**An Analysis of AI Ethics in the Wake of OpenAI’s Superalignment Controversy**

Jamie Kemman

*Department of Computer Sciences, Mathematics, and Engineering*

*Shepherd University*

*P.O. Box 3210*

*Shepherdstown, WV 25443y = 0*

*jkemma01@rams.shepherd.edu*

**ABSTRACT**

In May of 2024, the prominent researchers heading OpenAI’s superalignment team left the company following the culmination of silent internal power struggles and priority conflicts. Many junior engineers and researchers have done likewise. Since then, OpenAI’s public facing documentation of the team’s efforts to influence the ethical framework around Artificial General Intelligence have gone unrevised despite reports that the team has disbanded entirely. This paper seeks to understand the implications of this development by analyzing recent discussion in AI ethics with a focus on AGI superalignment in particular.

**KEYWORDS:** OpenAI Ethics AI AGI Superalignment

**REFERENCES** <No Section Number>

List and number all articles which you actually used and cited in the paper. When referenced in the text, enclose the citation number in square brackets, for example [1]. References must have full citation.

[1] Leike, Jan. Jan Leike. <https://jan.leike.name> (accessed September 2, 2024).

[2] Samuel, Sigal. “‘I lost trust’: Why the OpenAI team in charge of safeguarding humanity imploded.” Vox. <https://www.vox.com/future-perfect/2024/5/17/24158403/openai-resignations-ai-safety-ilya-sutskever-jan-leike-artificial-intelligence> (accessed September 2, 2024).

[3] OpenAI. “Introducing Superalignment.” <https://openai.com/index/introducing-superalignment/> (accessed September 2, 2024).

[4] Nourbakhsh, Illah Reza, “AI Ethics: A Call to Faculty,” *Communications of the ACM,* Vol. 64, no. 9, Sept. 2021, doi: 10.1145/3478516.

[5]Bleher, Hanna and Braun, Matthias, “Reflections on Putting AI Ethics into Practice: How Three AI Ethics Approaches Conceptualize Theory and Practice,” *Science and Engineering Ethics,* Vol. 29, no. 21, May 2023, <https://doi.org/10.1007/s11948-023-00443-3>.

[6]Hagendorff, Thilo, “The Ethics of AI Ethics: An Evaluation of Guidelines,” *Minds and Machines, Vol. 30, pp. 99-120,* Feb. 2020, https://doi.org/10.1007/s11023-020-09517-8

[7] Borenstein, Jason and Grodzinsky, Frances S. et al., “AI Ethics: A Long History and a Recent Burst of Attention,” *Computer,* Vol. 54, no. 1, pp. 96-102, Jan. 2021, doi: 10.1109/MC.2020.303495

[8] OpenAI. “Our approach to alignment research.” <https://openai.com/index/our-approach-to-alignment-research/> (accessed September 9, 2024).

[9] OpenAI. “Charter.” <https://openai.com/charter/> (accessed September 9, 2024).

[10] OpenAI. “Safety.” <https://openai.com/safety/> (accessed September 9, 2024).