciliation between factions focused on near-term and long-term artificial intelligence. *AI & Society*, 33(4), 565-572. https://doi.org/10.1007/s00146-017-0734-3.

Boddington, P. (2023). *AI Ethics. A Textbook.* Springer.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology,* 3(2), 77–101.

Braun, V., & Clarke, V. (2022). *Thematic analysis: A practical guide*. SAGE.

CBS Mornings. (2023, March 25). *Full interview: "Godfather of artificial intelligence" talks impact and potential of AI* [Video]. YouTube. https://www.youtube.com/watch?v=qpoRO378qRY.

Chouldechova, A. & Roth, A. (2020). A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM,* 63(5), 82–89. https://doi.org/10.1145/3376898.

Christian, B. (2020). *The Alignment Problem*. W.W. Norton & Company.

Davies, P. (2010). *The Eerie Silence: Renewing Our Search for Alien Intelligence*. Houghton Mifflin Harcourt.

DeepMind. (2025). *Responsibility & Safety*. Retrieved April 12, 2025, from https://deepmind.google/about/responsibility-safety/.

Dippnall, S. (2024). The Top 5 AI Challenges: Insights and Solutions. Retrieved November 18, 2024, from https://www.sandtech.com/insight/the-top-5-ai-challenges-insights-and-solutions/.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. Advocates for using inherently interpretable models rather than post hoc explanations for black-box models. *Machine Learning.* https://doi.org/10.48550/arXiv.1702.08608.

Dr Brian Keating. (2024, December 9). *Yann LeCun: AI Doomsday Fears Are Overblown [Ep. 473]* [Video]. YouTube. https://www.youtube.com/watch?v=u7e0YUcZYbE.

Dung, L. (2023). Current cases of AI misalignment and their implications for future risks. *Synthese* 202, 138. https://doi.org/10.1007/s11229-023-04367-0.

DW News. (2023, April 14). *Max Tegmark interview: Six months to save humanity from AI? | DW Business Special* [Video]. YouTube. https://www.youtube.com/watch?v=ewvpaXOQJoU

Etikan, I., Musa, S. A., & Alkassim, R. S. (2016). Comparison of convenience sampling and purposive sampling. *American Journal of Theoretical and Applied Statistics*, 5(1), 1. https://doi.org/10.11648/J.AJTAS.20160501.11.

Etzioni, A., Etzioni, O. (2017) Incorporating Ethics into Artificial Intelligence. *Journal of Ethics,* 21(4), 403-418. https://doi.org/10.1007/s10892-017-9252-2.

Felzmann, H., Villaronga, E. F., Lutz, C., & Tamò-Larrieux, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society,* 6(1). https://doi.org/10.1177/2053951719860542.

Floridi, L., & Cowls, J. (2019) A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*, 1(1). https://doi.org/10.1162/99608f92.8cd550d1.

Future of Life Institute. (2023). Pause giant AI experiments: An open letter. Retrieved October 11, 2024, from https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

Google AI. (2025). *Our AI Principles*. Retrieved April 12, 2025, from https://ai.google/responsibility/principles/.

Human Brain Project. (2023). *Human Brain Project & EBRAINS*. Retrieved November 11, 2024, from https://www.humanbrainproject.eu/en/about-hbp/human-brain-project-ebrains/.

Lovelock, J., & Margulis, L. (1974). Atmospheric homeostasis by and for the biosphere: The Gaia hypothesis. *Tellus*, 26(1-2), 2–10.

Luccioni, S. (2023). *AI is dangerous, but not for the reasons you think. TED*. Retrieved October 14, 2024, from https://www.youtube.com/watch?v=eXdVDhOGqoE.

Luccioni, S. (2024). The environmental impacts of AI -. Primer. Retrieved October 14, 2024, from https://huggingface.co/blog/sasha/ai-environment-primer.

Machine Learning Street Talk. (2025, January 15). *GODFATHER OF AI: MIGHT THE ROBOTS TAKE OVER?* [Video]. YouTube. https://www.youtube.com/watch?v=G1ARvwQntAU.

