Bálint Gyevnár

Towards Trustworthy Autonomous Systems via Conversations and Explanations

THE UNIVERSITY *of* EDINBURGH

Autonomous Agents
Research Group

- **Explainable AI (XAI) doesn’t work for people:**
  - Not for safety critical systems (Rudin; 2019)
  - Not for trust calibration and understanding (Miller; 2023)
  - And just in general (e.g., Wiegreffe and Yuval; 2019)
- **Explainability ≠ Transparency**
  - **Bridging the Transparency Gap: What Can Explainable AI Learn From the AI Act?** *Balint Gyevnar*, Nick Ferguson, Burkhard Schafer at ECAI 2023.
- **Explain like the people for the people**

Miller, Tim. 2023. “Explainable AI Is Dead, Long Live Explainable AI! Hypothesis-Driven Decision Support Using Evaluative AI.” In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 333–42. FAccT ’23. New York, NY, USA: Association for Computing Machinery.

Rudin, Cynthia. 2019. “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead.” *Nature Machine Intelligence* 1 (5): 206–15.

Wiegreffe, Sarah, and Yuval Pinter. 2019. “Attention Is Not Not Explanation.” In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, edited by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, 11–20. Hong Kong, China: Association for Computational Linguistics.

Counterfactual Causal Selection via Simulation

- Sample counterfactual worlds grounded in observation
- Variables most correlated have larger causal effect

Future work: Conversational Agent

- Iterative conversational framework
- User can guide the explanatory process

Causal Explanations for Sequential Decision-Making in Multi-Agent Systems. *Balint Gyevnar*, Cheng Wang, Christopher G. Lucas, Shay B. Cohen, Stefano V. Albrecht; 23rd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS), 2024.

HEADD: Human Explanations for Autonomous Driving Decisions. *Balint Gyevnar,* Cheng Wang, Christopher G. Lucas, Shay B. Cohen, Stefano V. Albrecht. [dataset]