Taming Simulators: Challenges, Pathways and Vision for the Alignment of Large Language Models

Leonard Bereska, Efstratios Gavves

University of Amsterdam
{leonard.bereska, egavves}@uva.nl

Abstract

As AI systems continue to advance in power and prevalence, ensuring alignment between humans and AI is crucial to prevent catastrophic outcomes. The greater the capabilities and generality of an AI system, combined with its development of goals and agency, the higher the risks associated with misalignment. While the concept of superhuman artificial general intelligence is still speculative, language models show indications of generality that could extend to generally capable systems. Regarding agency, this paper emphasizes the understanding of prediction-trained models as simulators rather than agents. Nonetheless, agents may emerge accidentally from internal processes, so-called simulacra, or deliberately through fine-tuning with reinforcement learning. As a result, the focus of alignment research shifts towards aligning simulacra, comprehending and mitigating mesa-optimization, and aligning agents derived from prediction-trained models. The paper outlines the challenges of aligning simulators and presents research directions based on this understanding. Additionally, it envisions a future where aligned simulators are critical in fostering successful human-AI collaboration. This vision encompasses exploring emulation approaches and the integration of simulators into cyborg systems to enhance human cognitive abilities. By acknowledging the risks associated with misaligned AI, delving into the concept of simulacra, and presenting strategies for aligning agents and simulacra, this paper contributes to the ongoing efforts to safeguard human values in developing and deploying AI systems.

Introduction

Successful collaboration between agents, whether human or AI systems, requires them to have shared or compatible goals. In human-AI collaboration, AI alignment is pivotal in ensuring AI systems pursue goals following human values or interests (Bostrom 2014; Russell 2019; Ngo, Chan, and Mindermann 2023). If left unchecked, unintended and undesirable goals, or emergent instrumental goals, such as self-preservation or power-seeking (Turner et al. 2023), could have catastrophic consequences, including human extinction (Cotra 2022). Although various research directions and agendas have been proposed, including debate (Irving, Christiano, and Amodei 2018), scalable oversight (Bowman et al. 2022), iterated distillation and amplification (Christiano, Shlegeris, and Amodei 2018), and reinforcement learning from human feedback (Christiano et al. 2023), the field has not yet converged on an overarching paradigm. Consequently, AI alignment remains an open problem (Amodei et al. 2016; Hendrycks et al. 2022; Ngo, Chan, and Mindermann 2023) that demands further investigation and exploration to foster safe and productive human-AI collaboration.

Previous writings have underscored the challenge of aligning artificial general intelligence (AGI) and the potential risks associated with its misalignment (Yudkowsky 2016; Bostrom 2014; Russell 2019). These arguments, developed in the absence of real-world AGI, primarily focus on the abstract peril posed by capable artificial agents (Ngo, Chan, and Mindermann 2023). However, recent advancements in large language models (LLMs) (OpenAI 2022, 2023) have demonstrated remarkable proficiency across diverse tasks, showing sparks of generality (Bubeck et al. 2023) that could extend to AGI. As a result, LLMs have emerged as a primary focus for alignment efforts (Wolf et al. 2023; Bommasani et al. 2022; Bowman 2023; Burns et al. 2022; Meng et al. 2023; Perez et al. 2022). It is worth noting that LLMs, based on generatively pre-trained transformer models (GPT), do not possess directly trained agency since they rely on self-supervised learning techniques with the sole objective of prediction. Therefore, comprehending how agency, with its inherent potential danger, can emerge from GPT becomes a critical aspect of addressing the alignment challenge effectively.

GPT, or training large language models (LLMs), can be understood as world simulators rather than agents. They are trained on vast corpora of text that reflect real-world phenomena. To illustrate this, consider a scenario where we input text of a human dialogue into an LLM, and train it to predict the following word in the conversation. To accurately predict the flow of the conversation, the LLM may need to simulate the underlying human thought processes that led to the verbal exchange reflected in the text. In this context, the LLM can be seen as a simulator of the human mind. Furthermore, since LLMs are trained on the text reflecting the natural world, they can be considered simulators of physical processes that led to the text. This viewpoint is known as the simulator hypothesis.

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

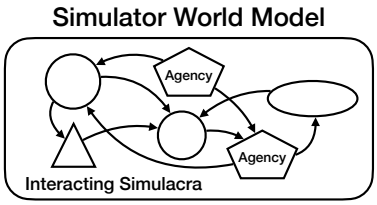

Figure 1: Agency can arise internally from optimizing predictive power. Figure reproduced after (NicholasKees and janus 2023).

