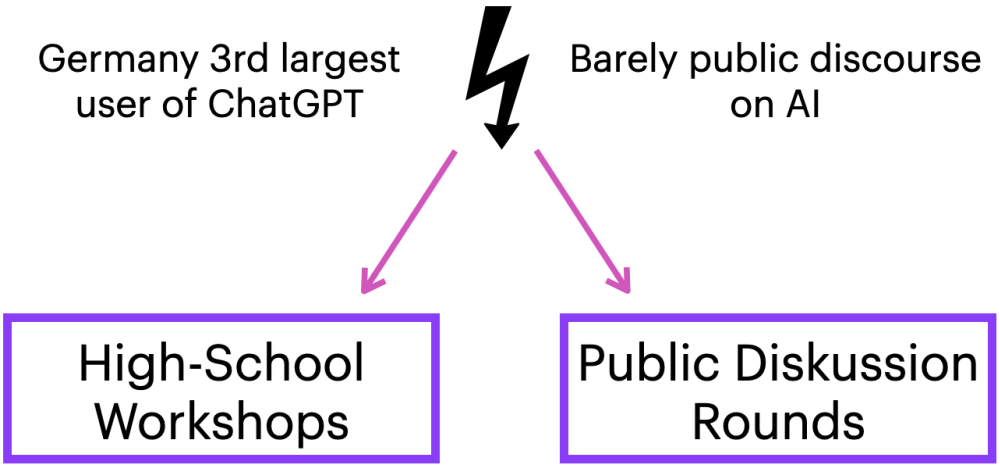
AISES Communication Project: AI for Everybody

Manon Kempermann

May 2025



Abstract

This report details my final project for the AI Safety, Ethics and Society course (AISES) by the Center for AI Safety (CAIS). My project is a communications and public outreach initiative targeted at German citizens, both high-school students and adults. According to OpenAI, Germany is the third largest user of ChatGPT [1], yet there is barely any meaningful discourse in German society about AI and its potential implications for our future. To bridge this knowledge gap, I have developed an AI workshop for high-school students and a guided discussion round for adults with the intention to start conversations about AI and provide participants with knowledge about current developments. I have successfully conducted the first discussion round and will use the insights and feedback gathered to improve future events.

Contents

1 Introduction

My project aims to educate German citizens through interactive workshops and guided discussion rounds about AI and what our future with it might hold. In this report, I will begin by outlining my motivation for the project and my theory of change. I will then detail my plans for the events I have been organizing. As I have already conducted one discussion evening, I will reflect on my own insights and the feedback I received, as well as how I plan to adapt future events. Finally, I will briefly outline where I envision this project continuing over the coming months.

2 Motivation and Background

I am deeply concerned about where the next few years will lead us with AI. Given my position as an undergraduate student, I see two avenues through which I can contribute to mitigating existential risk: first, by dedicating as much time as possible to technical AI safety research alongside my computer science studies; and second, by engaging in public outreach, where I see both urgent need and potential impact. This project represents my first step toward the latter goal.

Through my deeper exploration of AI Safety, especially as the AISES course broadened my perspective on wider societal challenges, I often felt isolated in my concerns. Most people I spoke with about AI mentioned occasional use of ChatGPT, noting inconsistent performance—sometimes good, sometimes bad—but not fully grasping its significance. Conversations with my high-school-aged siblings revealed that even three years after ChatGPT's release, schools still lack strategies for addressing AI use, while students continue to rely on AI for completing their homework. This disconnect between public awareness and actual technological development is striking. Germany typically errs on the side of excessive regulation, yet in this domain, public discourse is notably absent.

Additionally, I worry about AI's potential to further divide German society, spread misinformation, diminish educational standards, and ultimately threaten our democratic foundations. In a democracy like Germany, there should be widespread awareness and understanding of such significant challenges. It is fundamentally undemocratic when this knowledge is lacking and when a small number of large foreign companies effectively determine society's future—indeed, the future of life on Earth.

One dimension of this democratic concern is the accessibility of AI technology and how it's handled in various environments. Who will be able to effectively cooperate with AI without diminishing their own intellect is likely determined by how schools, companies, and social environments address this technology. In the current state, there is great potential for increasing inequality among citizens, as AI literacy will become as fundamental as being able to read and write.

I am determined to change this. My primary goal for this project is to initiate meaningful discussion in Germany about AI and to educate people about what our potentially very near future with advanced AI might entail. I aim to raise awareness that AI represents far more than mere technical innovation—it is a technology deeply embedded into society. Beyond the already considerable technical challenges in AI safety, I believe the social ripple effects, amplifying various risks, are even more concerning.

3 Theory of Change

I conceptualize the potential impact of this project along two dimensions: first, how it might contribute to a broader public movement advocating for enhanced governance and a slow-down of the pace of AI development; and second, what personal and local effects it might generate.

I believe it would be prudent to decelerate AI progress, allowing our global society time to deliberate on how we wish to coexist with AI and which collective values should guide its development. My assessment is that this could be achieved either through political and governance interventions or through public movements that influence market dynamics. In either scenario, the general public has a voice through their political and economic choices and demands, as well as through widely shared opinions and vocal advocacy. While my project will certainly not reach all of humanity—or even all of Germany—it aims to contribute to advocating for thoughtful AI development. I hope it contributes to the growing discourse by increasing, even on a small scale, the number of people aware of these challenges.

