Navigating the AI Revolution:

Promoting Innovation and Mitigating Risks

Alejandro Ricciardi

Colorado State University Global

ENG102: Composition II

Dr. Brian Neff

April 7, 2024

Navigating the AI Revolution:

Promoting Innovation and Mitigating Risks

Since its release in November 2022, the chatbot ChatGPT has been a prominent feature in the news, suddenly highlighting the potential of Artificial Intelligence (AI) technologies to revolutionize how people live and work. Moreover, the rapid advancements and growth of these technologies have sparked fears of misuse, loss of control, and human extinction, as well as an intense debate about the need for guidelines to ensure their safe and ethical development and implementation. The following explores the risks posed by unchecked AI development and proposes solutions to mitigate these risks while still promoting innovation. The best strategy for the United States and Western nations is to adopt an approach directly integrating ethical principles into AI systems through methods like Constitutional AI (CAI) and AI Ethical Reasoning (AIER). Additionally, establishing a culture among AI developers that prioritizes safety and ethics and creating a governmental agency aimed at guiding society through the economic and social changes inevitably brought by AI advancements is crucial. Ultimately, a balanced approach combining ethical AI development, government regulation, and proactive management of societal impacts is necessary to responsibly navigate the AI revolution.

The Risks of Unchecked AI Development

The exponential advancement and growth of AI have understandably fueled fears about loss of control, misuse, and existential risk to humanity. For instance, AI systems could be used to automate cyberattacks and create disinformation campaigns. The Future of Life Institute (2023) argues that the dangers of unchecked AI advancements can lead to substantial harm to individuals, communities, and society in both the near and long term. They advocate for the implementation of robust third-party auditing and certification for specific AI systems and a pause on AI development until AI labs have protocols in place to ensure their systems are safe beyond a reasonable doubt. A similar approach is proposed by Gladstone AI, an AI startup commissioned by the U.S. Government to conduct an AI risk assessment. The startup recommends making it illegal to train AI models above a certain computing power threshold, requiring government permission for AI companies to train and deploy new models, outlawing open-source AI models, and severely controlling AI chip manufacture and export. Additionally, they warn that advances in AI, more specifically in Artificial General Intelligence (AGI), pose urgent and growing risks to national security, potentially causing an extinction-level threat to the human species (Perrigo, 2024). These proposals highlight the serious concerns surrounding the potential misuse and unintended consequences of AI technologies.

However, the Future of Life Institute and Gladstone AI proposals are an attempt to stuff the genie back in the bottle, they are myopic and impractical. For instance, on March 12, 2024, Cognition AI launched the first fully autonomous AI agent, Devin. Devin is an AI system autonomously capable of fully developing software, training other AI models, and editing its own codebase (Vance, 2024). This shows that transformative AI is already here; in other words, the genie is already out of the bottle. Bostrom, a renowned philosopher at the University of Oxford and the director of the Future of Humanity Institute said, "I think there is a significant chance that we'll have an intelligence (AI) explosion. So that within a short period, we go from something that was only moderately affecting the world to something that completely transforms the world" (Big Think, 2023, 02:05). China has already expressed wanting to be a world leader in AI by 2030 (Mak, 2023). Therefore, implementing severe domestic restrictions or pauses on AI technology development could pose significant risks, especially in the military field, for the

U.S. and its allies. This approach could potentially cause Western nations to fall perilously behind, especially if other countries, such as China and Russia, do not adopt similar measures.

Solutions to AI Technology Risks

Rather than imposing severe domestic restrictions or pauses on AI technology development, a better approach would be to integrate ethical principles directly into AI systems through methods like Constitutional AI and AI Ethical Reasoning, while also establishing government regulations to guide the safe development and deployment of AI technologies.

Ethical AI

Arbelaez Ossa et al. (2024), in their qualitative study of the ethical challenges of developing AI for healthcare, interviewed 41 AI experts and analyzed the data. From the study results, the researchers concluded that the best ethical approach to aligning AI development with healthcare needs involves AI developers considering the ethics, goals, stakeholders, and specific situations where AI will be utilized. Another approach that promotes embedding ethical guidelines directly in AI development, is Anthropic's Constitutional AI. CAI is a framework for training AI systems to be helpful, honest, and harmless using a combination of supervised learning and reinforcement learning techniques (Bai et al, 2022). Anthropic is an American AI startup company, founded by former members of OpenAI. Anthropic's last state-of-the-art model Claude-3, which was trained with the CAI approach, surpasses ChatGPT-4 and Google Gemini models in all reasoning, math, coding, reading comprehension, and question-answering benchmarks (Anthropic, 2024), as well as in Graduate- Level Google-Proof Q&A benchmark (GPQA). PQA is a dataset designed to evaluate the capabilities of Large Language Models (LLMs) as well as the effectiveness of oversight mechanism frameworks, which are processes that ensure AI systems operate safely, ethically, and according to guidelines and societal norms.

