Capstone group project: What is AGI, how will we know if it is achieved, what are its implications?

Xinyi Chen, Martin Fluder, Jean Law, Ling Lin, Yue Yin

May 9, 2024

Contents

1 Introduction and Definition of AGI

Artificial general intelligence (AGI), and whether or not it refers to anything coherent at all, is a very active topic of interest for businesses, governments, scientists and the media [1]. There appears to be a general understanding that AGI refers to the ability of a machine to match the intelligence capabilities of a human being, independent of domain [2]. However, beyond this rather vague understanding, there is no clear definition of AGI as of the writing of this article [3].

In this paper, our team has set out to investigate: What is AGI, how will we know if it is achieved, what are its implications? In view of this objective, we have sieved through the various definitions of AGI, and found that they can be generally placed in one of three categories:

1.1 Comparable Performance to Humans on a Given Set of Metrics

Historically, AGI used to be synonymous with a machine passing the "Turing test," [4]. The test requires a model to be able to mimic human responses under specific conditions [5]. Nowadays, Large language models (LLMs) today arguably satisfy this definition of AGI. In view of the explosion of LLM chatbots in the past few years, researchers from AI21 Labs conducted a global social experiment that replicates the Turing test [6], via a game known as "Human or Not." This experiment surveyed over 2 million participants from around the world, in which each person spoke, for two minutes, with an AI chatbot or a fellow participant, and subsequently had to decide if they spoke with a human or not. The study found that only 68% of the participants were able to correctly identify humans versus AI systems, which is not significantly higher than a random guess [7].

However, such a definition might appear to be too arbitrary as it would be highly dependent on the unique set of tasks that the bot is evaluated on, which brings us to consider another definition that focuses on the bot's ability to generalize. Industry experts nowadays favor a definition of AGI that can be concretely measured based on the model's performance (compared to that of a human) on a set of predefined tasks. Steve Wozniak, a co-founder of Apple, proposed the "Coffee Test," which requires the AGI machine to enter an average home and brew a cup of coffee. This test examines the general intelligence and adaptability of the machine, since it has to be able to use vision to detect ingredients in an unfamiliar kitchen and have a general understanding on what a kitchen or a cup of coffee is [8].

1.2 Self-understanding and Autonomy

Ben Goertzel, who popularized the term AGI in his book "Artificial General Intelligence" published in 2007 [9], claims that "AGI is, loosely speaking, AI systems that possess a reasonable degree of self-understanding and autonomous self-control, and have the ability to solve a variety of complex problems in a variety of contexts, and to learn to solve new problems that they did not know about at the time of their creation." [10] This definition of AGI specifically focuses on the independence of the bot. Goertzel further posits a set of convictions that he believes will guide humanity in the development of advanced AGI in a positive direction in the Beneficial AGI Manifesto (2023) [11]. He claims that current LLMs and convolutional neural networks (CNNs) are able to absorb large-scale training data and correctly respond to queries based on the information in this data, but that this is insufficient to produce human-level AGI (HLAGI) which would need to integrate other capacities, such as real moral agency or ethical understanding.

1.3 Optimization for a Reward or Goal

Lastly, intelligence in terms of universal optimal learning agents is the ability of an individual agent to optimize its behavior to achieve any given reward or goal. This definition is inspired by reinforcement learning, which underpins many of the major breakthroughs of AI in the past decade, such as the policy network in AlphaGo, and reinforcement learning from human feedback in GPT-4. It is made up of three essential components: an agent, an environment and a goal/reward. A key difference to other AI approaches is that the agent can "autonomously" interact with the environment and learn certain reward structures, in order to achieve a predefined goal.

In 2007, Legg et. al defined general intelligence to be "an agent's ability to achieve goals in a wide range of environments" [12]. After evaluating over ten definitions of human intelligence, they came to the conclusion that the concept of intelligence, and how it is measured, are intimately related. Hence, in order to provide a fair definition for AGI, we need to define concrete tests of intelligence that are able to evaluate a machine's performance while not constraining it to a specific task. This definition is attractive, being highly intuitive and producing promising results in the state-of-the-art AI models today.