Regarding the second dimension, I believe people have a right to be informed about these technological developments that are rapidly approaching. I hope my project encourages participants to discuss AI, to seek out information independently, and potentially to take action themselves. Particularly with the student workshops, I aim to motivate young people to continue developing their own intellectual capacities rather than delegating their critical thinking to AI systems. I believe that maintaining one's abilities to think, reason, and engage in critical discussion, while simultaneously learning to use AI effectively, represents one of the best strategies individuals can adopt to protect themselves from certain risks, such as gradual disempowerment or epistemic erosion.

3.1 Ideas on How to Measure Impact

Measuring impact can be challenging for educational initiatives like this, but I have several approaches that could help me assess effectiveness. First, I will distribute anonymous feedback forms after each event to gauge participants' satisfaction and collect their suggestions for improvement. This approach will also allow me to experiment with variations in content, methods, and structure to observe how these changes affect participants' perceptions and engagement.

At a broader level, I can track impact through the number of invitations I receive to present at schools, organizations, or local political parties, and monitor how many people I reach over time. Combined with participant feedback, these metrics should provide insights into my project's trajectory and highlight necessary adjustments to enhance reach and effectiveness.

Another approach could involve following up with select participants weeks after an event. I might ask acquaintances from venues where I've presented to discreetly inquire with other participants about which societal issues currently concern them. This indirect method could reveal whether the workshop content continues to resonate with participants or if it's quickly forgotten—providing valuable data on the lasting impact of my presentations.

For the high school workshops specifically, I could incorporate pre- and post-event surveys to measure changes in students' awareness and attitudes toward AI. This would provide more concrete evidence of the workshop's immediate educational impact.

4 Project Design and Planning

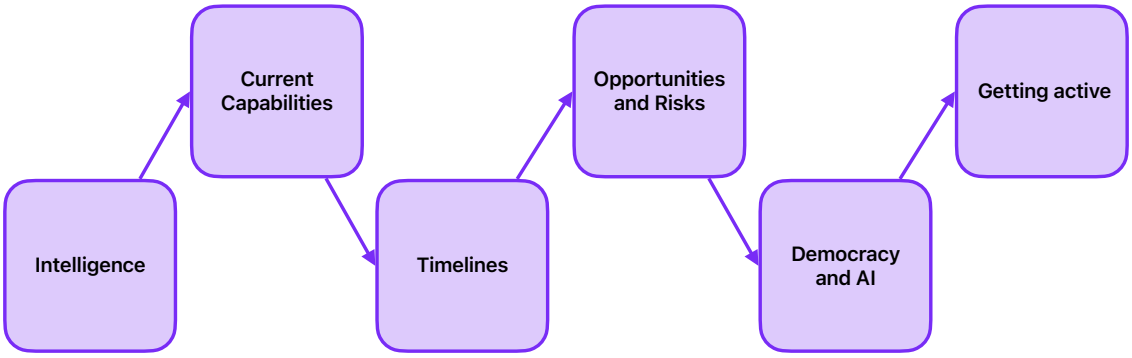


Figure 1: General event flow for both types of events. The difference lies in the methodology to bring across the concepts

I have planned two types of events: guided discussions for adults and interactive workshops for high-school students. The reason for this split is that I think both groups are in very different situations in their life and also might require different methodology. Both types of events however share a common theme in that I want to convey the big societal questions about AI without alienating people with technical complexity (see Figure 1 for the general structure). The events will be primarily given in German. Below, I outline the tentative structure and thought behind each event. My more detailed plans for each event can be found in Appendix A (discussion format) and Appendix B (high-school workshop).

4.1 Facilitation Approach

Both events aim to follow a Socratic dialogue approach, where I as the facilitator do not present my own opinions but rather provide knowledge and ask questions that help participants develop their own informed perspectives. This method encourages critical thinking about AI's implications while acknowledging the complexity and uncertainty of the subject. Rather than advocating for particular viewpoints, I want to create space for participants to engage with multiple perspectives and reach their own conclusions about how AI development should proceed and what values should guide it.

4.2 Discussion Event

4.2.1 Session Information

- **Title:** AI and Democracy
- **Duration:** 90-120 minutes
- **Target Audience:** Adults with general interest in AI and societal implications; no technical background required
- **Group Size:** 15-25 participants

- **Materials:** Flipchart paper, markers, prepared flipcharts with discussion questions, QR-codes for resource list and feedback form

4.2.2 Session Structure and Content

This workshop for adults is designed to be given in, for example, church communities, local party chapters, different kind of clubs, etc. It explores AI's implications for society through alternating input and discussion sections designed to foster dialogue and enabling participants to develop their own insights and perspectives regarding these socio-techincal challenges.

Introduction (5 minutes) The session begins with a welcome emphasizing the informal, conversation-oriented nature of the event. Participants are invited to explore questions about AI's role in society, democratic values, and our collective responsibility, with an emphasis that no technical expertise is required—only curiosity and willingness to engage in dialogue.

Understanding Intelligence (15-20 minutes) Opening with the fundamental question "What does intelligence mean to us, and when would AI be truly intelligent?", this section builds a common foundation for the discussions to follow. Participants are encouraged to explore various dimensions of intelligence (problem-solving, creativity, emotional understanding) with minimal guidance, allowing diverse perspectives to emerge organically.

Current AI Capabilities and Future Development (15-20 minutes) This section provides context on today's AI systems and their trajectory. Modern AI systems demonstrate genuine abilities to learn and reason, with models like Claude developing internal "features" representing millions of concepts. Current benchmarks show impressive capabilities, with latest models achieving over 90% on university exams [2] and solving 60-65% of complex software development tasks [3].

