Bálint Gyevnár

Edinburgh, UK | balint.gyevnar@ed.ac.uk | gbalint.me

Research Interests

PhD student focusing on multi-step LLM reasoning, explainable multi-agent systems, and AI safety with the goal of achieving trustworthy human-agent collaboration.

Skills

Programming: Python (PyTorch, vLLM, Pandas, uv, etc.), R, C++, C#, Haskell, etc;
Data analysis: qualitative coding, unsupervised topic modelling, graph analysis, mixed effects regression, statistical hypothesis testing, data visualization;
Languages: English, German, Japanese, Chinese, Hungarian.

Education

University of Edinburgh Sep. 2021 – May 2026 (est.)
PhD in Natural Language Processing with Integrated Studies *Edinburgh, UK*
Supervisors: Stefano Albrecht, Shay Cohen, Christopher Lucas

University of Edinburgh Sep. 2016 – May 2021
Integrated Master of Informatics *Edinburgh, UK*
Supervisor: Maria Wolters

Projects

Combining Multi-Step LLM Reasoning with World Simulators for Generating Complex Explanations Jan. 2025 – present

- Development of multi-step reasoning framework for complex explanation generation;
- Integration of LLM inference with world simulators in a RAG approach;
- Evaluation with a wide range of models (Llama, Qwen, Phi, GPT, etc.) and humans.

Bridging Shared Research Challenges Amid Responsible AI Wars Jul. 2024 – present

- Curation of corpus of 3K+ papers on AI safety and AI ethics;
- Qualitative data analysis and visualization (e.g., topic coding, graph analysis);
- Quantitative unsupervised topic modeling and analysis (e.g., BERTopic);

Causal Explanations for Decision-Making in Multi-Agent Systems Sep. 2021 – present

- Counterfactual reasoning with RL planning for causally-grounded explanations in natural language;
- Two large-scale human subjects studies to evaluate natural and automatically generated explanations;
- Curation of HEADD: the Human Explanations for Autonomous Driving Decisions dataset.

Experience

Teaching Assistant Sep. 2019 – present
University of Edinburgh *Edinburgh, UK*

- Teaching assistant for “Evaluating Sustainable Lands & Cities” and “Data Mobility & Infrastructure”;
- Supervision of master’s students and tutor for ∼12 students for machine learning;
- Marker for courses in natural language processing, reinforcement learning, and machine learning.

Research Intern May 2020 – Oct. 2020
Five AI Ltd. *Edinburgh, UK*

- Development and evaluation of goal-based interpretable prediction and planning for autonomous vehicles;
- Scenario-based and open-world testing and results collection;
- Main contributor of open-source implementation on GitHub with added support for CARLA.

Volunteering

Sports Club Executive Member Sep. 2022 – Jun. 2025
Edinburgh University Volleyball Club *Edinburgh, UK*

- (2024-25; Secretary) Public outreach and networking with alumni members and organizing an event series;
- (2023-24; VP) Large-scale events, public speaking, timetabling, HR management of 220+ members;
- (2022-23; Treasurer) Setting up an annual budget, and managing a cash flow of £70k.

Awards

Colours Award for Outstanding Volunteering Contribution to Sports Jun. 2024
Edinburgh University Sports Union *Edinburgh, UK*

AI100 Early Career Essay Competition Featured Essay Aug. 2023
One Hundred Year Study on Artificial Intelligence (AI100) *Stanford University*

Trustworthy Autonomous Systems Early Career Researcher Award Jun. 2023
4,000 GBP; UK Research & Innovation *Southampton, UK*

Shape the Future of ITS Competition; 3rd Place Aug. 2022
1,000 USD; IEEE Intelligent Transportation Systems Society *USA*

Selected Publications

- **AI Safety for Everyone**
  Nature Machine Intelligence, 2025;
  B. Gyevnar, A. Kasirzadeh
- **Objective Metrics for Human-Subjects Evaluation in Explainable Reinforcement Learning**
  Multi-Disciplinary Conference on Reinforcement Learning and Decision Making, RLDM 2025
  B. Gyevnar, M. Towers
- **People Attribute Purpose to Autonomous Vehicles When Explaining Their Behavior**
  ACM Conference on Human Factors in Computing Systems, CHI 2025;
  B. Gyevnar, S. Droop, T. Quillien, S.B. Cohen, N.R. Bramley, C.G. Lucas, S.V. Albrecht.
- **Causal Explanations for Sequential Decision-Making in Multi-Agent Systems**
  23rd International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2024;
  B. Gyevnar, C. Wang, S.B. Cohen, C.G. Lucas, S.V. Albrecht.
- **Explainable AI for Safe and Trustworthy Autonomous Driving: A Systematic Review**
  IEEE Transactions on Intelligent Transportation Systems, 2024;
  A. Kuznietsov*, B. Gyevnar*, C. Wang, S. Peters, S.V. Albrecht. [* equal contribution]
- **Bridging the Transparency Gap: What Can Explainable AI Learn From the AI Act?**
  26th European Conference on Artificial Intelligence, ECAI 2023;
  B. Gyevnar, N. Ferguson, B. Schafer.
- **A Human-Centric Method for Generating Causal Explanations in Natural Language for Autonomous Vehicle Motion Planning** [best paper runner-up]
  Workshop on Artificial Intelligence for Autonomous Driving, IJCAI 2022;
  B. Gyevnar, M. Tamborski, C. Wang, C.G. Lucas, S.B. Cohen, S.V. Albrecht.
- **Interpretable Goal-based Prediction and Planning for Autonomous Driving**
  International Conference on Robotics and Automation, ICRA 2021;
  S.V. Albrecht, C. Brewitt, J. Wilhelm, B. Gyevnar, F. Eiras, M. Dobre, S. Ramamoorthy.
- **GRIT: Fast, Interpretable, and Verifiable Goal Recognition with Learned Decision Trees for Autonomous Driving**
  IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2021;
  C. Brewitt, B. Gyevnar, S. Garcin., S.V. Albrecht.