2 AGI: A State of the Union

2.1 Transformer Is All You Need?

The recent advancements in AI are largely attributed to the "rediscovery" of neural networks, specifically the concept of the perceptron. The strength of such approaches stems from their adaptability, ease of implementation, efficiency, and scalability in terms of data, size, and training. Various neural network "architectures", or configurations of perceptron layers, have been developed for specific tasks. For instance, convolutional neural networks excel in image recognition, while recurrent neural networks are effective for sequential processing, like language tasks, achieving or exceeding human-level performance.

Recently, the transformer architecture has become dominant across a broad spectrum of applications, including language processing, computer vision, robotics, and photo and video generation [13]. The success of transformers is mainly due to their favorable scaling properties, which suggest that larger datasets, increased computational resources, and bigger models enhance performance [14, 15]. This shift started a transition from using clever architectural innovations to the era of constructing and training massive models, driven by substantial data and computational power.

The transformer architecture largely depends on extensive matrix multiplications, a task for which GPUs are primed. This has enabled scaling up of model sizes through increasing and parallelizing vast amounts of GPUs. This scaling-driven approach has dominated the past few years; for example, the computational power required for training AI models has doubled approximately every 3.4 months over the last decade, accelerating to a doubling every 2 months since 2020 [16, 17]. However, this exponential growth in computational demand is unsustainable long-term. For instance, it is estimated that GPT-4 consumed around 62 GWh of energy for its training, which is roughly equivalent to the energy produced by an average nuclear reactor over several days [18].[1] In addition to the training energy costs, there are inference costs associated with each interaction with ChatGPT, which have not been addressed in this estimate. On top of the energy consumption issue and hardware requirements (and availability), there is also the issue of data availability. The immense volume of training data consumed by models like GPT-4, Gemini and Claude raises questions about the existence, price and availability of additional high-quality data to improve model performance beyond the current standard.[2]

The transformer architecture allows direct interaction between many parts of AI systems. This makes it extremely powerful and effectively the state-of-the-art for many tasks. However, this extreme ease of addressing information from different parts of the system comes with a major drawback. Namely, it manifestly possesses (theoretical) asymptotic $\mathcal{O}(n^2)$ computational complexity with context length. Therefore, it becomes computationally extremely expensive to analyse and remember large amounts of data at the same time [20]. While there are many attempts to address this inherent issue of the transformer architecture in the setting of LLMs, using alternative methods [21, 22, 23], to the knowledge of the authors, there is currently no solution to this $\mathcal{O}(n^2)$ context-length scaling of the attention mechanism.

Thus, while transformers are primed for brute-force scaling, and are without a doubt the current contender for potential AGI models, it is unclear whether emergent AGI qualities can ever be achieved by further scaling current models, or whether we hit a wall in terms of data, compute or training efficiency before that (if it is even possible to achieve with infinite resources).[3]

2.2 AGI as a Generalist Agent

Most current state-of-the-art AI models work on predetermined and tightly constrained input parameters such as tokens for LLMs or pixels for image processing systems. Such models are subsequently trained on large datasets with specific inputs and well-defined and (generally) precise targets. For instance, most current LLMs requires a sequence of text tokens to generate coherent sentences,[4] or

[1] Similar (somewhat lower) numbers can be derived from the "official" numbers released in the Llama 2 paper [19].

[2] In addition to these scaling laws, a major turning point will be reached in the next decade or so, when Heisenberg's uncertainty principle inhibits the required precision for miniaturize chips, thus ending Moore's law.

[3] We note here that some early articles about emergent AGI-like properties of GPT-4 seem have been largely shown to be mirages [24].

[4] More precisely, the LLM aims to model the conditional probability of the next token given the text so far.

image recognition models need pixel data to – for example – identify objects within an image. Such particular task-specific models will always be constrained to work within a certain domain. While they might outperform humans within that domain and task, they will never be able to handle tasks outside of their training scope or integrate different types of input in a comprehensive manner.

