Mechanistic Interpretability –

Joonas Virtanen

February 5, 2025

1 Introduction

Mechanistic interpretability review for AI safety [1]

Literature

- ”Mechanistic Interpretability for AI Safety” [1]: Broad surface level review of MI in the context of AI safety
- ”Everything, Everywhere, All at Once: Is Mechanistic Interpretability Identifiable?” [3]: Main question of the paper: For a fixed behaviour under criterias that MI sets for itself, is it guaranteed that there exists unique explanaiton?
- ”Position: An Inner Interpretability Framework for AI Inspired by Lessons from Cognitive Neuroscience” [5]: Trying to answer possible critisism of MI, e.g. lack of rigor and general framework, the paper describes inner interpretability frameworks derived from the field of Cognitive Neuroscience.
- ”Mechanistic?” [6]: Describes the term ”mechanistic” in the context of MI, how is it used and presents a history of the NLP interpretability community and formation of the separate MI community.
- ”Explaining AI through mechanistic interpretability” [2]: ...
- ”A Systematic Literature Review on AI Safety: Identifying Trends, Challenges, and Future Directions” [4]: ...
- etc

Less academic sources, blogs, etc.

- Neel Nanda: ...
- Chris Olah: ...

References

[1] Leonard Bereska and Efstratios Gavves. *Mechanistic Interpretability for AI Safety – A Review.* 2024. arXiv: 2404.14082 [cs.AI]. URL: https://arxiv.org/abs/2404.14082.

[2] Lena Kästner and Barnaby Crook. "Explaining AI through mechanistic interpretability". In: *European Journal for Philosophy of Science* 14.4 (Oct. 2024), p. 52. ISSN: 1879-4920. DOI: 10.1007/s13194-024-00614-4. URL: https://doi.org/10.1007/s13194-024-00614-4.

[3] Maxime Méloux et al. "Everything, Everywhere, All at Once: Is Mechanistic Interpretability Identifiable?" In: *The Thirteenth International Conference on Learning Representations.* 2025. URL: https://openreview.net/forum?id=5IWJBStfU7.

[4] Wissam Salhab et al. "A Systematic Literature Review on AI Safety: Identifying Trends, Challenges, and Future Directions". In: *IEEE Access* 12 (2024), pp. 131762–131784. DOI: 10.1109/ACCESS.2024.3440647.

[5] Martina G. Vilas et al. *Position: An Inner Interpretability Framework for AI Inspired by Lessons from Cognitive Neuroscience.* 2024. arXiv: 2406.01352 [cs.AI]. URL: https://arxiv.org/abs/2406.01352.

[6] Sarah Wiegreffe and Naomi Saphra. "Mechanistic?" In: *The 7th BlackboxNLP Workshop.* 2024. URL: https://openreview.net/forum?id=schAf4BPtD.