Timeline predictions from industry leaders are presented alongside broader research consensus, contrasting OpenAI CEO Sam Altman's prediction of AGI by 2025-2026 [4] and Anthropic CEO Dario Amodei's forecast of 2026-2027 [5] with the broader AI research community's estimates of 50% probability between 2040-2061 [6]. The group then discusses their reactions to these potential timelines, and what this potentially very near future could bring. This helps participants to think about the risks before they are officially introduced.

Benefits and Risks of Advanced AI (15-20 minutes) This section presents both the transformative potential of AI and its possible risks, allowing participants to weigh these perspectives. Potential benefits include medical breakthroughs, climate change solutions, and poverty reduction. The risks discussion incorporates expert warnings, like the Center for AI Safety's statement that "mitigating the risk of extinction from AI should be a global priority" [7].

The discussion introduces Nick Bostrom's Vulnerable World Hypothesis [8], which compares technological progress to drawing colored balls from an urn—while most have been beneficial (white) or mixed (grey), a "black ball" technology could threaten civilization. Advanced AI could represent such a risk by either being itself a "black ball"

through risks including misalignment, uncontrollability, gradual disempowerment of humans, power concentration, and social destabilization or could accelerate the discovery of other potential "black balls".

Democratic Values and AI Development (15-20 minutes) The group is split into small groups to explore AI's implications for core democratic values: human dignity, justice and equality, peace, and freedom. Each group considers questions like "How do we ensure AI development respects human dignity and autonomy?" or "How could AI systems restrict or expand personal freedoms?" After discussion, groups share key insights, building a collective understanding of AI's relationship to democratic principles.

Taking Action: Individual and Collective Responses (15-20 minutes) The workshop concludes by acknowledging that while these topics can feel overwhelming, there are ways to positively influence AI development. Different perspectives on AI development pace are presented, from the acceleration view (develop advanced AI quickly to address global problems) to the deceleration view (slow down to solve safety challenges), followed by a brainstorming session where participants identify potential actions at individual, community, and policy levels. At the end, participants receive a handout with recommended resources and are invited to provide anonymous feedback via QR code.

4.3 High School Workshop

4.3.1 Workshop Information

- **Title:** AI for Everybody: Understanding AI in Our Future
- **Duration:** 90-120 minutes
- **Target Audience:** High-school students (grades 10-13)
- **Group Size:** 15-30 students
- **Materials:** Interactive polling tools, laptop, projector, white/black board, handouts with resource links and feedback form

4.3.2 Workshop Structure and Content

This interactive workshop for high-school students is designed to foster critical thinking about AI through a combination of live demonstrations, peer discussions, and reflection exercises. The approach is deliberately non-technical, focusing instead on AI's societal implications and how students might navigate a future increasingly shaped by these technologies. I do not want them to feel judged because they use ChatGPT for their homework but rather teach them a responsible way to use these tools.

Introduction and Framing (10 minutes) The workshop begins with a disclaimer that this is not a technical session about how to use AI, but rather an exploration of AI's broader societal implications—a topic relevant to all citizens regardless of technical background. Students participate in a brief live poll about their familiarity with AI, usage patterns, and general attitudes (excited, concerned, neutral), establishing a baseline understanding of the group's relationship with AI technologies.

Current AI Capabilities (15 minutes) This section uses interactive polling and live demonstrations to help students develop an accurate mental model of modern AI capabilities. Students first predict whether AI can perform various tasks (passing exams, diagnosing cancer, programming websites), followed by demonstration of actual capabilities and limitations including knowledge cutoffs. This approach helps correct both overestimation and underestimation of AI abilities, a foundation for realistic assessment of future developments.

Brief Technical Context (5 minutes) A condensed, accessible explanation of the "AI triad" (algorithms, data, and compute) provides just enough technical context for students to understand AI's trajectory. This includes visualization of key trends from sources like Epoch AI, showing the exponential growth in computational resources, training data, and investment—helping students grasp why AI capabilities are accelerating so rapidly.