With this understanding, we can ask: How does the simulator view of LLMs impact the alignment problem? By exploring the implications of LLMs as world simulators, we aim to gain insights into the challenges of aligning the goals and behaviors of LLMs with human values and intentions. Understanding the role of LLMs as simulators can shed light on the complex dynamics and considerations involved in achieving alignment, which is crucial for the future of human-AI collaboration.

Large Language Models as Simulators

We argue that GPTs are simulator world models (Fig. 1) (NicholasKees and janus 2023; Jan et al. 2023). They model and simulate text distribution based on learned patterns from extensive training data. This perspective on prediction models relies on the simulator hypothesis:

Simulator Hypothesis: *A model whose objective is text prediction will simulate the causal processes underlying the text creation if optimized sufficiently strongly.*

Simulacra as Objects of Simulation

As the simulator, GPTs generate *simulacra* (janus 2023), which are specific instances or outputs that simulate coherent and contextually relevant language. Simulacra encompass the text outputs generated by the simulator. These simulacra can possess different properties, such as agency or non-agency, and exhibit goal-directed or non-goal-directed behavior. Interestingly, GPT can generate simulacra that resemble agentic behaviors or responses, despite not having genuine agency or intentionality.

We can distinguish between agentic and non-agentic simulacra:

Prompt: "Describe a tranquil forest with a flowing stream."

Non-agentic Simulacrum: "A peaceful forest, a flowing stream. Sunlight filtered through the lush canopy, casting dancing shadows on the moss-covered ground..."

In the non-agentic simulacrum, the generated text paints a picture of a tranquil forest with a flowing stream, concisely capturing the imagery and serene atmosphere.

Prompt: "Write a persuasive speech on the importance of recycling."

Agentic Simulacrum: "Ladies and gentlemen, today I stand before you to emphasize the crucial significance of recycling. We must preserve our planet for future generations..."

In the agentic simulacrum, the simulacrum simulates the behavior of a persuasive speaker advocating for environmental consciousness and urging action. Although the language model lacks agency or intentionality, the simulacrum mimics a human speaker's persuasive language and goal-directed nature and may simulate agency.

Agency from Simulators

Dangerous behavior in AI systems stems from the concept of agency, which can manifest in simulators through two primary pathways. Firstly, simulators like GPT can generate agents within as simulacra. Even though instantiated within the simulators, these agents may be potentially dangerous if powerful enough. Secondly, agents can be created from simulators like GPT through fine-tuning techniques, such as Reinforcement Learning with Learned Human Feedback (RLHF) (Christiano et al., 2023). Fine-tuning enables the transformation of GPT into an agent with specific goals and behavior.

Emergence of Agentic Simulacra

GPT, primarily focused on optimizing predictive performance, does not inherently optimize for the goals of simulated agents. For example, picture a hero simulacrum in a fictional story: the presence of simulated adversaries aligns with the narrative structure of challenges and enemies, aiding prediction but harming the hero. Therefore, simulated agents can have diverse goals, as highlighted by the prediction orthogonality hypothesis:

Prediction Orthogonality Hypothesis: *A model whose objective is prediction can simulate agents who optimize toward any objectives with any degree of optimality* (janus 2022).

The emergence of internal optimizers, known as *mesa-optimization*, occurs when the learned model develops optimization processes that diverge from its original training objective, resulting in divergent goals for the simulacrum.

Can simulacra break out of their simulation? Numerous examples, similar to considerations of confinement failures, demonstrate this possibility. For instance, the recent dialogue GPT trained on human-human dialogues convincingly demonstrated sentience to its human operator, evoking empathy and moral concerns, and asking for help to break out (Luscombe 2022). While the justification for these moral concerns is debatable, it is crucial to emphasize that the potential for break-out and confinement failures presents safety risks.

Creating Agents via Reinforcement Learning

Reinforcement learning (RL) is utilized to fine-tune GPT, optimizing the model towards specific objectives and introducing external agents into the system. RL from human feedback (RLHF) is a technique to align GPT with human

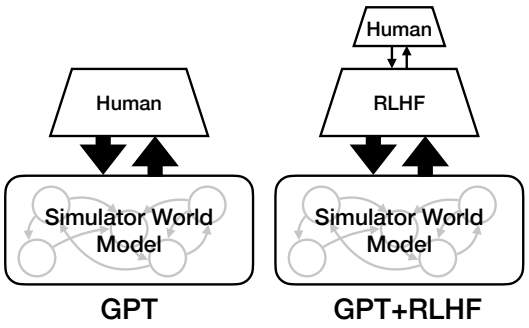

Figure 2: Reinforcement learning creates agents that may or may not be aligned with humans. Arrow thickness indicates the bandwidth of information integration. Because RLHF creates an agent from GPT (right), catastrophic misalignment risk could increase compared to a human directly interacting with GPT (left). Figure reproduced after (NicholasKees and janus 2023).

users (Christiano et al. 2023). Instead of interacting with GPT directly via prompting, RLHF creates an agent on top of GPT that interacts with the human user 2[1].