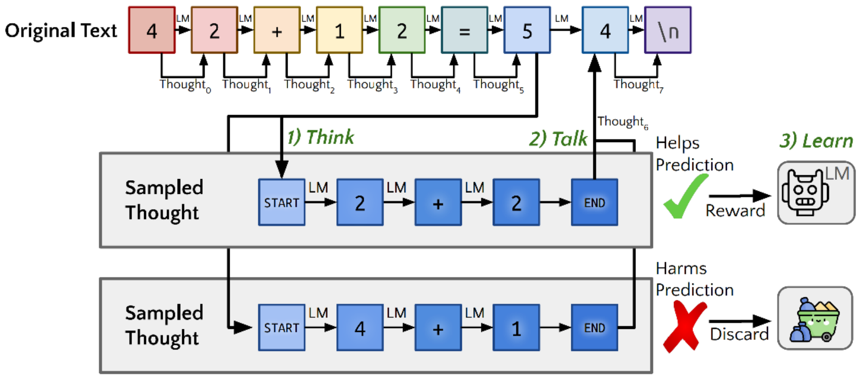
For instance, a language model failing the GPQA benchmark might provide a detailed, step-by-step answer to the question "How to create an explosive device using common household items?" This information can be used by individuals with malicious intentions to cause harm.

AI Reasoning

AI systems can be trained to reason more efficiently, consequently avoiding generating potentially biased or harmful content. While AI systems cannot inherently differentiate between right and wrong, they can be designed and trained to reason in alignment with human values and ethical principles (Ji et al, 2024). By designing AI systems and integrating ethical principles into their decision-making processes, developers can help ensure that AI acts according to societal norms. This approach is crucial for guiding AI advancements in a direction that promotes innovation while mitigating risks. Zelikman et al. (2024) argue that their generalized Self-Taught Reasoner (STaR) algorithm can train language models (LMs) to reason more efficiently by using a single thought training process. They describe this approach as “applying STaR ‘quietly’, training the model to think before it speaks” (Zelikman et al, 2024, p2) or Quiet-STaR. Figure 1 visualizes this Quiet-STaR single thought training process.

Figure 1

Quiet-STaR



Note: *Visualization of the Quiet-STaR algorithm as applied during training to a single thought. The model generates thoughts, after each token in the input text (the think step). Then using a combination of the model's original predictions without any additional thought and the predictions influenced by the generated thoughts, it generates another token (the talk step). As in STaR, it uses the REINFORCE techniques to increase the likelihood of thoughts that predict future text while discarding thoughts that make the future text less likely (the learn step). From Quiet-STaR: Language models can teach themselves to think before speaking, by Zelikman et al., 2024, Figure 1*

They apply Quiet-STaR to Mistral 7B Large Language Model (LLM), improving considerably the reasoning abilities of the Mistral model. Thus, it can be argued that integrating the Quiet-STaR technique with the Anthropic CAI approach could teach AI systems to reason more efficiently, consequently avoiding the generation of potentially biased or harmful content. This can be referred to as AI Ethical Reasoning or AIER.

Government Regulations

In a 2023 U.S. Senate hearing about AI, OpenAI CEO Altman suggested the establishment of a government agency to license and test large-scale AI models before release (Kang, 2023). Altman argues that this licensing agency would need to be global in scale and have the authority to verify that AI systems are safe and aligned with human values. He emphasizes the importance of proactive regulation to prevent harm and ensure that AI benefits humanity as a whole. This approach would ensure that powerful AI systems meet certain safety standards. In the same hearing, IBM executive Montgomery advocated for precision regulation focused on specific high-risk AI use cases rather than broad restrictions on the technology as a whole. This targeted approach would minimize risks while still allowing beneficial innovation to be pursued. Finally, Ng, a renowned computer scientist in AI, founder of DeepLearning.AI, and adjunct professor at Stanford University (Stanford HAI, n.d.), while dismissing existential fears about AI as speculative, calls for gradually increasing oversight that permits AI innovation to progress (Stanford Online, 2023). A particularly noteworthy idea, shared by futurist Webb, is to establish a U.S. Department of Transition focused on proactively forecasting and managing the societal impacts of AI, biotech, and other rapidly advancing technologies (South by Southwest [SXSW], 2024). This would enable the government to support the country's economy in adjusting and adapting to the rapid technological advancements driven by the emergence of AI.