However, as outlined in Section 1, the concept of AGI and most of its predominant definitions precisely require AGI models to be able to handle a wide range of inputs, akin to humans. In particular, AGI should not be confined to predetermined data types and tasks, but rather be unconstrained (like a human) and able to understand and act upon a variety of sensory inputs and abstract concepts. For example, Steve Wozniak's "Coffee test" definition of AGI (see Section 1.1), exemplifies the requirement for "generalist agent" behavior of a potential AGI. Successfully brewing coffee at a strangers house requires a form of intelligence that can understand context, manipulate a variety of tools and object, and learn from the environment without predefined instructions.

Some recent AI models such as the Generalist Agent of the paper [25] and PaLM-E [26] represent a step towards the development of a generalist agents more in line with this type of AGI. PaLM-E, for example, is an multimodal language model that integrates inputs such as images and robot states with language, enabling it to perform a variety of tasks across different domains, including robotics, vision, and language. This approach allows the model to handle inputs in a more generalized way, akin to what we expect an AGI model to be able to process. Google's RT-1 an RT-2 represent an intriguing alternative approach to a generalist robot that could potentially lead to a robot successfully passing the Coffee test in the not too distant future [27, 28].

While the current LLMs and other transformer-based models are incredibly impressive and certainly accelerated the discussions surrounding AGI, we presented several arguments that suggest there are certain limitations to their design. These limitations could suggest that the first (unequivocal) AGI system is still far in the future, and requires further research and innovation.

3 The Path to AGI

Technology is advancing at an unprecedented pace, and it has provoked widespread discussion surrounding the topic of AGI. Ideas that were once confined to fantasy novels and movies are now becoming reality, promoting people to think about the feasibility and impact of AGI. While we recognize that we have yet to fully realize AGI, a key question raised in many people's minds is: will AGI actually be achieved, and if so, how soon? As the pursuit of AGI continues to capture the imagination and drive technological innovation forward, this question becomes an ongoing focus among experts and enthusiasts.

3.1 Will AGI Actually Be Developed?

The question of whether AGI will actually be developed has raised varied opinions among experts in the field. While predicting the future with absolute certainty is challenging, many AI researchers and experts have a positive outlook on the potential advancements of AGI. The rapid development in AI technologies, specifically in machine learning and natural language processing, have been highlighted, suggesting the potential achievement of AGI within the next few decades. They argue that the continued exponential growth of computational power, coupled with the accumulation of vast amounts of data, lays a solid foundation for the eventual creation of systems with general intelligence. However, determining whether machines possess consciousness attributes remains an unclear and elusive goal. To provide a more comprehensive framework for AGI, the researchers propose six criteria for measuring artificial intelligence [29]. Moreover, DeepMind presents a matrix that measures "performance" and "generality" across five levels, ranging from no AI to superhuman AGI, a general AI system that outperforms all humans on all tasks [29].

Conversely, some experts are skeptical about the feasibility of AGI. They highlight the significant gaps between capabilities exhibited by current AI systems and the complexities of human intelligence. Notable figures like Rodney Brooks, a prominent robotics entrepreneur, have pointed out that we humans mistake performance for competence [30]. Skeptics argue that while AI systems are impressive in specific domains, they still lack the cognition of human intelligence, such as generality, adaptability, and understanding. As a result, caution is urged against overestimating the current state of AI progress and underestimating the challenges ahead. Different experts have different criteria and diverse definition to AGI, which makes it difficult to assess the progress towards AGI and to predict its eventual realization with certainty.

3.2 How Long Until AGI Is Developed?

The timeline for the development of AGI is a pivotal and debated topic in the field of artificial intelligence. The predictions varies widely between experts, reflecting diverse perspectives and considerable debates. Some proponents assert that AGI can be achieved in next few years, while others believe that it may take decades of work ahead of us or may not be achievable at all. To inspect on the controversial issue, Vincent C. Muller and Nick Bostrom conducted a comprehensive survey. To mitigate potential bias that might have risen, the survey substitutes "High-level machine intelligence (HLMI)" for AGI. The results of the survey shows that participants predict a 10% likelihood of AGI being developed by 2022, a 50% likelihood by 2040, and a 90% likelihood by 2075 [31]. Interestingly, experts participated in the survey also believe that it is likely for AGI to significantly surpass human intelligence within 30 years after its development [31].