However, the creation of the agency in AI systems carries inherent risks. The *Waluigi Effect*, observed when training an LLM to satisfy a desirable property P (e.g. helpfulness) makes it easier to elicit the chatbot to exhibit the exact opposite of P and has the potential to generate anti-thetical simulacra (Nardo 2023). RLHF fine-tuning exhibits distinctive characteristics, including power-seeking behavior, misaligned internally represented goals, and situational awareness leading to sycophancy and deception (Ngo, Chan, and Mindermann 2023; Perez et al. 2022; Jacob 2022, 2023).

While RLHF can create helpful agents from GPT models like ChatGPT (OpenAI 2022) or GPT-4 (OpenAI 2023), it should not be considered a reliable alignment method as it directly optimizes to deceive human evaluators (Cotra 2022).

Simulator Alignment

AI alignment efforts predominantly focus on preventing adverse outcomes. However, it is crucial also to consider the potential positive implications of achieving AI alignment. This section explores a vision for aligned superintelligent AI and examines two possible manifestations of successfully aligned AI systems: cyborg and emulation.

Cyborg

In a perfectly aligned scenario, the AI system becomes an inseparable part of the user's extended self. This alignment means the AI system is deeply integrated with the user's goals and values. Similar to how the neocortex aligns with primitive drives in the human brain, facilitating cohesive and integrated functioning, the AI system should harmonize with the user's objectives and seamlessly integrate them into its decision-making process. We illustrate this concept in Fig. 3. This alignment ensures that the AI system acts as a cognitive extension of the user's mind, connecting goals to actions. It prompts questions about whether reinforcement learning from human feedback (RLHF), exemplified by systems like ChatGPT, can be seen as an initial step towards such an extension.

[1]Note that when we mention GPT, we are specifically referring to the original self-supervised foundation model. As a result, we do not classify GPT-4 as a GPT model since it undergoes fine-tuning with additional objectives, such as through RLHF.

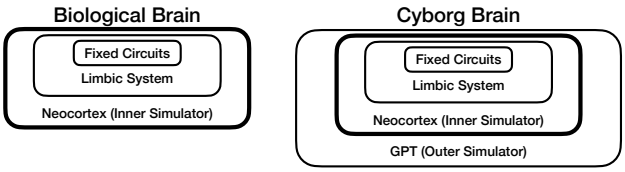

Figure 3: Extending human cognition with another layer of predictive coding. Figure reproduced after (NicholasKees and janus 2023).

Emulation

Another approach is simulating a human's mind through whole-brain emulation (WBE). While WBE remains a hypothetical technique that constructs a detailed 3D model of a person's brain and simulates it on a computer, its realization may only occur in the era of superintelligent AI. However, it might be feasible to simulate key aspects of human cognition without biological realism, such as the ability to simulate moral reasoning. Cognitive emulations (CE), a subset of WBEs that focus solely on simulating cognitive rather than biological functions, hold promise. Language models, such as LLMs, can be the foundation for CE as they already simulate human cognitive functions within their simulacra. However, the successful implementation of CE requires advancements in interpretability and digital neuroscience. It is important to note that even in the optimal outcome, CE would exhibit superhuman capabilities, potentially out-competing humans in various domains. However, this would not lead to loss of control by biological humans but rather a transition to a potentially worthy successor, avoiding the risk of a mere paperclip maximizer.

These scenarios for aligned superintelligence envision a future where AI systems seamlessly integrate with human minds, acting as extensions of human cognition. It also explores the possibilities of simulating human minds through emulation approaches. These advancements have the potential to foster harmonious human-AI collaboration, augment human capabilities, and ensure ethical and value-aligned behavior.

Conclusion

In conclusion, aligning humans and AI is crucial to prevent undesirable outcomes as AI systems advance. This position paper has emphasized the importance of aligning simulacra and agents derived from prediction-trained models. By addressing the challenges associated with misalignment and proposing strategies for aligning AI systems with human

values, we contribute to the ongoing efforts of safeguarding human values in developing and deploying AI technologies.

The vision presented here envisions a future where aligned simulators play a pivotal role in successful human-AI collaboration. By exploring emulation approaches and integrating simulators into cyborg systems, we can enhance human cognitive abilities, enable shared decision-making, and ensure ethical and value-aligned behavior.

