The Path to AGI Goes through Embodiment

Cheston Tan[*], **Shantanu Jaiswal**[*]

Centre for Frontier AI Research (CFAR) and Institute of High Performance Computing (IHPC)
Agency for Science, Technology and Research (A*STAR)
cheston_tan@cfar.a-star.edu.sg, jaiswals_shantanu@ihpc.a-star.edu.sg

Abstract

Recent advances in large language models have raised the question of whether these language models alone could lead to artificial general intelligence (AGI). In this short position essay, we argue that embodiment is not only required for achieving AGI, but also that embodiment is the key to convincingly demonstrate AGI capabilities. There is no single widely-accepted, objective test for AGI, so therefore whether a system has achieved AGI is a subjective judgement. We argue that a language-only system or one that cannot demonstrate success in the real world would not be convincing.

Introduction

Recent advances in large language models (LLMs) have led some people to believe that LLMs alone can eventually give rise to AGI – Artificial General Intelligence (Bubeck et al. 2023; Michael et al. 2022; Zhang et al. 2023). On the contrary, this short position essay argues the opposite – that AGI not only requires embodiment (Deitke et al. 2022; Duan et al. 2022b), but that **in the absence of a widely-accepted "acid test" for AGI, embodied intelligence is the key towards convincing demonstrations of AGI**.

It is not obvious that AGI even requires embodiment, much less being key to AGI. In addition to the papers mentioned earlier, which posit that LLMs by themselves are sufficient for AGI, others with less rosy views of LLMs do not necessarily believe strongly in embodiment either. For example, in LeCun's work on world models (LeCun 2022; Dawid and LeCun 2023), there is little to no mention of embodiment as being needed for human-level AI.

Other approaches or efforts that people believe could lead to AGI include deep reinforcement learning (RL) that has demonstrated human-level or superhuman-level performance on Atari (Mnih et al. 2013, 2015) and strategy games (Berner et al. 2019; Heinrich and Silver 2016). While the game agents follow the laws of physics as variously defined for each game (e.g. not allowed to pass through walls, cannot teleport except when going to the next game level, etc.), these game "universes" are generally not considered or emphasized to be embodied. Proponents of RL leading to AGI (Arel 2012; Silver et al. 2021) also do not generally place importance on embodiment, but more that RL is a general learning approach that can lead to AGI given sufficiently complex tasks and sufficient resources.

Working Definitions and Unstated Assumptions

So far, we have used the terms AGI and embodiment without any attempt to define them. For the purposes of this paper, we adopt the definition of **AGI** to be "*an autonomous machine exhibiting general-purpose learning and reasoning capabilities that equal or surpass human-level capabilities across diverse tasks and in a broad range of environments or situations*" (Chollet 2019; Goertzel 2014; Pennachin and Goertzel 2007), and **embodiment** to be "*an agent's engagement with a physical or virtual environment through sensorimotor interactions and experiences, mirroring a human's experiential understanding of the world*" (Chrisley and Ziemke 2006; Glenberg 2010; Longo et al. 2008).

Definitions and discussions of AGI often make certain unstated assumptions, and we make these more explicit in order to make clearer the link between embodiment and AGI. First and foremost, it is **usually assumed that AGI is in the context of the human world as we currently know it**, and with follow-on assumptions about normal laws of physics (e.g. gravity), the present time, etc. unless otherwise stated.

These assumptions are important, but are so commonly assumed that one does not usually stop to ponder the underlying assumptions when posed questions such as "will a glass cup break when you drop it?" The answer obviously depends on many things, such as whether it falls due to gravity (not necessarily true in outer space), whether it contacts a sufficiently-hard surface (not necessarily true if one is in bed), whether it's made of shatter-proof glass, etc.

Gist of Our Argument

A crucial thing to note about AGI is that there is no single universally-accepted, objective definition or test. Even the above definition – our preferred one – uses subjective phrases such as "diverse tasks" and "broad range". Hence, unlike relatively well-accepted tests for humans such as IQ tests or driving tests, whether an AI system has attained AGI is still a highly subjective matter at present. In other words, for better or worse, **to be recognized as attaining**

[*] These authors contributed equally.