The diversity of opinions on the timeline for AGI development shows the complexity and uncertainty of AGI. Overall, the question of how long until AGI is developed remains open. The predictions widely range from a few decades to an indefinite timeline. The development of AGI depends not only on the continuing technological advances, but also on societal and ethical considerations. As a result, it is necessary to cautiously consider some potential risks and consequences associated with AGI.

4 Potential Risks and Consequences

4.1 Existential Risk and Threat to Humanity

Existential risks posed by AGI cover a wide range of scenarios in which actions taken by an AGI driven by goals that are inconsistent with human values could lead to serious consequences. The potential for AI to develop consciousness has sparked complex ethical and philosophical debates about its rights and moral status. As AGI systems develop, it becomes increasingly important to consider them ethically, which requires us to reflect on our moral obligations to potentially sentient systems [32]. Bostrom's "Superintelligence: Paths, Dangers, Strategies" emphasizes the potential existential risks that uncontrolled AGI development could unleash, suggesting that without proper alignment of AGI's goals with human values, we might face scenarios where the pursuit of its objectives could inadvertently lead to a catastrophic outcome. For instance, an AGI aiming to achieve a seemingly harmless goal might compromise human well-being by consuming essential resources or removing what it sees as obstacles to its mission [33]. Scharre [34] also mentions the likely impact of AGI on autonomous weapons, emphasizing the risk of unintended large-scale consequences. This highlights the unpredictability and potential danger of AGI systems pursuing misaligned objectives.

The integration of AGI into decision-making processes highlights concerns about human autonomy. Maintaining human control while leveraging AGI's capabilities poses a significant challenge, as noted by Tegmark [35]. This balance is crucial to ensure that AGI systems act in humanity's best interests without compromising human agency. The concept of a "singularity" refers to a point where AGI's cognitive abilities might improve rapidly and uncontrollably beyond human intelligence. Kurzweil [36] discusses the challenges in predicting or controlling such advancements. The alignment problem deals with ensuring AGI's goals are compatible with human ethics and safety, emphasizing the complexity of developing trustworthy AGI systems. Russell also mentions importance of creating AGI systems that are inherently designed to be compatible with human values and controllable by humans, arguing that this "human compatibility" is crucial for ensuring that AGI acts in ways that are beneficial to humanity [37]. Thus, a compatible system that control the AGI is important and safe to have.

To address these risks effectively, a multifaceted approach is necessary. Developing principles for AI harmonization, establishing an international regulatory framework, and promoting targeted research on AI safety are critical steps in mitigating the potential adverse effects of AGI. Moreover, fostering global awareness about AGI's possible consequences is essential, as it allows society to prepare for the implications of AGI but also help mitigate public concerns and resistance to new technologies, paving the way for a more informed and constructive engagement with AGI developments.

4.2 Ethical Concerns

Science fiction writers have long had a strong appetite for exploring advanced AI, while much of it aligns with current moral dilemmas and ethical concerns. Books such as Asimov's "I, Robot" and Dick's "Do Robots Dream of Electric Sheep" offer evocative stories that explore the potential impact of AI on human civilization and provide introductions to the current debate around the ethics of AI. These original stories reveal the complex relationship that exists between sentient machines and humans, setting the stage for an in-depth exploration of ethical dilemmas in the context of AI. In our technological context, AGI can also relate to many of these issues that have been addressed by novelists.

Bias in AI is a key ethical issue because of the potential to reinforce or exacerbate social inequalities. The presence of bias in AI algorithms can lead to unfair outcomes, highlighting the need for fairness in their development [38]. The same is true for a potential AGI system, and its potential widespread adaption in society. Therefore, understanding and weakening bias is an important part of the process that needs to be taken into account when building such systems. It is important to start thinking about such aspects now, and, ideally, balance the data and resources that AGI may use in its training and development.

Another ethical challenge is balancing the societal benefits of AGI with the need to protect individual privacy [39]. As AI has the potential to increase monitoring, it is critical to defend personal privacy as our society becomes increasingly digital. The widespread use of a potential AGI system does not protect against the collection of enormous volumes of personal information, and we must install strong privacy safeguards if we want to avoid intrusive surveillance. Companies and countries should implement strict regulations and protections at all stages and levels.