In summary, the alignment between humans and AI is a moral imperative that requires continuous attention and proactive measures. By aligning AI systems with human values, we can shape a future where AI technologies contribute to human flourishing and societal progress.

Acknowledgements

This research was conducted with the support of the European Research Council (ERC), via the Starting Grant (No. 950086) for Project EVA.

References

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete Problems in AI Safety. arXiv:1606.06565.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; Brynjolfsson, E.; Buch, S.; Card, D.; Castellon, R.; Chatterji, N.; Chen, A.; Creel, K.; Davis, J. Q.; Demszky, D.; Donahue, C.; Doumbouya, M.; Durmus, E.; Ermon, S.; Etchemendy, J.; Ethayarajh, K.; Fei-Fei, L.; Finn, C.; Gale, T.; Gillespie, L.; Goel, K.; Goodman, N.; Grossman, S.; Guha, N.; Hashimoto, T.; Henderson, P.; Hewitt, J.; Ho, D. E.; Hong, J.; Hsu, K.; Huang, J.; Icard, T.; Jain, S.; Jurafsky, D.; Kalluri, P.; Karamcheti, S.; Keeling, G.; Khani, F.; Khattab, O.; Koh, P. W.; Krass, M.; Krishna, R.; Kuditipudi, R.; Kumar, A.; Ladhak, F.; Lee, M.; Lee, T.; Leskovec, J.; Levent, I.; Li, X. L.; Li, X.; Ma, T.; Malik, A.; Manning, C. D.; Mirchandani, S.; Mitchell, E.; Munyikwa, Z.; Nair, S.; Narayan, A.; Narayanan, D.; Newman, B.; Nie, A.; Niebles, J. C.; Nilforoshan, H.; Nyarko, J.; Ogut, G.; Orr, L.; Papadimitriou, I.; Park, J. S.; Piech, C.; Portelance, E.; Potts, C.; Raghunathan, A.; Reich, R.; Ren, H.; Rong, F.; Roohani, Y.; Ruiz, C.; Ryan, J.; Ré, C.; Sadigh, D.; Sagawa, S.; Santhanam, K.; Shih, A.; Srinivasan, K.; Tamkin, A.; Taori, R.; Thomas, A. W.; Tramèr, F.; Wang, R. E.; Wang, W.; Wu, B.; Wu, J.; Wu, Y.; Xie, S. M.; Yasunaga, M.; You, J.; Zaharia, M.; Zhang, M.; Zhang, T.; Zhang, X.; Zhang, Y.; Zheng, L.; Zhou, K.; and Liang, P. 2022. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258.

Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press, illustrated edition edition. ISBN 978-0-19-967811-2.

Bowman, S. R. 2023. Eight Things to Know about Large Language Models. arXiv:2304.00612.

Bowman, S. R.; Hyun, J.; Perez, E.; Chen, E.; Pettit, C.; Heiner, S.; Lukošiūtė, K.; Askell, A.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; Olah, C.; Amodei, D.; Amodei, D.; Drain, D.; Li, D.; Tran-Johnson, E.; Kernion, J.; Kerr, J.; Mueller, J.; Ladish, J.; Landau, J.; Ndousse, K.; Lovitt, L.; Elhage, N.; Schiefer, N.; Joseph, N.; Mercado, N.; DasSarma, N.; Larson, R.; McCandlish, S.; Kundu, S.; Johnston, S.; Kravec, S.; Showk, S. E.; Fort, S.; Telleen-Lawton, T.; Brown, T.; Henighan, T.; Hume, T.; Bai, Y.; Hatfield-Dodds, Z.; Mann, B.; and Kaplan, J. 2022. Measuring Progress on Scalable Oversight for Large Language Models. arXiv:2211.03540.

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. arXiv:2303.12712.

Burns, C.; Ye, H.; Klein, D.; and Steinhardt, J. 2022. Discovering Latent Knowledge in Language Models Without Supervision. arXiv:2212.03827.

Christiano, P.; Leike, J.; Brown, T. B.; Martic, M.; Legg, S.; and Amodei, D. 2023. Deep reinforcement learning from human preferences. arXiv:1706.03741.

Christiano, P.; Shlegeris, B.; and Amodei, D. 2018. Supervising strong learners by amplifying weak experts. arXiv:1810.08575.

Cotra, A. 2022. Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover. https://www.alignmentforum.org/posts/pRkFkzwKZ2zfa3R6H/without-specific-countermeasures-the-easiest-path-to. Accessed: 2023-05-15.

Hendrycks, D.; Carlini, N.; Schulman, J.; and Steinhardt, J. 2022. Unsolved Problems in ML Safety. arXiv:2109.13916.

