Bálint Gyevnár

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

Research Interests

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

Skills

Programming: Python (PyTorch, vLLM, Transformers, uv, etc.), R, C++, C#, Haskell, etc;
Human-computer interaction: human-subjects experiments, online crowdsourcing, survey design;
Data analysis: qualitative coding, unsupervised topic modelling, graph analysis, mixed effects regression, statistical hypothesis testing, data visualization;
Languages: English, German, Japanese, Hungarian.

Education

PhD in Natural Language Processing 09/2021 – 05/2025 (est.)
University of Edinburgh *Edinburgh, UK*
Supervised by Stefano V. Albrecht, Shay B. Cohen, and Christopher G. Lucas

Bachelor and Master of Informatics 09/2016 – 05/2021
University of Edinburgh *Edinburgh, UK*
Supervised by Maria Wolters

Academic Exchange in Computer Science 08/2018 – 06/2019
Nanyang Technological University *Singapore*

Experience

Research Assistant Jul. 2023 – Dec. 2024
University of Edinburgh *Edinburgh, UK*

- Researching the intersection of AI safety and AI ethics to build bridges among research problems;
- Quantitative literature analysis with unsupervised natural language processing and network analysis;
- Curation, topic coding, and qualitative analysis of large corpora of papers.

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 against baselines;
- Open-source implementation on GitHub with added support for CARLA.

Volunteering

Sports Club Executive Committee 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.

Teaching

- Teaching assistant for machine learning, computing systems, data analysis courses (2020-present);
- Assistant supervisor to two master's students (2022-2023);
- Marker for natural language processing, reinforcement learning, and machine learning courses (2020-present).

Research 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 modelling and analysis (e.g., BERTopic);

Understanding how humans explain multi-agent systems May 2024 – Present

- Large-scale human-agent interaction study of human explanatory modes.
- Curation and release of HEADD: the Human Explanations for Autonomous Driving Decisions dataset.
- Quantitative analysis and statistical hypothesis testing in R.

Generating causal explanations for sequential decision-making in multi-agent systems Sep. 2021 – May 2024

- Conversion of Monte Carlo Tree Search to probabilistic graph for counterfactual inference;
- 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.

Studying how humans acquire/communicate colour naming systems Sep. 2021 – Nov. 2022

- Understanding the effects of communicative efficiency and acquisition on the patterns of human colour naming systems via computational information-theoretic measures.
- Simulating patterns of colour term acquisition using self-organising maps and the World Colour Survey.

Interpretable goal-based prediction and planning for autonomous vehicles May 2020 – May 2021

- Integration of rational inverse planning-based prediction module with Monte Carlo Tree Search for interpretable autonomous vehicle planning;
- Functional scenario-based and open-world evaluation against multiple baselines;
- Main developer and maintainer of open-source Python implementation.

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*

Peer-Reviewed 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: Insights From Cognitive Science for Explainable AI**
  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.
- **Towards Trustworthy Autonomous Systems via Conversations and Explanations**
  AAAI Conference on Artificial Intelligence, AAAI 2024;
  B. Gyevnar.
- **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.
- **Building Trustworthy Human-Centric Autonomous Systems Via Explanations**
  23rd International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2024;
  B. Gyevnar.
- **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.
- **Deep Reinforcement Learning for Multi-Agent Interaction**
  AI Communications, 35(4), 357-368, 2022;
  I.H. Ahmed, C. Brewitt, I. Carlucho, F. Christianos, M. Dunion, E. Fosong, S. Garcin, S. Guo, B. Gyevnar, T. McInroe, G. Papoudakis, A. Rahman, L. Schäfer, M. Tamborski, G. Vecchio, C. Wang, S.V. Albrecht.
- **Communicative Efficiency or Iconic Learning: Do Communicative and Acquisition Pressures Interact to Shape Colour-Naming Systems?**
  Entropy, 24(11), 1542, 2022;
  B. Gyevnar, G. Dagan, C. Haley, S. Guo, F. Mollica.
- **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.

Event Organiser

- **Evaluating Explainable AI and Complex Decision-Making**
  Workshop at the 28th European Conference on Artificial Intelligence, ECAI 2025;
  H. Baier, B. Gyevnar, M. Towers, Y. Zhang
- **The Explainable Reinforcement Learning Competition**
  Under review at NeurIPS 2025;
  M. Towers, B. Gyevnar, A. Nowé, C.G. Lucas, D. Abel, H. Baier, T. Miller, T.J. Norman, T. Huber, T. Bewley, S.V. Albrecht

Invited Talks

- **How do we make explainable AI work for people?**
  Machine Learning and Modelling Seminar Series, Charles University of Prague, 2024.

Peer Review

- IEEE Transactions on Intelligent Transportation Systems – *Reviewer (2025)*
- 3^{rd} World Conference on eXplainable Artificial Intelligence – *Program Chair (2025)*
- IEEE International Conference on Robotics and Automation – *Reviewer (2025)*
- IEEE/RSJ International Conference on Intelligent Robots and Systems – *Reviewer (2023)*