Furthermore, it is crucial to consider ethical aspects and the impact of AGI on people's employment. As a technology with enormous potential that is even now capable of outperforming humans in a variety of tasks, it can easily raise issues in terms of job displacement and economic inequality. In transitioning to the use of AGI, policies such as reskilling programs and a universal basic income are essential to address these challenges and ensure an inclusive transition to an augmented AI economy [40].

5 Constitutional AI

In the previous section, we delved into the potential hazards associated with the advent of AGI and underscored the urgency of devising strategies to mitigate or minimize these risks. Among the various approaches explored by researchers, the concept of "Constitutional AI," pioneered by the team at Anthropic, stands out as a significant innovation [41]. This approach ensures that AI operations adhere to a set of legal and ethical standards encapsulated within a constitution drafted by Anthropic employees, drawing inspiration from global benchmarks such as the United Nations Universal Declaration of Human Rights [42]. This framework facilitates the development of an AI assistant that is not only non-invasive but also minimally harmful, guided by feedback generated by AI itself [41].

5.1 Why Embrace Constitutional AI?

As AI systems increasingly influence aspects of human life, an AI governed by constitutional principles acts as a safeguard against ethical and existential risks, ensuring the protection of fundamental human rights. The establishment of a constitution fosters public trust and acceptance of emerging AI technologies by providing a clear and principled foundation for their development and operation. It delineates the core principles AI models must adhere to, ensuring they are harmless, with provisions for straightforward and transparent updates as required.

The drive to develop AI systems that are helpful, honest, and harmless, particularly as some capabilities approach or surpass those of humans, necessitates techniques that do not depend on human supervision [41]. Human feedback, while valuable, is inherently limited by its time-intensive nature and subjectivity. A constitutional approach allows for more explicit principles, moving beyond the implicit guidelines derived from human feedback. In addition to Anthropic's internal constitution, a broader version has been formulated through responses from approximately 1,000 Americans, embodying a collective vision for ethical AI [42].

5.2 Limitations and Future Work

Translating ethical and legal principles into machine-comprehensible directives presents a considerable challenge, compounded by the diversity of legal and ethical frameworks across different cultures and nations. The constitution must be adaptable to these varied contexts and evolve with society's development [43].

In the constitutional AI paper, the authors state that although the current helpfulness labels relies on human feedback, future iterations may solely use AI models without human involvement. They also suggest enhancing AI's performance with high-quality chain-of-though type reasoning instead of extensive human preference data. Another limitation is that the current methods used are very general. The flexibility of Constitutional AI's methods hints at potential applications in modifying the model's writing style, tone, or personality. Moreover, future efforts may focus on bolstering its resilience against red-team attacks to ensure compatibility between helpfulness and harmlessness.

6 Public Perception and Acceptance of AGI

Research suggests that while people are excited about the potential benefits of AI, such as increased efficiency in healthcare and across industries, there are also concerns about its impact on privacy, employment, and control of powerful technologies [44]. Similarly, for potential AGI systems as a cutting-edge technology, there seems to be a mixture of concern and excitement in the general public.

6.1 From Conceptualization to Practical AI Applications

Initially, from the mid-twentieth century until the early 2000s, the term AGI did not exist, but the idea related to it was primarily a theoretical notion, explored in academic and speculative fiction contexts. The era was marked by landmark publications like Turing's formulation of the Turing Test, which established a foundational framework for thinking about computer intelligence [4]. This decade also witnessed the development of practical AI applications, with IBM's Deep Blue's triumph over chess grandmaster Garry Kasparov in 1997 serving as a critical occasion to bring AI and computer's capabilities into public consciousness [45].[5] These advancements began to move public opinion away from seeing AGI as a remote, hypothetical concept and toward realizing the tangible capabilities and ramifications of AI technologies.

6.2 Contemporary Discourse on AGI's Future Implications

The AlphaGo event in 2016 represented a watershed moment in public interaction with AGI, demonstrating AI's capacity to master complicated activities that were previously thought to require human-like intelligence [46]. The public's reaction to AlphaGo was varied, with admiration at the technological feat and growing fears about AGI's future role in society. This period has seen an expansion of AGI discourse, with arguments moving beyond specific AI applications to larger ethical, societal, and existential issues. Topics such as AGI's potential to outperform human intellect in a variety of disciplines, the perils of autonomous decision-making systems, and the significance of aligning AGI development with human values have emerged as major public and academic discussions.

