Whoever owns human-level AGI (Artificial General Intelligence) will own the global means of production for all goods and services [Brynjolfsson (2022)]. Superintelligent AGI has been conservatively estimated to have a net present value of ~$15 quadrillion [Russell (2024)]. Accordingly, the major profit-motivated AI labs (and their associated sovereign states) are currently engaged in an AGI arms race [Ramamoorthy and Yampolskiy (2018)], each in pursuit of their own short-term self-interest, seemingly oblivious to the long-term best interest of the human species as a whole [Hardin (1968); Dawes (1980); Alexander (2014)].

The currently dominant opinion among AI and AI safety researchers [AI Safety Institute (2024)] seems to be that multimodal Large Language Models (LLMs) [Zhao et al. (2023); Wu et al. (2023)], built using the Transformer neural network model [Vaswani et al. (2023)] (or similar), massively scaled [Kaplan et al. (2020)], and aligned with human preferences via RLHF (Reinforcement Learning through Human Feedback) and other methods [Shen et al. (2023)], represent the most promising path to AGI and beyond [Bubeck et al. (2023)], with the median estimate for when human-level AGI will arrive ranging from 2026 [Metaculus (2024a)] to 2031 [Metaculus (2024b)], with superintelligent AGI arriving ~2 years later [Metaculus (2024c)].

At the same time, many AI researchers hold variously negative opinions about LLMs, including that "we have no idea how they work" [Nanda (2023)], that they are "stochastic parrots" [Bender et al. (2021)] capable of at most weak reasoning [Berglund et al. (2023); Nezhurina et al. (2024)] over shallow world models [Li et al. (2023)], that LLM hallucinations [Magesh et al. (2024)] are inevitable [Xu, Jain, and Kankanhalli (2024)], that the scaling laws are illusory [Udandarao et al. (2024)], and that reliable LLM alignment is impossible [Wolf et al. (2024)]. In a recent survey [AI Impacts (2022)], the median AI researcher appeared to put a 5-10% chance on the extinction risk from misaligned AGI [Hendrycks and Mazeika (2022); Bengio (2024)].

We propose an alternative to the de facto LLM-based short-term-self-interest-driven approach to AGI. Rather than engage in a race, over the next 10-20 years, towards an AGI future that is likely to be, at best, *hugely sub-optimal for all mankind for all eternity* (due to the trapdoor nature of superintelligence [Yudkowsky (2021)]), and, at worst, catastrophic [Hendrycks, Mazeika, and Woodside (2023); Yampolskiy (2024)], we propose spending 50-100 years doing it properly [Russell (2022); Russell (2023); Tegmark and Omohundro (2023)], in the best interest of all mankind, in order to achieve an AGI endgame that is (as close as possible to) maximally-beneficent for all mankind, while at the same time using the additional breathing space to mitigate to the maximum extent possible the inevitable pain of such a profound transition [Altman (2023)[1]].

Our overall approach is to try to imagine the *ideal endgame* (from the perspective of the human species as a whole), and to work backwards from there in order to make it (or something close to it) actually happen. This is largely equivalent to imagining the ideal (or "Gold Standard") superintelligent AGI, and then working backwards to actually build it. To this end, we seek to design, develop, and deploy a *provably maximally-aligned maximally-superintelligent* AGI (called BigMother/BigMom) that is ultimately owned by all mankind (via the United Nations), and whose operation benefits all mankind, without favouring any subset thereof (such as the citizens of any particular country or countries, or the shareholders of any particular company or companies).

The proposed BigMother cognitive architecture comprises the following hybrid (symbolic + connectionist) AGI stack: NBG set theory as the knowledge representation language (KRL); cognitive primitives include induction, deduction, and abduction (IDA) manipulating beliefs expressed in the KRL; generic problem-solving, program synthesis, continuous learning (from observation of the real world), and continuous planning all constructed on top of IDA; all major components formally specified, with code and proofs generated via program synthesis; code and data distributed across massively parallel (and provably correct) hardware; performance further enhanced via FPGAs, ASICs, VLSI, neural networks, GPUs, quantum, etc; comprehensive, multi-decade, primary, secondary, and tertiary machine education (including to PhD level and beyond).

In *The BigMother Manifesto: A Roadmap to Provably Maximally-Aligned Maximally-Superintelligent AGI (Part 1)* [Turner (forthcoming)], we will describe the BigMother cognitive architecture and associated BigMother project in detail. Together, these define an AGI research agenda for the next 50-100 years.

[1] "A gradual transition to a world with AGI is better than a sudden one. ... A gradual transition gives people, policymakers, and institutions time to understand what's happening, personally experience the benefits and downsides of these systems, adapt our economy, and to put regulation in place. It also allows for society and AI to co-evolve, and for people collectively to figure out what they want while the stakes are relatively low. ... A slower takeoff is easier to make safe ..."

References

AI Impacts. 2022. Is AI an existential risk to humanity? https://wiki.aiimpacts.org/doku.php?id=arguments_for_ai_risk:is_ai_an_existential_threat_to_humanity:start.

AI Safety Institute. 2024. International Scientific Report on the Safety of Advanced AI. https://drive.google.com/drive/folders/1pv-m1qTfP7osJE-bXOnNpULsn-_L2xcE.