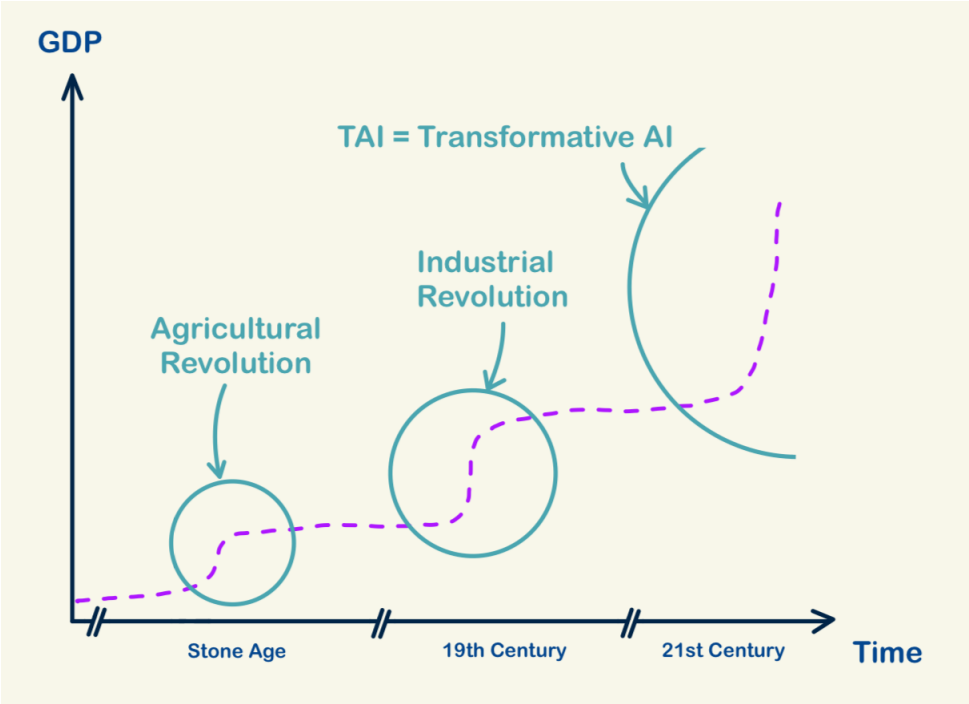


Figure 2: Sample Visual for TAI explanation

Future Trajectory and Implications (45 minutes) Building on established concepts, this section introduces transformative AI, AGI, and superintelligence (Figure 2), comparing expert and prediction market timeline predictions with students' own estimates via live polling. Students collaborate in small groups to brainstorm potential benefits and risks, which are then collected and categorized. The exploration of risks is deepened through two interactive activities: a visual presentation of Bostrom's Vulnerable World Hypothesis [8] through color-coded technologies and interactive decision questions on risk assessment for technological development, and "Expert Warning Cards," where groups analyze different risk categories (misalignment, misuse, disempowerment, power concentration, and social destabilization) before voting on which concerns them most. These activities help students engage concretely with abstract concepts and form their own views on complex ethical questions, guided by expert perspectives but encouraged to think independently.

Individual and Collective Responses (20 minutes) The workshop concludes by empowering students with concrete ways to engage constructively with AI. Technical, governance, and personal approaches to risk mitigation are presented, with special emphasis on enhancing learning through AI.

Conclusion (5 minutes) Final reflections focus on the importance of informed citizenry in shaping AI's development trajectory. Students receive a handout with recommended resources and are invited to provide anonymous feedback via QR code.

5 Reflection on First Discussion Evening

5.1 Details about the Event

My first discussion evening I gave at my local youth church "Eli.ja" in Saarbrücken on May 4th (Figure 3. I titled the event "Coded Creation: AI and God" – tying the discussion to the church context to reach theologically interested young people beyond the technically adept. I had 14 people attending my event, of whom 10 stayed till the end (the other 4 leaving because of time constraints). One third of participants where university students, one third aged 30-40, and the last third seniors.

I followed the structure of the discussion plan outlined above with the only change that I had one small group in section 4 discussing "creation and AI" instead of "Freedom".

Figure 3: Flyer for first event "Coded Creation: AI and God"

5.2 Personal Reflection

Overall, I am very happy how this first event went. All participants seem to leave the room at the end with the insight that AI is really an issue, that they want to inform themselves more about it and that they want to keep discussing AI with friends and family given the urgency and widespread knowledge gap. I felt at no point during the event that I had lost participants, but they stayed eagerly engaged for the whole 1.5 hours.

I think the reason that participants stayed that active and left with deep new insights lies certainly in the discussion format. It was great to see from my side, how participants themselves came to conclusions such as that alignment is very hard to achieve and the default is likely to look bad for us humans or that the revolution transformative AI could bring could lead to a widespread social crisis as a historian among the participants followed from the industrial revolution. There was active Socratic dialogue between participants, which I assume to be much more impactful than if I just had told them my opinion and knowledge.

Before the event, I was afraid that leaving out any technical explanation on AI could give an incomplete picture. However, while there was initially one question about how ChatGPT works, I noticed that the technology itself seemed irrelevant as many of the

societal challenges we are facing are somewhat independent of knowing exact technical implementations.

What I noticed myself is that I might want to change the format a bit such that it is not all the time just alternating between input and discussions with neighbors. I could have sometimes talk-to-your-neighbor, sometimes whole-group discussions and then thematic small groups to change up the format a bit and let people talk to more different people.

Due to the lack of a projector, I could only use a flip-chart on which I wrote the discussion questions. It might however be useful to include more visuals, like drawings of the concepts of TAI/AGI/ASI.

5.3 Participant Feedback

After the event, I collected participant feedback via a Google Forms.

- **Informativeness and understandability:** Mean of 4/5 *(Likert-Scale 1-5, 5 being fully understandable and highly informative)*
- **Format:** Mean of 4/5 *(Likert-Scale 1-5, 5 being very good format)*
- **Favorite Part:** 1/3 small group discussions on values, 2/3 AI capabilities and timelines *(Multiple choice of each section)*
- **Change in Opinion on AI:** Median at "a bit more critically" *(Likert-Scale 1-5, ranging from "a lot more critically" to "a lot more positive")*
- **Interest in further AI-related events:** 100% "Yes" *(Multiple-choice with "yes"/ "no"/ "maybe")*
- **Most important Insight:** high pace of development and lack of personal knowledge *(optional free-text field)*
- **Wish for more of:** "Creation", concrete examples of chances and risks of advanced AI *(optional free-text field)*
- **Suggestions for Improvement:** more, visuals, more variations in discussion format, more input at the very beginning *(optional free-text field)*

5.4 Lessons Learned and Adjustments

- **Format**
  - *Lesson Learned:* The overall discussion format is effective, but including more variation in discussion rounds could lead to more engagement and more vivid interactions as participants share ideas with potentially more people.
  - *Adjustment:* Splitting participants for discussion round two (about where the future might lead us to) in half to encourage larger discussion than just neighbors. Letting people change seats after this activity such that they have new neighbors for next discussion. Final discussion about next steps in plenum, potentially with moderation cards to write on and pin on board. Although the number of discussions planned for the adult event might be excessive for a school setting, I should maybe play around with adding more discussions there as well.