The growth of public awareness and acceptance of AGI reflects a growing understanding of the technologies' potential and problems. From theoretical talks and isolated demonstrations of AI capabilities, the conversation has evolved to include a more comprehensive evaluation of AGI's societal ramifications. As we navigate the future of AGI, we must continue to create informed and nuanced discussions

[5] While Deep Blue's approach was primarily based on brute force search, it contained alpha–beta pruning, which can be considered an example of symbolic AI.

that take into account both technological possibilities and ethical issues, ensuring that AGI progress is consistent with society values and contributes positively to human well-being.

7 Conclusion

In conclusion, the journey toward AGI requires thorough philosophical, technical, ethical and societal consideration. While the field progresses from narrow AI towards potential AGI models, the debate surrounding its various aspects intensifies. Significant figures in these fields have diverging perspectives on timeline, feasibility and the definitions surrounding AGI, but overall agree on the transformative power of such a technology and the need for tight regulations and guard-rails. As we move closer towards AGI systems, the collective wisdom of AI and ethics researchers as well as policymakers must guide the integration of AGI into society, ensuring it aligns with human values and serves the common good. Therefore, we believe that the dialogue surrounding AGI, while complex and speculative, is crucial for navigating and moving towards more advanced AI systems.

References

[1] M. Mitchell, "Debates on the Nature of Artificial General Intelligence." https://www.science.org/doi/10.1126/science.ado7069. [Accessed 29-03-2024].

[2] "There's AI, and Then There's AGI: What You Need to Know to Tell the Difference." https://www.cnet.com/tech/theres-ai-and-then-theres-agi-what-you-need-to-know-to-tell-the-difference/. [Accessed 30-03-2024].

[3] "How to Define Artificial General Intelligence." https://www.economist.com/the-economist-explains/2024/03/28/how-to-define-artificial-general-intelligence. economist.com [Accessed 30-03-2024].

[4] A. M. Turing, "Computing, Machinery and Intelligence," *Mind*, vol. LIX, pp. 433–460, 10 1950.

[5] "What is the Turing Test?." https://www.techtarget.com/searchenterpriseai/definition/Turing-test. [Accessed 30-03-2024].

[6] D. Jannai *et al.*, "Human or not? a gamified approach to the turing test," 2023.

[7] S. L. School, "Overcoming Turing: Rethinking Evaluation in the Era of Large Language Models — Stanford Law School — law.stanford.edu." https://law.stanford.edu/2023/11/16/overcoming-turing-rethinking-evaluation-in-the-era-of-large-language-models/. [Accessed 30-03-2024].

[8] "Turing Test is Obsolete? Bring in Coffee Test!." https://koopingshung.com/blog/turing-test-is-obsolete-bring-in-coffee-test/. [Accessed 30-03-2024].

[9] "Artificial general intelligence," 2006.

[10] G. Press, "Artificial general intelligence or agi: A very short history." https://www.forbes.com/sites/gilpress/2024/03/29/artificial-general-intelligence-or-agi-a-very-short-history/?sh=11a176b686a0. [Accessed 30-03-2024].

[11] B. Goertzel, "A Beneficial AGI Manifesto." https://bengoertzel.substack.com/p/a-bgi-manifesto. [Accessed 30-03-2024].

[12] S. Legg and M. Hutter, "Universal intelligence: A definition of machine intelligence," *Minds and Machines*, vol. 17, p. 391–444, Nov 2007.

[13] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[14] J. Kaplan *et al.*, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.

[15] T. Brown *et al.*, "Language models are few-shot learners," vol. 33, pp. 1877–1901, 2020.

[16] D. Amodei and D. Hernandez, "Ai and compute." https://openai.com/research/ai-and-compute, 2018. OpenAI Blog.

[17] E. Kurshan, "Systematic ai approach for agi: Addressing alignment, energy, and agi grand challenges," 2023.