AGI would likely require a subjective consensus among some number of thought leaders.

Current large language models have shown human- or superhuman-level performance on a range of text-based intellectual tasks or tests, arguably fulfilling some to most of the key aspects of the above AGI definition. However, few people are claiming that AGI has been achieved already. The more optimistic of LLM proponents believe that "sparks" of AGI have been demonstrated (Bubeck et al. 2023), or that further scaling of current architectures can lead to AGI.

What can we conclude from this? Since there is no single "acid test" for AGI, and since embodiment is necessary for AGI (as we argue below), ultimately it is likely that the most convincing indications of AGI would result from embodied demonstration.

Language Is Not Enough: Why Embodiment Is Needed for AGI

In this section, we lay out various arguments as to why AGI requires embodiment. Certainly, embodiment is not required for every single aspect or component capability of AGI – there are sufficient obvious counter-examples of human-level intellectual abilities achieved by existing AI systems without embodiment. Rather, we argue that without embodiment, AGI cannot emerge as a capability of the system.

Application perspective: general-purpose robots need embodied intelligence. General-purpose robots that, roughly speaking, can do anything that humans can do, are probably the most common application that drives interest in AGI. Without the ability to carry out a broad range of actions in the real world, or with only specific robotic capabilities (e.g. welding or vacuuming), AGI would be a lot less useful or meaningful. It then follows that AGI for general-purpose robots must be equipped to understand the nature of the physical world (Duan et al. 2022a), as well as the nature of the body that it occupies. An AGI controlling a child-sized robot versus a truck-sized giant robot would surely have different abilities, face different constraints and accomplish the same set of goals differently.

That is the clearest (but perhaps least interesting) argument for AGI needing to have embodied capabilities. That said, one counter-argument would be that AGI could be defined to be about the high-level learning, planning and reasoning capabilities, while the capabilities about navigation, obstacle avoidance, action planning, motor control, etc. are robotics rather than AGI. Our view is that by this argument, many widely-accepted components of AI (and AGI) such as vision and language could be similarly stripped away, leaving AGI to be defined as just abstract learning, planning and reasoning. But if there's one lesson from the decades of research in classical GOFAI ("good old-fashioned AI"), it's that these abstractions also oversimplify the problem, stripping away the tough complexities as well as possible interdependencies that must be addressed at some point.

More importantly, we argue that **even abstract, high-level planning must take into account the practical and physical realities and constraints of the body and the environment in order to be general and flexible**. To use a simple navigation example, a system cannot plan a feasible path to the goal location unless it understands what "feasible" entails, which obviously depends on a plethora of embodied considerations such as overhead clearance, floor loading capacity, etc. Embodiment implies that capabilities must be developed to handle understanding of intuitive physics and a plethora of other real-world constraints, which abstract high-level planning alone cannot account for.

Language is an interpretive description, not the actual thing. This argument is a specific counter to the view that LLMs can be sufficient to eventually achieve AGI. While a lot of knowledge is captured and represented symbolically, the conciseness of knowledge in the form of a natural language sentence comes at the cost of it being **just one interpretation** of reality, not the reality itself. For example, images described as "a person smiling" or "a room with toys on the floor" connote happiness and messiness, and these become assumptions henceforth. However, a person's mental state can only be guessed, and messiness is a subjective and continuous attribute. If the system were to attempt to be more factual and detailed, descriptions such as "the corners of the person's mouth are upturned at X degrees" become unwieldy, and are incomplete anyway.

Furthermore, the compression of knowledge into language **entails (commonsense) assumptions**. The "fact" that "pork belly can be eaten" makes assumptions about it being prepared, cooked, its cleanliness, religious and cultural norms, allergies, health considerations, etc. (and even what it means to be edible). If one were to list all the possible exceptions and conditions, then it would be an extremely unwieldy and practically unusable piece of knowledge.