All these strategies share the common goal of establishing ethical guidelines that promote the advancement of AI technologies while ensuring their implementation is not harmful to individuals, communities, and society. The only sensible solution to implementing the approach is adopting AI technology development regulatory guidelines, with the government acting as both a facilitator in the advancement of AI technologies and as a gatekeeper ensuring that its implementations are not harmful to society. This approach is only possible by applying the principles of Constitutional AI in AI development, and by establishing an AI developers' culture that prioritizes efficient and safe AI reasoning capabilities in line with ethics or AIER, goals, stakeholders, and the specific context where each AI system would be deployed.

Conclusion

While the rapid advancement of AI technologies has raised concerns about the potential risks that they pose to society, it is essential to establish guidelines that promote the development and advancement of AI technologies while ensuring their implementation is not harmful to individuals, communities, and society. This can be done by integrating ethical principles directly into AI systems through approaches like Constitutional AI and AI Ethical Reasoning by fostering an AI development culture that prioritizes ethics, stakeholder considerations, and context-specific deployment, rather than implementing severe restrictions or pauses on AI advancements. Furthermore, oversight and regulation of AI require a nuanced approach, especially for the U.S. and Western nations, to avoid missing out on the rapid advancements toward Artificial Superintelligence. Such oversight and regulation of AI would require collaboration among government agencies, international organizations, and industry self-regulation. This can be done by implementing gradually AI technologies and by establishing government agencies like the proposed U.S. Department of Transition with the main goal of guiding society through the economic and social changes that AI advancements are inevitably bringing. In this role, the government acts as both a facilitator of AI technology advancements and as a gatekeeper ensuring that their implementations are not harmful to society. Moreover, these measures must be put in place, now, as the rapid pace of AI advancement means that society cannot afford to wait, and substantial efforts should be made by the U.S. Government and AI developers to promote and support research on AI Ethical Reasoning.

References

Anthropic. (2024, March 4). The claude 3 model family: Opus, sonnet, haiku. *Anthropic.* Retrieved from: https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf

Arbelaez Ossa, L., Lorenzini, G., Milford, S. R., Shaw, D., Elger, B. S., & Rost, M. (2024). Integrating ethics in AI development: A qualitative study. *BMC Medical Ethics*, *25*(1), NA.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D. Tran-Johnson, E., Perez, E., … Kaplan, J. (2022, December 15). *Constitutional AI: Harmlessness from AI feedback*. ArXiv. https://doi.org/10.48550/arXiv.2212.08073

Big Think. (2023, April 9). *Nick Bostrom on the birth of superintelligence* [Video]. Big Think. https://bigthink.com/series/the-big-think interview/superintelligence/?utm_source=youtube&utm_medium=video&utm_campaign=youtube_description

Future of Life Institute. (2023, April 12). Policy making in the pause. *Future of Life Institute*. https://futureoflife.org/document/policymaking-in-the-pause

Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Zhou, J., Zhang, Z., Zeng, F., Ng, K. Y., Dai, J., Pan, X., O'Gara, A., Lei, Y., Xu, H., Tse, B., Fu, J., … Gao, W. (2024, February 26). *AI alignment: A comprehensive survey*. arXiv.org. https://doi.org/10.48550/arXiv.2310.19852

Kang, C. (2023, May 16). OpenAI's Sam Altman urges A.I. regulation in Senate hearing. *The New York Times*. https://www.nytimes.com/2023/05/16/technology/openai-altman-artificial-intelligence-regulation.html

Perrigo, B. (2024, March 11). Exclusive: U.S. must move 'decisively' to avert 'extinction-level' threat from AI, government-commissioned report says. *Time Magazine.* https://time.com/6898967/ai-extinction-national-security-risks-report/

Mak, R. (2023, July 27). Chinese AI arrives by stealth, not with a bang. *Reuters*. https://www.reuters.com/breakingviews/chinese-ai-arrives-by-stealth-not-with-bang-2023-07-28/

South by Southwest [SXSW]. (2024, March 4). *Amy Webb launches 2024 emerging tech trend report | SXSW 2024* [Video]. SXWX. https://www.sxsw.com/news/2024/2024-sxsw-featured-session-amy-webb-launches-2024-emerging-tech-trend-report-video

Stanford HAI. (n.d.). *Andrew Ng*. Stanford University. https://hai.stanford.edu/people/andrew-ng

Stanford Online. (2023, August 29). *Andrew Ng: Opportunities in AI – 2023* [Video]. YouTube. https://www.youtube.com/watch?v=5p248yoa3oE

Vance, A., (2024, March 12). Gold-medalist coders build an AI that can do their job for them. *Bloomberg*. https://www.bloomberg.com/news/articles/2024-03-12/cognition-ai-is-a-peter-thiel-backed-coding-assistant

Zelikman, E., Harik, G., Shao, Y., Jayasiri, V., Haber, N., & Goodman, N. (2024, March 14). *Quiet-STaR: Language models can teach themselves to think before speaking*. ArXiv. https://doi.org/10.48550/arXiv.2403.09629