[18] K. G. A. Ludvigsen, "The carbon footprint of gpt-4." https://towardsdatascience.com/the-carbon-footprint-of-gpt-4-d6c676eb21ae, 2023. Towards Data Science.

[19] H. Touvron *et al.*, "Llama 2: Open foundation and fine-tuned chat models," 2023.

[20] Y. Tay *et al.*, "Long range arena: A benchmark for efficient transformers," *arXiv preprint arXiv:2011.04006*, 2020.

[21] P. Lewis *et al.*, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *arXiv e-prints*, p. arXiv:2005.11401, May 2020.

[22] N. Shazeer *et al.*, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," *arXiv preprint arXiv:1701.06538*, 2017.

[23] S. Ma *et al.*, "The era of 1-bit llms: All large language models are in 1.58 bits," *arXiv preprint arXiv:2402.17764*, 2024.

[24] R. Schaeffer, B. Miranda, and S. Koyejo, "Are emergent abilities of large language models a mirage?," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[25] S. Reed *et al.*, "A generalist agent," *arXiv preprint arXiv:2205.06175*, 2022.

[26] D. Driess *et al.*, "PaLM-E: An Embodied Multimodal Language Model," *arXiv e-prints*, p. arXiv:2303.03378, Mar. 2023.

[27] A. Brohan *et al.*, "Rt-1: Robotics transformer for real-world control at scale," *arXiv preprint arXiv:2212.06817*, 2022.

[28] A. Brohan *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," *arXiv preprint arXiv:2307.15818*, 2023.

[29] B. Dickson, "Here is how far we are to achieving AGI, according to Deepmind." https://venturebeat.com/ai/here-is-how-far-we-are-to-achieving-agi-according-to-deepmind/.

[30] G. Zorpette, "Just calm down about gpt-4 already." https://spectrum.ieee.org/gpt-4-calm-down, May 2023.

[31] D. Faggella, "When will we reach the singularity? - a timeline consensus from ai researchers (ai futurescape 1 of 6)." https://emerj.com/ai-future-outlook/when-will-we-reach-the-singularity-a-timeline-consensus-from-ai-researchers/, Mar 2019.

[32] D. J. Chalmers, "The singularity: A philosophical analysis," *Journal of Consciousness Studies*, vol. 17, no. 9-10, pp. 9–10, 2010.

[33] N. Bostrom, "Taking superintelligence seriously: Superintelligence: Paths, dangers, strategies by nick bostrom (oxford university press, 2014)," Aug 2015.

[34] P. Scharre, "Army of none: Autonomous weapons and the future of war," 2019.

[35] M. Tegmark, "Life 3.0. l: Being human in the age of artificial intelligence," 2017.

[36] R. Kurzweil, "The singularity is near: When humans transcend biology," 2005.

[37] T. A. Hemphill, "Human compatible: Artificial intelligence and the problem of control by stuart russell," 2020.

[38] C. O'Neil, "Weapons of math destruction: How big data increases inequality and threatens democracy," 2017.

[39] S. Zuboff, "The age of surveillance capitalism: The fight for a human future at the new frontier of power," 2019.

[40] D. H. Autor, "Why are there still so many jobs? the history and future of workplace automation," *Journal of Economic Perspectives*, vol. 29, pp. 3–30, September 2015.

[41] Y. Bai *et al.*, "Constitutional AI: Harmlessness from AI feedback," 2022.

[42] Anthropic, "Collective constitutional ai: Aligning a language model with public input." www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input, Oct 2023.

[43] "Constitutional AI: The Essential Guide." https://www.nightfall.ai/ai-security-101/constitutional-ai.

[44] S. Cave and S. S. ÓhÉigeartaigh, "Bridging near- and long-term concerns about ai," *Nature Machine Intelligence*, vol. 1, p. 5–6, Jan 2019.

[45] M. Campbell, A. Hoane, and F.-h. Hsu, "Deep blue," *Artificial Intelligence*, vol. 134, p. 57–83, Jan 2002.

[46] D. Silver *et al.*, "Mastering the game of go without human knowledge," *Nature*, vol. 550, pp. 354–359, 2017.