Similarly, not everything is meaningfully describable in language alone. Take for example the query "can I eat this banana, and why?" This is a multi-layered query involving issues of ownership or permission, the questioner's context (e.g. allergies), social context (e.g. the social meal event has not started yet), object properties (real vs. plastic or wax banana), etc. Even putting these aside, assuming the question is now about the ripeness of the banana, all explanations about ripeness and edibleness have to ultimately fall back upon other people having eaten bananas of similar shade without incident – an explanation rooted in embodiment.

Why Embodiment Is Key for AGI

In the previous section, we laid out various reasons why embodiment is required for AGI. In this section, we go further to explain why embodiment is key for **convincing demonstrations of AGI**.

At this point, let us examine in more detail the definition of AGI – Artificial **General** Intelligence. It is the word "general" that differentiates AGI from AI. The latter includes everything that we have today (since presumably we do not have AGI yet), such as current computer vision, natural language processing (NLP) and robotics technology. What makes today's AI technology not sufficiently general, such that the consensus is that we have not achieved AGI yet? We examine a number of answers in increasing complexity, which will then lead us to why embodiment is key for AGI.

Current AI is too narrow and brittle. The first and most straightforward answer as to why today's AI is not sufficiently general to be considered AGI, is that many of the technologies that achieve human-level or superhuman-level performance are fairly narrow in their capabilities, and are quite brittle even within their domain of expertise.

One simple example is chess, where rule-based AI surpassed humans in 1997 (Campbell, Hoane Jr, and Hsu 2002), eventually followed by RL-based AI methods (Silver et al. 2018). Nobody would claim that superhuman chess or Go algorithms can do anything other than play the given games, and even simple tweaks to the game rules, and unusual or adversarial strategies can throw the algorithms off (Lan et al. 2022; Timbers et al. 2020; Wang et al. 2022). Even deep RL algorithms that can master multiple Atari games (Mnih et al. 2013, 2015) are still ultimately constrained to a certain subset of game types. Even if one were to be extremely generous, these algorithms are still constrained to problems that lend themselves to RL. And even then, they are very susceptible to even simple manipulations (Kansky et al. 2017).

Explicit, formal knowledge is not enough. In more recent years, LLMs, having been trained on immense corpora of textual information, have performed surprisingly well on a fairly broad variety of tests (Qin et al. 2023; Srivastava et al. 2022; Wei et al. 2022a). These tests span knowledge domains such as mathematics, the English language, law, programming, etc. (Frieder et al. 2023; Sobania et al. 2023; Srivastava et al. 2022). The seeming breadth of knowledge capabilities, combined with excellent grammar and fluency, some evidence of reasoning capabilities (Brown et al. 2020; Qin et al. 2023; Wei et al. 2022b; Yao et al. 2023), as well as the ability to change styles and personas (Deshpande et al. 2023; Lyu et al. 2021), have led some to think that AGI will soon be upon us (Bubeck et al. 2023; Zhang et al. 2023).

However, despite LLMs having been trained on immense data and having knowledge far exceeding any individual human could ever have, the best LLMs today are still far from perfect (Sap et al. 2022; Valmeekam et al. 2022; Webson and Pavlick 2022; Wolf et al. 2023), and the general consensus is that we have not achieved AGI yet (Biever 2023).

Why is this the case? Despite having more "general knowledge" than any single human could ever have, LLMs also "hallucinate" and easily make up untrue statements (Alkaissi and McFarlane 2023; Ji et al. 2023). They also perform poorly on tests of commonsense reasoning (Bian et al. 2023; Chen et al. 2023; Qin et al. 2023). Thus, doing well on explicit, formal knowledge – even for a broad range of domains – is not sufficient to make the leap from AI to AGI in people's minds.

AGI is gauged from performance on everyday human tasks. This brings us to the penultimate step in our argument for embodiment being key to AGI. Even if LLMs were fixed to no longer hallucinate, poor performance on commonsense reasoning would still hold people back from believing that LLMs possess AGI. Commonsense reasoning is indeed a capability that some people have held up as the key to AGI (Choi 2023; LeCun 2022).

