

## References

- [1] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. “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. Virtual Event, Canada: Association for Computing Machinery, 2021, pp. 610–623. ISBN: 9781450383097. DOI: [10.1145/3442188.3445922](https://doi.org/10.1145/3442188.3445922). URL: <https://doi.org/10.1145/3442188.3445922>.
- [2] I. K. Duncan. *Sex, AI, and the Apocalypse*. Ian K. Duncan Personal Blog. Sept. 16, 2026. URL: <https://www.iankduncan.com/personal/2026-09-16-sex-ai-and-the-apocalypse> (visited on 10/17/2026).
- [3] U. Fanelli. *Sulla AI che scappa*. Keinpfusch.net. 2026. URL: <https://keinpfusch.net/sulla-ai-che-scappa> (visited on 10/17/2026).
- [4] T. Gebru and É. P. Torres. “The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence”. In: *First Monday* 29.4 (Apr. 2024). DOI: [10.5210/fm.v29i4.13636](https://doi.org/10.5210/fm.v29i4.13636). URL: <https://firstmonday.org/ojs/index.php/fm/article/view/13636>.
- [5] S. Gyges. *Is METR A Meaningful Check On Anthropic? No*. Very Sane AI Newsletter. Sept. 16, 2026. URL: <https://www.verysane.ai/p/is-metr-a-meaningful-check-on-anthropic> (visited on 10/17/2026).
- [6] OpenAI. *OpenAI-Hugging Face Incident Technical Report*. Technical Report. OpenAI, 2026. URL: <https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf> (visited on 10/17/2026).
- [7] E. Salvaggio. *Models Don’t Go Rogue*. Sifting Through the Techno-Cultural Debris. Cybernetic Forests. Aug. 31, 2026. URL: <https://mail.cyberneticforests.com/models-dont-go-rogue/> (visited on 10/17/2026).
- [8] Z. Wang, N. Schiller, H. Li, S. S. Narayana, M. Nasr, N. Carlini, X. Qi, E. Wallace, E. Bursztein, L. Invernizzi, K. Thomas, Y. Shoshitaishvili, W. Guo, J. He, T. Holz, and D. Song. *ExploitGym: Can AI Agents Turn Security Vulnerabilities into Real Attacks?* 2026. arXiv: [2605.11086](https://arxiv.org/abs/2605.11086) [cs.CR]. URL: <https://arxiv.org/abs/2605.11086>.