Mitchell, M. (2019). *Artificial intelligence: A guide for thinking humans*. Farrar, Straus and Giroux.

OpenAI. (2025). *Safety Approach*. Retrieved April 14, 2025, from https://openai.com/safety/.

Policy-Relevant Science & Technology. (2023, June 24). *Munk Debate on Artificial Intelligence | Bengio & Tegmark vs. Mitchell & LeCun* [Video]. YouTube. https://www.youtube.com/watch?v=144uOfr4SYA.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4),515–526.

Priestley, M, O'Donnell, F, & Simperl, E. (2023). A Survey of Data Quality Requirements That Matter in ML Development Pipelines. *Journal of Data and Information Quality* 15 (2), 1-39. https://doi.org/10.1145/3592616.

Rapley, T. (2007). *Doing conversation, discourse and document analysis*. SAGE Publications.

Reddit. (n.d.). r/ControlProblem [Online community]. Reddit. https://www.reddit.com/r/ControlProblem/.

Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach (4th ed.)*. Pearson Education.

Russell, Stuart J. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.

Scripps Research. (2023, August 4). *Melanie Mitchell: Straight Talk on A.I. Large Language Models* [Video]. YouTube. https://www.youtube.com/watch?v=0PxygSrMCv4.

Sindin, X. (2017). Secondary data. In *The sage encyclopedia of communication research methods* (Vol. 4, pp. 1578-1579). SAGE Publications, Inc, https://doi.org/10.4135/9781483381411.n557.

Taylor, J. (2025). *'No consent': Australian authors 'livid' that Meta may have used their books to train AI.* Retrieved April 12, 2025, from https://www.theguardian.com/technology/2025/mar/25/no-consent-australian-authors-livid-that-meta-may-have-used-their-books-to-train-ai-ntwnfb.

Tegmark, M. (2014). *Our Mathematical Universe: My Quest for the Ultimate Nature of Reality*. Knopf Doubleday Publishing.

Tegmark, M. (2017). *Life 3.0: Being Human in the Age of Artificial Intelligence.* Knopf.

Turing, A. (1951). Can Digital Computers Think? In B J Copeland (ed.) (2004) *The Essential Turing.* Oxford University Press. https://doi.org/10.1093/oso/9780198250791.003.0019.

Yudkowsky, E. (2001). *Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures*. The Singularity Institute.

Zhang, D., Xia, B., Liu, Y., Xu, X., Hoang, T., Xing, Z., Staples, M., Lu, Q., & Zhu, L. (2024). Privacy and copyright protection in generative AI: A lifecycle perspective. Proceedings of the 2024 IEEE/ACM 3rd International Conference on AI

Engineering: Software Engineering for AI (CAIN) https://doi.org/10.1145/3644815.3644952.

Appendix 1. Interview questions

Introduction

- In this section, we aim to learn about the participants' background in AI to find out if having different backgrounds (for example, AI professional vs general AI enthusiast) affects their opinion regarding AI.

1. **(Aim: get to know the participant in regards to their background in AI.)**
   a. Could you tell me a bit more about your **background** regarding AI? Are you **professionally** involved with AI, or is it more of a **personal interes**t?
      I. What got you interested in AI in the first place?

2. **(Aim: to understand the participant's opinion of AI and superintelligence in general)**
   a. What is your **general outlook** on current AI development efforts? Are you concerned, excited, indifferent? Why?
      I. What excites or worries you the most?
   b. Do you think a superintelligent AI is **inevitable**?
      I. (If yes) On what kind of **timeline** do you think this would be? Months, years, decades?

Safety and alignment

In this section, we wish to see the participants' views on AI safety and whether they think the current efforts in ensuring it are enough.

Some people believe that AI could potentially be a risk to humanity and threaten its survival, and push for AI safety, while others think such beliefs belong to the realm of science fiction.

1. **(Aim: To see what the participant considers "AI safety")**
   a. What is your **understanding of AI safety**?
      I. Do you think AI safety is more about preventing **purposeful misuse or preventing unintended consequences?**

2. **(Aim: to see whether the participant is content with current safety protocols)**
   a. Do you believe the **current measures** for ensuring AI safety are sufficient or are there shortcomings?
      I. If there are shortcomings, what gaps do you see? Are there specific examples that concern you, or just the overall state of things?