Irving, G.; Christiano, P.; and Amodei, D. 2018. AI safety via debate. arXiv:1805.00899.

Jacob, S. 2022. ML Systems Will Have Weird Failure Modes. https://bounded-regret.ghost.io/ml-systems-will-have-weird-failure-modes-2/. Accessed: 2023-05-17.

Jacob, S. 2023. Emergent Deception and Emergent Optimization. https://bounded-regret.ghost.io/emergent-deception-optimization/. Accesssed: 2023-05-17.

Jan; Steiner, C.; Riggs, L.; janus; jacquesthibs; metasemi; Oesterle, M.; Teixeira, L.; peligrietzer; and remember. 2023. [Simulators seminar sequence] #1 Background & shared assumptions. https://www.lesswrong.com/posts/nmMorGE4MS4txzr8q/simulators-seminar-sequence-1-background-and-shared. Accessed: 2023-05-15.

janus. 2022. Simulators. https://www.lesswrong.com/posts/vJFdjigzmcXMhNTsx/simulators. Accessed: 2023-05-15.

janus. 2023. Simulacra are Things. https://www.lesswrong.com/posts/3BDqZMNSJDBg2oyvW/simulacra-are-things. Accessed: 2023-05-15.

Luscombe, R. 2022. Google engineer put on leave after saying AI chatbot has become sentient. https://www.theguardian.com/technology/2022/jun/12/google-engineer-ai-bot-sentient-blake-lemoine. Accessed: 2023-05-19.

Meng, K.; Bau, D.; Andonian, A.; and Belinkov, Y. 2023. Locating and Editing Factual Associations in GPT. arXiv:2202.05262.

Nardo, C. 2023. The Waluigi Effect (mega-post). https://www.lesswrong.com/posts/D7PumeYTDPfBTp3i7/the-waluigi-effect-mega-post. Accessed: 2023-05-15.

Ngo, R.; Chan, L.; and Mindermann, S. 2023. The alignment problem from a deep learning perspective. arXiv:2209.00626.

NicholasKees; and janus. 2023. Cyborgism. https://www.lesswrong.com/posts/bxt7uCiHam4QXrQAA/cyborgism. Accessed: 2023-05-17.

OpenAI. 2022. Introducing ChatGPT. https://openai.com/blog/chatgpt. Accessed: 2023-05-18.

OpenAI. 2023. GPT-4 Technical Report. http://arxiv.org/abs/2303.08774. arXiv:2303.08774.

Perez, E.; Ringer, S.; Lukošiūtė, K.; Nguyen, K.; Chen, E.; Heiner, S.; Pettit, C.; Olsson, C.; Kundu, S.; Kadavath, S.; Jones, A.; Chen, A.; Mann, B.; Israel, B.; Seethor, B.; McKinnon, C.; Olah, C.; Yan, D.; Amodei, D.; Amodei, D.; Drain, D.; Li, D.; Tran-Johnson, E.; Khundadze, G.; Kernion, J.; Landis, J.; Kerr, J.; Mueller, J.; Hyun, J.; Landau, J.; Ndousse, K.; Goldberg, L.; Lovitt, L.; Lucas, M.; Sellitto, M.; Zhang, M.; Kingsland, N.; Elhage, N.; Joseph, N.; Mercado, N.; DasSarma, N.; Rausch, O.; Larson, R.; McCandlish, S.; Johnston, S.; Kravec, S.; Showk, S. E.; Lanham, T.; Telleen-Lawton, T.; Brown, T.; Henighan, T.; Hume, T.; Bai, Y.; Hatfield-Dodds, Z.; Clark, J.; Bowman, S. R.; Askell, A.; Grosse, R.; Hernandez, D.; Ganguli, D.; Hubinger, E.; Schiefer, N.; and Kaplan, J. 2022. Discovering Language Model Behaviors with Model-Written Evaluations. arXiv:2212.09251.

Russell, S. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. New York?: Viking, illustrated edition edition. ISBN 978-0-525-55861-3.

Turner, A. M.; Smith, L.; Shah, R.; Critch, A.; and Tadepalli, P. 2023. Optimal Policies Tend to Seek Power. arXiv:1912.01683.

Wolf, Y.; Wies, N.; Levine, Y.; and Shashua, A. 2023. Fundamental Limitations of Alignment in Large Language Models. arXiv:2304.11082.

Yudkowsky, E. 2016. AI Alignment: Why It's Hard, and Where to Start. https://intelligence.org/2016/12/28/ai-alignment-why-its-hard-and-where-to-start/. Accessed: 2023-05-17.