Alexander, S. 2014. Meditations on Moloch. https://slatestarcodex.com/2014/07/30/meditations-on-moloch/.

Altman, S. 2023. Planning for AGI and beyond. https://openai.com/blog/planning-for-agi-and-beyond.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 610–623. New York, NY, USA: Association for Computing Machinery.

Bengio, Y. 2024. Why a Forefather of AI Fears the Future. https://www.youtube.com/watch?v=KcbTbTxPMLc.

Berglund, L.; Tong, M.; Kaufmann, M.; Balesni, M.; Stickland, A. C.; Korbak, T.; and Evans, O. 2023. The Reversal Curse: LLMs trained on "A is B" fail to learn "B is A".

Brynjolfsson, E. 2022. The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; Nori, H.; Palangi, H.; Ribeiro, M. T.; and Zhang, Y. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4.

Dawes, R. M. 1980. Social Dilemmas. *Annual Review of Psychology* 31.

Hardin, G. 1968. The Tragedy of the Commons. *Science* 162:1243–1248.

Hendrycks, D., and Mazeika, M. 2022. X-Risk Analysis for AI Research.

Hendrycks, D.; Mazeika, M.; and Woodside, T. 2023. An Overview of Catastrophic AI Risks.

Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling Laws for Neural Language Models.

Li, K.; Hopkins, A. K.; Bau, D.; Viégas, F.; Pfister, H.; and Wattenberg, M. 2023. Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task.

Magesh, V.; Surani, F.; Dahl, M.; Suzgun, M.; Manning, C. D.; and Ho, D. E. 2024. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools.

Metaculus. 2024a. When will the first weakly general AI system be devised, tested, and publicly announced? https://www.metaculus.com/questions/3479/date-weakly-general-ai-is-publicly-known.

Metaculus. 2024b. When will the first general AI system be devised, tested, and publicly announced? https://www.metaculus.com/questions/5121/date-of-artificial-general-intelligence.

Metaculus. 2024c. After a (weak) AGI is created, how many months will it be before the first superintelligent AI is created? https://www.metaculus.com/questions/9062/time-from-weak-agi-to-superintelligence.

Nanda, N. 2023. What is a Transformer? (Transformer Walkthrough Part 1/2). https://www.youtube.com/watch?v=bOYE6E8JrtU.

Nezhurina, M.; Cipolina-Kun, L.; Cherti, M.; and Jitsev, J. 2024. Alice in Wonderland: Simple Tasks Showing Complete Reasoning Breakdown in State-Of-the-Art Large Language Models.

Ramamoorthy, A., and Yampolskiy, R. V. 2018. Beyond MAD?: The Race for Artificial General Intelligence. https://www.itu.int/en/journal/001/Documents/itu2018-9.pdf.

Russell, S. 2022. Provably Beneficial Artificial Intelligence. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, IUI '22, 3. New York, NY, USA: Association for Computing Machinery.

Russell, S. 2023. How Not To Destroy the World With AI. https://www.youtube.com/watch?v=ISkAkiAkK7A.

Russell, S. 2024. AI: What If We Succeed. https://www.youtube.com/watch?v=UvvdFZkhhqE.

Shen, T.; Jin, R.; Huang, Y.; Liu, C.; Dong, W.; Guo, Z.; Wu, X.; Liu, Y.; and Xiong, D. 2023. Large Language Model Alignment: A Survey.

Tegmark, M., and Omohundro, S. 2023. Provably safe systems: the only path to controllable AGI.

Turner, A. (forthcoming). The BigMother Manifesto: A Roadmap to Provably Maximally-Aligned Maximally-Superintelligent AGI (Part 1). https://www.bigmother.ai; https://www.bigmom.ai.

Udandarao, V.; Prabhu, A.; Ghosh, A.; Sharma, Y.; Torr, P. H. S.; Bibi, A.; Albanie, S.; and Bethge, M. 2024. No "Zero-Shot" Without Exponential Data: Pretraining Concept Frequency Determines Multimodal Model Performance.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2023. Attention Is All You Need.

Wolf, Y.; Wies, N.; Avnery, O.; Levine, Y.; and Shashua, A. 2024. Fundamental Limitations of Alignment in Large Language Models.

Wu, J.; Gan, W.; Chen, Z.; Wan, S.; and Yu, P. S. 2023. Multimodal Large Language Models: A Survey.

Xu, Z.; Jain, S.; and Kankanhalli, M. 2024. Hallucination is Inevitable: An Innate Limitation of Large Language Models.

Yampolskiy, R. V. 2024. *AI: Unexplainable, Unpredictable, Uncontrollable*. CRC Press.

Yudkowsky, E. 2021. The real core of the argument for 'AGI risk' (AGI ruin). https://twitter.com/ESYudkowsky/status/1405580521237745665.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; Du, Y.; Yang, C.; Chen, Y.; Chen, Z.; Jiang, J.; Ren, R.; Li, Y.; Tang, X.; Liu, Z.; Liu, P.; Nie, J.-Y.; and Wen, J.-R. 2023. A Survey of Large Language Models.