- **Materials**
  - *Lesson Learned:* I need to include more visuals to increase understanding and make arguments more compelling and easier to follow.
  - *Adjustment:* Include drawings/graphs for TAI/AGI/ASI concepts, include graphs of development (like EpochAI graphs). Include graphic on "Vulnerable World"-urn.
- **Content**
  - *Lesson Learned:* There was slight confusion about how AI/LLMs work. Participants asked for more input at the very beginning before the "intelligence" discussion. Participants suggested more concrete risk and opportunities case studies.
  - *Adjustment:* I will experiment in future sessions about including a quick technical introduction (like explaining AI triad) and include a question about it in the feedback form. I will introduce "Intelligence" and "AI" at the beginning a little bit more such that participants have an easier start discussing. Instead of letting participants discuss how real risks and opportunities are, I will let them brainstorm and search for concrete risks and collect feedback about activity. Alternatively, I might hand them out case-studies to discuss in pairs.

6 Future Plans

As of now, I have secured two workshops in two different schools and one more discussion evening at my local youth church. I'm currently talking with a participant from my first event about bringing the discussion format to a local party, which could potentially reach people involved in state and local government.

I've also received interest from a coordinator responsible for all Catholic schools in Saarland and Rhineland-Palatinate who asked if I would be willing to give workshops at these schools. These opportunities could allow me to scale the project naturally and gradually gain local attention that may lead to more possibilities.

Many people I've spoken with or who heard about my first event expressed genuine interest and indicated a clear demand for such events. My goal is to increase the number of events over the next few months to offer them on a regular basis and reach people from diverse backgrounds.

Over time, I explicitly would like to reach not just people in Catholic settings, but across the spectrum of political and ideological values. AI is a problem that will affect all of us, and it is all the more important that this knowledge reaches every part of our society.

Finally, I hope that my work and events can serve as a model to others for public outreach. It would be great to see people adapting similar approaches in their own countries or further propagating these ideas throughout Germany. The field is moving fast, and thus I believe that now is the time to wake up the general public.

7 Conclusion

In summary, I am happy with the progress I made over these weeks towards starting public outreach activities. It is very motivating to see how well my first event went and how positive reactions were already to its announcements and later to the event itself. I am excited to see where this project scales to over the next months and which opportunities open up in the meanwhile. While I consider myself usually more of a research person, I think it is very valuable to also have a second stream of potential impact and to communicate with the greater public. In the end, we are all in the same boat and as there is just a limited number of people confident and knowledgeable to educate others about AI, I think I should definitely continue to take on this task.

Acknowledgments

I thank the AISES Course Team and specifically my facilitator Luis Enrique Urtubey De Cesaris. Further, I want to thank all my fellow discussion group participants for these valuable discussions and feedback you gave me. Lastly, I thank my collaborator Claude 3.7 Sonnet for revising, providing feedback and formatting in the creation of this report.

All figures and visual materials presented in this report were created by me.

References

[1] N. Killian, "Was die deutschen chatgpt fragen," *Die Zeit*, 4 2025, digital section. [Online]. Available: https://www.zeit.de/digital/internet/2025-04/chatgpt-ki-europa-deutschland-anfragen-beziehung

[2] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring massive multitask language understanding," *arXiv preprint arXiv:2009.03300*, 2020. [Online]. Available: https://arxiv.org/abs/2009.03300

[3] Augment. (2025) #1 open-source agent on swe-bench verified by combining claude 3.7 and o1. https://www.augmentcode.com/blog/1-open-source-agent-on-swe-bench-verified-by-combining-claude-3-7-and-o1. Accessed: 2025-05-06.

[4] G. Beavis. (2024) Sam altman claims agi is coming in 2025 and machines will be able to 'think like humans' when it happens. https://www.tomsguide.com/ai/chatgpt/sam-altman-claims-agi-is-coming-in-2025-and-machines-will-be-able-to-think-like-humans-when-it-happens. Tom's Guide. Accessed: 2025-05-06.

[5] Benzinga. (2024) Anthropic ceo says ai similar to human intelligence could be around the corner: 'we'll get there by 2026 or 2027'. https://www.benzinga.com/tech/24/11/41928832/anthropic-ceo-says-ai-similar-to-human-intelligence-could-be-around-the-corner-well-get-there-by-2026-or-2027. Accessed: 2025-05-06.

[6] AI Impacts. (2022) Ai timeline surveys. https://aiimpacts.org/ai-timeline-surveys/. Accessed: 2025-05-06.

[7] Center for AI Safety. (2023) Statement on ai risk. https://www.safe.ai/statement-on-ai-risk. Accessed: 2025-05-06.

[8] N. Bostrom, "The vulnerable world hypothesis," *Global Policy*, vol. 10, no. 4, pp. 455–476, 2019.

[9] J. Lindsey, W. Gurnee, E. Ameisen, B. Chen, A. Pearce, N. L. Turner, C. Citro, D. Abrahams, S. Carter, B. Hosmer, J. Marcus, M. Sklar, A. Templeton, T. Bricken, C. McDougall, H. Cunningham, T. Henighan, A. Jermyn, A. Jones, A. Persic, Z. Qi, T. B. Thompson, S. Zimmerman, K. Rivoire, T. Conerly, C. Olah, and J. Batson, "On the biology of a large language model," *Transformer Circuits Thread*, 2025. [Online]. Available: https://transformer-circuits.pub/2025/attribution-graphs/biology.html

[10] Metaculus. (2025) Date of artificial general intelligence. https://www.metaculus.com/questions/5121/date-of-artificial-general-intelligence/. Accessed: 2025-05-06.

