Bálint Gyevnár

Carnegie Mellon University, Pittsburgh, PA | bgyevnar@cmu.edu | gbalint.me

Research Interests

AI Scientists; Agentic Scientific Discovery; Metascience; Explainable Multi-Agent Systems; Responsible AI; AI Safety; Human-AI Interaction.

Education

PhD in Natural Language Processing Sep. 2021 – Jan. 2026
University of Edinburgh *Edinburgh, UK*
Thesis: Action Explanation of Multi-Agent Systems via Counterfactual Reasoning
Supervisors: Stefano V. Albrecht, Shay B. Cohen, and Christopher G. Lucas

Integrated Master of Informatics Sep. 2016 – May 2021
University of Edinburgh *Edinburgh, UK*
Thesis: Comparison of Account Survival on Twitter During the First Wave of COVID-19
Supervisor: Maria Wolters

Exchange Year in Computer Science Aug. 2018 – Jun. 2019
Nanyang Technological University *Singapore*

Academic Experience

Postdoctoral Research Associate Sep. 2025 – present
Carnegie Mellon University *Pittsburgh, PA*

- Investigating failure modes of agentic AI scientists, such as reward hacking, data poisoning, spin doctoring;
- Understanding divergent behaviours of human and AI during formalization of mathematics in Lean.

Teaching Assistant Sep. 2020 – May 2025
University of Edinburgh *Edinburgh, UK*

- Teaching assistant for courses in machine learning, computing systems, and two data analysis courses;
- Marker for natural language processing, reinforcement learning, machine learning courses.

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

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

Service

Academic Reviewing

- *Program committee member*: AAAI 2025, AIES 2025, FAccT 2026, XAI 2025;
- *Reviewer:* Nature Communications 2025, Artificial Intelligence Review 2025, Transactions on Intelligent Transportation Systems 2025, CHI 2026, ECAI 2023, EMNLP2025, ICRA 2025, IASEAI 2026, IROS 2025, NeurIPS 2025.

Workshop Organiser Oct. 2025
Evaluating Explainable AI and Complex Decision-Making *ECAI 2025*

Teaching

- *Teaching Assistant (Spring 2024):* Evaluating Sustainable Lands and Cities; Data, Mobility, Infrastructure;
- *Marker (Spring 2022-24):* Machine Learning Theory, Reinforcement Learning, Natural Language Processing
- *Tutor (Autumn 2020):* Introductory Applied Machine Learning; Introduction to Computing Systems;

Awards

- Colours Award for Outstanding Volunteering Contribution to Sports;
  Edinburgh University Sports Union; Jun. 2024.
- AI100 Early Career Essay Competition Featured Essay;
  One Hundred Year Study on Artificial Intelligence at Stanford University; Aug. 2023.
- £4,000 GBP Trustworthy Autonomous Systems Early Career Researcher Award;
  UK Research & Innovation; Jun. 2023.
- $1,000 USD Shape the Future of ITS Competition Award;
  IEEE Intelligent Transportation Systems Society; Aug. 2022.

Publications

- **Integrating Counterfactual Simulations with Language Models for Explaining Multi-Agent Behaviour**
  25th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2026;
  B. Gyevnar, Christopher G. Lucas, Stefano V. Albrecht, Shay B. Cohen.
- **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.

Invited Talks

- **AI safety for everyone**
  Talk at the 9th Workshop of the Center for Human-Compatible AI, 2025.
- **How do we make explainable AI work for people?**
  Invited talk at the Machine Learning and Modelling Seminar Series, Charles University of Prague, 2024.