Bálint Gyevnár

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

Research Statement: I am interested in the *metascience of AI*: how AI is used in the scientific process, its effects on scientific integrity, and how it changes the way scientists think about research.

Education

PhD. Natural Language Processing 09/2021 – 01/2026
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

MInf. Integrated Master of Informatics 09/2016 – 05/2021
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 08/2018 – 06/2019
Nanyang Technological University *Singapore*

Academic Experience

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

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

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, AIES, XAI;
- *Reviewer:* Nature Communications, Artificial Intelligence Review, Transactions on Intelligent Transportation Systems, CHI, ECAI, EMNLP, ICRA, IASEAI, IROS, NeurIPS.

Workshop Organiser October, 2025
Evaluating Explainable AI and Complex Decision-Making *ECAI 2025*

Sports Union 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.

Teaching

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

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*

Publications

- **Integrating Counterfactual Simulations with Language Models for Explaining Multi-Agent Behaviour**
  Under review
  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.