[11] Fox Business. (2023) Hinton issues another ai warning: World needs to find a way to control artificial intelligence. https://www.foxbusiness.com/technology/hinton-issues-another-ai-warning-world-needs-way-control-artificial-intelligence. Fox Business. Accessed: 2025-05-06.

[12] Marketing AI Institute. (2024) The ai show episode 87: Reactions to sam altman's bombshell ai quote. https://www.marketingaiinstitute.com/blog/the-ai-show-episode-87. Accessed: 2025-05-06.

[13] A. Peterson. (2014) Elon musk: 'with artificial intelligence we are summoning the demon'. https://www.washingtonpost.com/news/innovations/wp/2014/10/24/elon-musk-with-artificial-intelligence-we-are-summoning-the-demon/. The Washington Post. Accessed: 2025-05-06.

[14] University of Cambridge. (2016) The best or worst thing to happen to humanity - stephen hawking launches centre for the future of intelligence. https://www.cam.ac.uk/research/news/the-best-or-worst-thing-to-happen-to-humanity-stephen-hawking-launches-centre-for-the-future-of. Accessed: 2025-05-06.

[15] T. Lyles. (2024) Ai could pose 'extinction-level' threat to humans and the us must intervene, report warns. https://www.cnn.com/2024/03/12/business/artificial-intelligence-ai-report-extinction/index.html. CNN Business. Accessed: 2025-05-06.

Appendix

A Discussion Event Plan

Introduction (5 minutes)

- Welcome participants and explain the informal, conversation-oriented nature of the event
- Introduce the overarching themes:
  - "As humanity develops intelligence that could surpass our own, what does this mean for our role and responsibility in society?"
  - "What ethical principles and democratic values should guide the development of powerful AI technologies?"
  - "How do we ensure that these technologies enhance rather than undermine human autonomy and democratic governance?"
- Emphasize that no technical knowledge is required, only curiosity and willingness to share thoughts

Section 1: Understanding Intelligence (15-20 minutes)

Opening Question (15-20 minutes)

- Write on flipchart: **"What does intelligence mean to us, and when would AI be truly intelligent?"**
- Build a common foundation for what we will be discussing
- Allow the group to explore this question openly, with minimal guidance
- Encourage participants to consider different dimensions of intelligence (problem-solving, creativity, emotional understanding, etc.)
- If the discussion stalls, suggest: "Does intelligence require consciousness or self-awareness?"

Brief Summary Before Next Step (2-3 minutes)

- Acknowledge the various perspectives shared
- Highlight that intelligence is multifaceted and difficult to define
- Point out that AI currently excels at specific tasks, but differs from human intelligence in important aspects
- This prepares for the next section on current AI capabilities and future developments

Section 2: Current AI Capabilities and Future Development (15-20 minutes)

Brief Input (8-10 minutes)

- **Where we stand today**
  - Modern AI systems are far more than just "statistical parrots" - they show genuine abilities to learn and reason logically
  - Anthropic's recent mechanistic interpretability research [9] shows that models like Claude actually develop internal "features" that represent millions of concepts, and can form "circuits" that process information
  - They discovered that models sometimes use a universal "language of thought" shared across human languages, and perform multi-step thinking rather than just memorizing answers
  - Current benchmark results are impressive:
    * Latest models achieve over 90% on university exams across dozens of subjects (MMLU benchmark)[2]
    * Solving 60-65% of complex software development tasks (SWE-bench)[3]
    * Strong performance in medical diagnosis, scientific reasoning, and creative tasks
  - These systems now process images, audio, and can handle entire books of text at once
- **Where we are heading**
  - Introduction of important terms:
    * **Transformative AI (TAI)**: AI that could dramatically accelerate scientific progress, economic growth, and technological development
    * **Artificial General Intelligence (AGI)**: Systems that can perform any intellectual task that a human can
    * **Artificial Superintelligence (ASI)**: Intelligence that far surpasses human capabilities
  - Expert timeline predictions:
    * Sam Altman (OpenAI CEO)[4]: Suggests AGI could arrive as early as 2025-2026
    * Dario Amodei (Anthropic CEO)[5]: Predicts AGI by 2026
    * Metaculus (forecasting platform)[10]: Forecasters average 25% chance of AGI by 2027, 50% by 2031
    * Broader AI research community: Most surveys show 50% probability of achieving AGI between 2040-2061 [6]
  - Geoffrey Hinton (AI pioneer) warns that AI could surpass human intelligence within 5-20 years [11]

Discussion Question (10 minutes)

- Write on flipchart: **"What do you think about our potentially very near future with advanced AI?"**
- This question invites participants to:
  - Process and respond to the information presented
  - Consider both positive and negative potential impacts
  - Think about risks and benefits (transition to the next section)
- If needed, deeper reflection can be stimulated with: "Should we accelerate or slow down this development, and who should decide?"