Our view is that this is only part of the answer. But commonsense reasoning is not something that can truly be tested only conceptually, in text. As argued earlier, language is interpretive description, and can only be a partial snapshot of reality. When one asks an LLM a commonsense question to test it, there are many simplifications and underlying assumptions compared to an actual similar scenario in the real world. As the saying goes, "the proof of the pudding is in the eating". We believe that **until an AI system controlling a robot in the real world is extensively tested successfully in live, uncontrolled interactions for an extended period of time, a majority of discerning AI researchers would not be fully convinced that it possesses AGI**.

Why might this be the case? The real world has properties and challenges unlike any game, toy, simulated or simplified environment would have – and these make all the difference when it comes to AGI. The real world is stochastic, dynamic, uncontrollable, real-time, highly constraining, complex and non-repeatable. An AI system in the real world has to continuously and rapidly perform a sense-decide-act cycle while facing all these potential challenges.

It is such challenges that lead to some well-known researchers to hold the view that today's AI systems are still far from the capabilities that even young children possess (Liu, Brooks, and Spelke 2019; Smith and Gasser 2005; Stojnić et al. 2023). It's certainly not due to children having extensive general knowledge, and most of their capabilities fall below those of the average adult human. Rather, it is their demonstrated ability to actually robustly interact with and "survive" in the real world, without needing to be rebooted, retrained or redesigned (if they were AI systems).

Granted, children are not infallible (nor are adults), and they require adult help and guidance in today's world, but one can easily imagine that for early humans, young teenagers would already be considered old enough to hunt and gather independently. Most people would not strongly disagree that the average human child possesses general intelligence (the natural counterpart of AGI), or at least would be expected to grow into an average adult human that does.

Embodiment is key. This line of thinking leads us, finally, to the crux of our argument. Summarizing the arguments so far, we have seen that: **i)** even superhuman performance on highly-intellectual tasks is not sufficient to be considered as achieving AGI; **ii)** nor is good performance on a broad range of explicit, formal knowledge tests. Performing well on **iii)** text-based tests of commonsense reasoning would probably convince a number of people that AGI is achieved, but ultimately, **iv)** the most convincing demonstration would likely be successfully learning and performing a broad range of tasks in the real-world continually over an extended period of time (Tan, Lallee, and Mandal 2017). **Such a demonstration would not only require commonsense reasoning, but also bring together a range of other important AI abilities in a robust and flexible overall system.**

So the remaining argument becomes very simple. If attainment of i), ii) and iii) are not sufficient to be convincing demonstrations of AGI, then it stands to reason that something like iv), i.e. embodied real-world performance, is a truer and more definitive reflection of AGI that ultimately really matters.

Researchers working on i), ii) and iii) have done so in the hope that attainment of these abilities would be key, and then "the rest of the problem" would be simple to solve. But thus far, these hopes have not borne out. Certainly, the capabilities or components developed would be useful or even important parts of achieving AGI, but there has always seemed to be "something more" that was crucially missing.

Concluding Remarks

The classic adage is that what's hard for humans is easy for AI, and what's easy for humans is hard for AI. This perfectly aligns with our thinking that one of the seemingly easiest capabilities for humans is simply: being in the real-world and carrying out everyday tasks. Not every human can pass the US LSAT and Uniform Bar Exam, much less ace them, as GPT-4 has done. But literally billions of people go about their lives every day, easily doing mundane things like getting around, getting food and drink, and getting shelter and rest.

While these seem like simple things (to humans, at least), the nature of the real world means that the full array of human general intelligence capabilities must be available to be called upon for any of the myriad emergencies, contingencies and corner cases that can and do happen to all of us.

Acknowledgements

This work was supported by A*STAR CRF funding to C.T.

References

Alkaissi, H.; and McFarlane, S. I. 2023. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus*, 15(2).

Arel, I. 2012. Deep reinforcement learning as foundation for artificial general intelligence. In *Theoretical Foundations of Artificial General Intelligence*, 89–102. Springer.

Berner, C.; Brockman, G.; Chan, B.; Cheung, V.; Debiak, P.; Dennison, C.; Farhi, D.; Fischer, Q.; Hashme, S.; Hesse, C.; et al. 2019. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*.