3. **(Aim: to see the participant's stance on AI alignment and its likelihood)**
 a. There's a lot of talk about AI alignment, ensuring that AI follows human values. How realistic do you think it is to keep AI truly **aligned with human goals**?

Misalignment

In this section, participants can express their ideas on the risks of misaligned AI.

One concern is that AI would not act in the best interest of people, not even out of malice but because there is a miscommunication between us and AI, so the AI is misaligned. Do you think that's a potential risk?

1. **(Aim: To see if the participant has any concerns about misaligned AI)**

    a. What are the **risks** you see with **misaligned AI**?
        I. Is there a particular risk that worries you the most?

    b. Some experts suggest that AI development should be **halted** to ensure safety, take, for example, the **Open Letter to pause Giant AI experiments**. Do you think that's a good idea, or is it unrealistic?

Technical aspects

In this section, we wish to give the participant an opportunity to give their ideas on what the technical difficulties are in regard to AI alignment.

In talks about AI safety, a lot of focus is on the policy side of things, but it could also be a **technical problem.**

1. **(Aim: To see if the participant has some technical background with AI and, if so ,what they consider the main technical difficulties in achieving AI alignment)**

    a. If you have had **previous experience** with AI development, what do you think are the main t**echnical difficulties** in achieving AI alignment or AI safety in general?

Ethics and Policy

- In this section, we wish to see the participants' stance on the ethics of AI development.

The public's perception of AI varies a lot, some love it, some hate it. How well do you think the general public understands the potential ethical dilemmas of AI?

1. **(Aim: learn what the participant thinks should be the basis of developing AI)**

    a. The United Nations' Universal Declaration of Human Rights defines some **core human values** and rights such as dignity, liberty, equality, the right to life and the prohibition of slavery and torture. Do you think an AI should **reflect those global values** or adapt to **specific cultural and societal norms?**

(e.g. Like for example how topics such as women's rights and even what foods are considered ethical are very different in different cultures?)

2. **(Aim: learn if the participant thinks there are any other ethical dilemmas regarding AI development)**

    a. What **ethical dilemmas** do you see in AI development today?

I. What could be the possible solutions if you can think of any?

Tool vs. Sentience

- In this section, we aim to find whether the participant views AI development as a way to get a tool for humans to use or a sentient entity with its own goals and aims.

Some say AI will always be just a **tool** for people to use, while others see AI as something that will one day have its own **goals**. What's your perspective?

1. **(Aim: To learn about the participant's views on whether AI for them is a simple tool or a potential consciousness)**

   a. Do you believe AI is and always will be akin to a **smart tool,** or does it have the potential to achieve **consciousness**?
      I. (If yes) Could it just be **mimicking human behaviour** from the data it was trained on?

2. **(Aim: To learn about the participant's perception of whether an AI could be considered sentient enough to have its own rights)**

   b. If we ever developed an AI that claimed to be sentient, how should we **react**?
      I. Should a sentient AI be given **rights similar to humans**?

Conclusion

In this section, we wish to give the participant an opportunity to add anything they wish that they could not mention during the rest of the interview and see which topics they consider most memorable or important

1. **(Aim: To see if the participant has more topics that they want to discuss or anything else to say)**

   a. Are there any other topics related to AI and its alignment that **we haven't covered** that you think are important?

2. **(Aim: to see which topic is most memorable/important to the participant)**
   a. What do you consider the **most important and memorable topic** discussed today?

Non-exclusive licence to reproduce the thesis and make the thesis public

I, Evelin Jõemaa,

1. grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the digital archives of the University of Tartu until the expiry of the term of copyright, my thesis

   Alignment and safety challenges in a superintelligent AI landscape,

   supervised by Grace Eden;

2. grant the University of Tartu a permit to make the thesis specified in point 1 available to the public via the web environment of the University of Tartu, including via the digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright;

3. am aware of the fact that the author retains the rights specified in points 1 and 2;

4. confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Evelin Jõemaa

15/05/2025