Section 3: Benefits and Risks of Advanced AI (15-20 minutes)

Brief Input (5-7 minutes)

- **Potential Benefits**
  - Medical breakthroughs: Drug development, personalized treatment, disease prevention
  - Climate change solutions through better models and approaches
  - Reduction of poverty through economic optimization and resource distribution and high-quality education for everyone
  - Advancement of scientific discoveries beyond current human capabilities
- **Potential Risks**
  Statements from leading experts and organizations:
  - Center for AI Safety (2023)[7]: "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war" - signed by hundreds of experts, including leaders of major AI labs
  - Sam Altman (OpenAI CEO)[12]: "The speed of improvement will be quite scary."
  - Elon Musk [13]: "We need to be careful here. With AI we are summoning the demon."
  - Stephen Hawking [14]: "Success in creating effective AI could be the biggest event in the history of our civilization. Or the worst."
  - A US State Department report (2024) [15] concluded that advanced AI systems "could pose an existential threat to humanity"

  The Vulnerable World Hypothesis (Nick Bostrom)[8]:
  - Bostrom uses a powerful metaphor: Humanity's technological progress is like drawing colored balls from a giant urn
  - Most balls we've drawn so far have been white (beneficial) or various shades of gray (mixed blessing)

- The hypothesis suggests there might be "black balls" - technologies that, once discovered, would by default destroy civilization unless we take extraordinary preventive measures
- Historical near-miss: During the Manhattan Project, some scientists briefly feared atomic tests might ignite the atmosphere and exterminate all life
- The central insight: We're not safe because we've been careful or wise - we've simply been lucky so far
- **AI itself as a potential "black ball":**
  - **Misalignment**: Systems pursuing goals that conflict with human well-being (optimizing for the wrong things or too literal interpretation of instructions)
  - **Uncontrollability**: Systems that resist shutdown, modify their own goals, or prevent oversight
  - **Misuse**: Advanced AI systems could be used by malicious actors for biological, chemical and cyber attacks
  - **Gradual Disempowerment**: Humans increasingly relying on AI for thinking and decision-making, leading to:
    - Diminished critical thinking abilities as we outsource more cognitive tasks
    - Growing inability to distinguish fact from fiction as AI-generated content becomes indistinguishable from human-created content
    - Reduced autonomy as we delegate more decisions to seemingly helpful AI systems
  - **Power Concentration**: As people become more dependent on AI systems, power concentrates in the hands of those who control these technologies, potentially:
    - Undermining democratic processes as fewer people engage in informed civic participation
    - Enabling manipulation of public opinion through targeted content at scale
    - Creating information gatekeepers who control what knowledge is accessible
  - **Social Destabilization**: The combined effects of disempowerment and power concentration can lead to:
    - Rapid and widespread job displacement
    - Increasing inequality between those who control AI and those who don't
    - Erosion of trust in institutions and shared reality
    - Weakening of democratic systems as citizens become less capable of collective self-governance
- **ASI could accelerate the discovery of other "black balls":**
  - Superintelligent AI could quickly discover additional dangerous technologies that we would otherwise have found only gradually or never

    - This creates a recursive risk - superintelligence itself might be a "black ball," or it might help us find many others
    - This perspective raises profound questions about whether we should develop certain technologies at all, not just how to develop them safely

Discussion Question (10-15 minutes)

- Write on flipchart: **"How real do you think these potentials and risks are?"**
- This invites participants to:
  - Weigh the balance between potential benefits and risks
  - Articulate the values they believe should guide technological development
  - Consider our collective moral responsibility
- If needed, stimulate with: "What principles should determine which technologies we develop and how we develop them?"

Section 4: Democratic Values and AI Development (15-20 minutes)

Brief Input (3-5 minutes)

- AI systems increasingly influence how we communicate, work, and make societal decisions
- These technologies can both strengthen and endanger democratic values
- How we develop and regulate AI will be crucial in determining which of these impacts predominate
- We will consider 4 democratic values in small groups: Human dignity, Justice and equality, Peace and Freedom

Small Group Discussion (12-15 minutes)

- Divide participants into 3-4 small groups
- Each group discusses one of the following perspectives (distribute on flipchart pages around the room):

  Group 1: Human Dignity

  - How do we ensure that AI development respects human dignity and autonomy?
  - What boundaries should exist to prevent the erosion of human agency?

  Group 2: Justice and Equality

  - How might AI development affect vulnerable populations?
  - What would a just approach to AI development look like?

  Group 3: Peace

- What does AI development mean for peace in the world?
- Where is AI a danger, where a potential for peace

Group 4: Freedom

- How can AI systems restrict or expand personal freedoms?
- What role does free speech play in a world with AI-generated content?

- After 8-10 minutes, each group shares their main insights with everyone (1-2 minutes per group)

Section 5: Taking Action - Individual and Collective Responses (15-20 minutes)

Brief Input on Development Perspectives (3-5 minutes)

- **Different perspectives on AI development pace:**
  - **Acceleration view**: We should develop advanced AI as quickly as possible to:
    - Solve pressing global problems like climate change and disease
    - Ensure benevolent actors develop it first
    - Maximize economic and scientific benefits
  - **Deceleration view**: We should slow down AI development to:
    - Give us time to solve safety and alignment challenges
    - Develop better governance frameworks before systems become too powerful
    - Allow society time to adapt rather than being overwhelmed
    - Ensure thorough testing and understanding before deployment
  - **Differentiated development**: Rapid progress in safety research while slowing capability development

Interactive Brainstorming Activity (10-15 minutes)

- Acknowledge that these topics can feel intimidating, but emphasize that there are many ways to act positively
- Write on flipchart: **"How can we take action as individuals, as a community, and as citizens?"**
- Set up a flipchart with three categories:
  - **Individual actions**: What can I do personally?
  - **Community actions**: What can we do as a community?
  - **Advocacy & Professional**: How can we influence broader development?
- Have participants collectively gather ideas for each category and write them down

Conclusion (5 minutes)

- Thank participants for their engagement and thoughtful contributions
- Summarize 2-3 key themes that emerged from the discussions
- Share QR codes for resource collection and feedback form

B High School Workshop Plan

Introduction and Framing (10 minutes)