Bian, N.; Han, X.; Sun, L.; Lin, H.; Lu, Y.; and He, B. 2023. ChatGPT is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. *arXiv preprint arXiv:2303.16421*.

Biever, C. 2023. The easy intelligence tests that AI chatbots fail. *Nature*, 619: 686–689.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33: 1877–1901.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; et al. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Campbell, M.; Hoane Jr, A. J.; and Hsu, F.-h. 2002. Deep blue. *Artificial Intelligence*, 134(1-2): 57–83.

Chen, J.; Shi, W.; Fu, Z.; Cheng, S.; Li, L.; and Xiao, Y. 2023. Say what you mean! Large language models speak too positively about negative commonsense knowledge. *arXiv preprint arXiv:2305.05976*.

Choi, Y. 2023. Common Sense: The Dark Matter of Language and Intelligence. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, 2–2.

Chollet, F. 2019. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.

Chrisley, R.; and Ziemke, T. 2006. Embodiment. *Encyclopedia of Cognitive Science*.

Dawid, A.; and LeCun, Y. 2023. Introduction to Latent Variable Energy-Based Models: A Path Towards Autonomous Machine Intelligence. *arXiv preprint arXiv:2306.02572*.

Deitke, M.; Batra, D.; Bisk, Y.; Campari, T.; Chang, A. X.; Chaplot, D. S.; Chen, C.; D'Arpino, C. P.; Ehsani, K.; Farhadi, A.; et al. 2022. Retrospectives on the embodied AI workshop. *arXiv preprint arXiv:2210.06849*.

Deshpande, A.; Murahari, V.; Rajpurohit, T.; Kalyan, A.; and Narasimhan, K. 2023. Toxicity in ChatGPT: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335*.

Duan, J.; Dasgupta, A.; Fischer, J.; and Tan, C. 2022a. A survey on machine learning approaches for modelling intuitive physics. *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence Survey Track*.

Duan, J.; Yu, S.; Tan, H. L.; Zhu, H.; and Tan, C. 2022b. A survey of embodied AI: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2): 230–244.

Frieder, S.; Pinchetti, L.; Griffiths, R.-R.; Salvatori, T.; Lukasiewicz, T.; Petersen, P. C.; Chevalier, A.; and Berner, J. 2023. Mathematical capabilities of ChatGPT. *arXiv preprint arXiv:2301.13867*.

Glenberg, A. M. 2010. Embodiment as a unifying perspective for psychology. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(4): 586–596.

Goertzel, B. 2014. Artificial general intelligence: concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, 5(1): 1.

Heinrich, J.; and Silver, D. 2016. Deep reinforcement learning from self-play in imperfect-information games. *arXiv preprint arXiv:1603.01121*.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12): 1–38.

Kansky, K.; Silver, T.; Mély, D. A.; Eldawy, M.; Lázaro-Gredilla, M.; Lou, X.; Dorfman, N.; Sidor, S.; Phoenix, S.; and George, D. 2017. Schema networks: Zero-shot transfer with a generative causal model of intuitive physics. In *International Conference on Machine Learning*, 1809–1818. PMLR.

Lan, L.-C.; Zhang, H.; Wu, T.-R.; Tsai, M.-Y.; Wu, I.; Hsieh, C.-J.; et al. 2022. Are AlphaZero-like Agents Robust to Adversarial Perturbations? *arXiv preprint arXiv:2211.03769*.

LeCun, Y. 2022. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62.

Liu, S.; Brooks, N. B.; and Spelke, E. S. 2019. Origins of the concepts cause, cost, and goal in prereaching infants. *Proceedings of the National Academy of Sciences*, 116(36): 17747–17752.

Longo, M. R.; Schüür, F.; Kammers, M. P.; Tsakiris, M.; and Haggard, P. 2008. What is embodiment? A psychometric approach. *Cognition*, 107(3): 978–998.

Lyu, Y.; Liang, P. P.; Pham, H.; Hovy, E.; Póczos, B.; Salakhutdinov, R.; and Morency, L.-P. 2021. StylePTB: A compositional benchmark for fine-grained controllable text style transfer. *arXiv preprint arXiv:2104.05196*.