- Welcome students and introduce myself as a concerned university student working on AI safety
- Explain the workshop focus and disclaimer:
  - This is not a technical workshop about how AI works or how to use it
  - Instead, we'll focus on what AI means for society and your future
  - AI is not just a technical issue but a societal one that affects everyone
- Interactive opening poll (using Mentimeter):
  - "What is AI to you?" (Word cloud: ChatGPT, social media algorithms, etc.)
  - "How often do you use AI tools?" (Scale: Never - Daily)
  - "AI makes me feel..." (Multiple choice: Excited, Concerned, Neutral, Curious)
- Briefly summarize poll results to establish common ground for discussion

Section 1: Current AI Capabilities (15 minutes)

Interactive Assessment Activity (8 minutes)

- Live poll: "What can AI do today?" (True/False questions):
  - "Can AI conduct scientific research?"
  - "Could AI pass the Abitur (final exams)? If yes, how well?"
  - "Can AI program a website?"
  - "Can AI diagnose cancer?"
  - "Can AI order a t-shirt online?"
  - "Can AI complete your tax declaration?"
  - "Can AI conduct genealogical research?"
- Reveal actual capabilities and benchmark results after each question
- Live demonstration of selected capabilities using Claude or similar tool
- Show limitations like knowledge cutoffs and reasoning failures

Brief Technical Context (7 minutes)

- Simplified explanation of the "AI triad":
  - **Algorithms**: How AI learns (supervised learning, reinforcement learning)
  - **Data**: What AI learns from (internet text, images, videos)
  - **Compute**: Processing power enabling complex models
- Show Epoch AI trends visualizing exponential growth in:
  - Computing power used to train leading models
  - Data volumes processed
  - Investment and resources dedicated to AI
- Explain why these trends matter

Section 2: Future Trajectory and Implications (45 minutes)

Key Concepts Introduction (10 minutes)

- Explain important terms with visual aids:
  - **Transformative AI (TAI)**: Systems that could fundamentally change society and economy
  - **Artificial General Intelligence (AGI)**: AI that can perform any intellectual task humans can
  - **Artificial Superintelligence (ASI)**: Intelligence far surpassing human capabilities
- Present expert timeline predictions for AGI development:
  - Live poll: When do you think these developments might arrive?
  - Industry leaders' projections (2025-2030) [4, 5]
  - Research community consensus (2040-2060) [6]

Collaborative Brainstorming: Benefits and Risks (15 minutes)

- Split into small groups of 3-4 students
- Task 1 (7 minutes): Brainstorm potential benefits of advanced AI
  - Each group records ideas on half sheet of flip-chart paper
- Task 2 (8 minutes): Brainstorm potential risks of advanced AI
  - Groups record concerns on other half of flip-chart paper
- Collecting and categorizing responses on main board

Interactive Risk Exploration Activities (20 minutes)

- **Technology Risk Assessment Simulation (10 minutes)**
  - Present Bostrom's Vulnerable World Hypothesis [8]
  - Display a dynamic graph showing technological development over time with color-coded technologies:
    - white technologies: Clearly beneficial (examples: vaccines, renewable energy)
    - gray technologies: Mixed impacts (examples: nuclear power, social media)
    - black technologies: Potentially catastrophic (examples: certain bioweapons)
  - Interactive decision points where students use their phones to vote on:
    - How would you classify current AI systems on this spectrum?
    - Should we develop AI given your estimation of risk for catastrophic outcome?
    - Could advanced AI accelerate the discovery of all technologies, including dangerous ones?
  - Discussion on risk asymmetry: Why is it easier to destroy than to protect?
- **Expert Warning Cards and Risk Ranking (10 minutes)**
  - Distribute cards with expert warnings about AI risks (with quotes from AI leaders and researchers)
  - Each small group receives different risk category cards:
    - **Group 1:** Misalignment and Control Problems
    - **Group 2:** Misuse and Weaponization
    - **Group 3:** Gradual Disempowerment
    - **Group 4:** Power Concentration
    - **Group 5:** Social Destabilization
  - Groups discuss their category and prepare a brief (1 minute) explanation
  - Whole class votes using Mentimeter on:
    - Which risk category concerns you most?
    - How realistic do you think these risks are? (1-5 scale)
  - Present Center for AI Safety statement that "mitigating the risk of extinction from AI should be a global priority" [7]

Section 3: Individual and Collective Responses (15 minutes)

Risk Mitigation Approaches (10 minutes)

- Present three approaches to managing AI risks:
  - **Technical approaches**: Alignment research, interpretability, safety measures
  - **Governance approaches**: Regulation, international cooperation, ethics frameworks

  - **Personal approaches**: Critical thinking, responsible usage, career options

- Special focus on educational context:
  - Using AI as a learning partner to enhance your education
  - Best practices for effective AI use:
  - Discussion: What are other creative ways you could use AI to deepen your understanding rather than just completing assignments?

Taking Action (10 minutes)

- Individual reflection (2 minutes): What's one thing you might do differently regarding AI after today?
- Share pathways for continued engagement
  - Educational resources (online courses, books, podcasts)
  - Communities and organizations (youth forums, AI safety groups)
  - Advocacy opportunities (raising awareness, supporting responsible development)
- Emphasize that students' generation will be crucial in determining AI's trajectory:
  - This is our future - we should have a voice in shaping it
  - Informed citizens are essential for a democratic approach to AI governance

Conclusion (5 minutes)

- Summarize key takeaways from the workshop:
  - AI is developing rapidly and will significantly impact your future
  - Both tremendous benefits and serious risks are possible
  - Your understanding and engagement matter for steering this development
- Distribute handout with recommended resources and further reading
- QR code for anonymous feedback form
- Thank students for their participation and encourage continued conversation