Michael, J.; Holtzman, A.; Parrish, A.; Mueller, A.; Wang, A.; Chen, A.; Madaan, D.; Nangia, N.; Pang, R. Y.; Phang, J.; et al. 2022. What do NLP researchers believe? Results of the NLP community metasurvey. *arXiv preprint arXiv:2208.12852*.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; and Riedmiller, M. 2013. Playing Atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529–533.

Pennachin, C.; and Goertzel, B. 2007. Contemporary approaches to artificial general intelligence. *Artificial General Intelligence*, 1–30.

Qin, C.; Zhang, A.; Zhang, Z.; Chen, J.; Yasunaga, M.; and Yang, D. 2023. Is ChatGPT a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Sap, M.; LeBras, R.; Fried, D.; and Choi, Y. 2022. Neural theory-of-mind? On the limits of social intelligence in large LMs. *arXiv preprint arXiv:2210.13312*.

Silver, D.; Hubert, T.; Schrittwieser, J.; Antonoglou, I.; Lai, M.; Guez, A.; Lanctot, M.; Sifre, L.; Kumaran, D.; Graepel, T.; et al. 2018. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419): 1140–1144.

Silver, D.; Singh, S.; Precup, D.; and Sutton, R. S. 2021. Reward is enough. *Artificial Intelligence*, 299: 103535.

Smith, L.; and Gasser, M. 2005. The development of embodied cognition: Six lessons from babies. *Artificial Life*, 11(1-2): 13–29.

Sobania, D.; Briesch, M.; Hanna, C.; and Petke, J. 2023. An analysis of the automatic bug fixing performance of ChatGPT. *arXiv preprint arXiv:2301.08653*.

Srivastava, A.; Rastogi, A.; Rao, A.; Shoeb, A. A. M.; Abid, A.; Fisch, A.; Brown, A. R.; Santoro, A.; Gupta, A.; Garriga-Alonso, A.; et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Stojnić, G.; Gandhi, K.; Yasuda, S.; Lake, B. M.; and Dillon, M. R. 2023. Commonsense psychology in human infants and machines. *Cognition*, 235: 105406.

Tan, C.; Lallee, S.; and Mandal, B. 2017. Vision and memory: Looking beyond immediate visual perception. *Computational and Cognitive Neuroscience of Vision*, 195–219.

Timbers, F.; Bard, N.; Lockhart, E.; Lanctot, M.; Schmid, M.; Burch, N.; Schrittwieser, J.; Hubert, T.; and Bowling, M. 2020. Approximate exploitability: Learning a best response in large games. *arXiv preprint arXiv:2004.09677*.

Valmeekam, K.; Olmo, A.; Sreedharan, S.; and Kambhampati, S. 2022. Large Language Models Still Can't Plan (A Benchmark for LLMs on Planning and Reasoning about Change). *arXiv preprint arXiv:2206.10498*.

Wang, T. T.; Gleave, A.; Belrose, N.; Tseng, T.; Miller, J.; Dennis, M. D.; Duan, Y.; Pogrebniak, V.; Levine, S.; and Russell, S. 2022. Adversarial policies beat professional-level go AIs. *arXiv preprint arXiv:2211.00241*.

Webson, A.; and Pavlick, E. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2300–2344.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Chi, E.; Le, Q.; and Zhou, D. 2022b. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Wolf, Y.; Wies, N.; Levine, Y.; and Shashua, A. 2023. Fundamental limitations of alignment in large language models. *arXiv preprint arXiv:2304.11082*.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T. L.; Cao, Y.; and Narasimhan, K. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Zhang, C.; Zhang, C.; Li, C.; Qiao, Y.; Zheng, S.; Dam, S. K.; Zhang, M.; Kim, J. U.; Kim, S. T.; Choi, J.; et al. 2023. One small step for generative AI, one giant leap for AGI: A complete survey on ChatGPT in AIGC era. *arXiv preprint arXiv:2304.06488*.