Bálint Gyevnár

Autonomous Agents Research Group
School of Informatics, University of Edinburgh
balint.gyevnar@ed.ac.uk | gbalint.me | github.com/gyevnarb

Education

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

Master of Informatics 09/2016 – 05/2021
University of Edinburgh *Edinburgh, UK*
Supervisor: Maria Wolters

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

Work Experience

Research Assistant 10/2023 – Present
University of Edinburgh (PI: Atoosa Kasirzadeh) *Edinburgh, UK*

- Assisting in the production of a **literature review** about the social and ethical opportunities and risks of the development and deployment of **generative models**.
- Formatting, cataloguing, and circulating findings to other stakeholders in line with project timescales and objectives.

Vice President 09/2022 – 06/2024
Edinburgh University Volleyball Club *Edinburgh, UK*

- (2023-24) Vice president responsible for large-scale event organisation, scheduling, public outreach, and the management of 8 teams, 11 coaches, and 220+ members.
- (2022-23) Treasurer managing a cash flow of approximately £70k, setting up an annual budget, and managing thousands of transactions.

Research Assistant 05/2020 – 10/2020
Five AI Ltd. (PI: Stefano Albrecht) *Edinburgh, UK*

- Developed and evaluated a goal-based interpretable prediction and planning system for autonomous vehicles with intuitive explanations.
- Rigorous scenario-based and open-world testing and evaluation of the proposed system.

Research Projects

Human-Centric Explanations in Natural Language for Trustworthy Autonomous Systems 09/2021 – Present

- **Inter-disciplinary** collaboration for combing research of autonomous agents, cognitive science, and natural language processing;
- Multi-stage large-scale **human evaluation** with online participants to measure effects of explanations on trust and understanding;
- Awarded by IEEE Intelligent Transportation Systems Society and UKRI Trustworthy Autonomous Hub;

Interpretable Goal-Based Prediction and Planning for Autonomous Vehicles 05/2020 – Present

- Helped develop and evaluate an integrated motion prediction and planning system based on rational inverse planning and **Monte Carlo Tree Search** (MCTS).
- Main developer and maintainer of open-source **Python** implementation with support for the **CARLA simulator** and extensive documentation (code).

Research Output

Awards

- AI100 Early Career Essay Competition Featured Essay, "Love, Sex, and AI", *Standing Committee of the One Hundred Year Study on Artificial Intelligence (AI100)*, Stanford University, 2023;
- Trustworthy Autonomous Systems Early Career Researcher Awards for £**4000**, Knowledge Transfer Track, *UK Research & Innovation*, 2023;
- Shape the Future of ITS Competition for $**1000**, "Cars that Explain: Building Trust in Autonomous Vehicles through Explanations and Conversations", *IEEE Intelligent Transportation Systems Society*, 2022.

Conference

- **B. Gyevnar**, C. Wang, S.B. Cohen, C.G. Lucas, S.V. Albrecht. "Causal Explanations for Sequential Decision-Making in Multi-Agent Systems", *Association for the Advancement of Artificial Intelligence (AAMAS)*, 2024;
- **B. Gyevnar**, N. Ferguson, B. Schafer. "Get Your Act Together: A Comparative View on Transparency in the AI Act and Technology", *European Conference on Artificial Intelligence (ECAI)*, 2023;
- S.V. Albrecht, C. Brewitt, J. Wilhelm, **B. Gyevnar**, F. Eiras, M. Dobre, S. Ramamoorthy. "Interpretable Goal-based Prediction and Planning for Autonomous Driving", *International Conference on Robotics and Automation (ICRA)*, 2021;
- C. Brewitt, **B. Gyevnar**, S. Garcin., S.V. Albrecht. "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.

Journal

- **B. Gyevnar**, G. Dagan, C. Haley, S. Guo, F. Mollica. "Communicative Efficiency or Iconic Learning: Do communicative and acquisition pressures interact to shape colour-naming systems?", *Entropy*, 24(11), 1542, 2022.

Codebase & Dataset

- **B. Gyevnar**, C. Wang, S.B. Cohen, C.G. Lucas, S.V. Albrecht. "HEADD: Human Explanations for Autonomous Driving Decisions", *University of Edinburgh*, 2024;
- **B. Gyevnar**, C. Brewit, S. Garcin, M. Tamborski, and S.V. Albrecht. Code Repository for Interpretable Goal-based Prediction and Planning (IGP2); *Github*, 2022.

Workshop

- **B. Gyevnar**, N. Ferguson. "Aligning Explainable AI and the Law: The European Perspective", *AAMAS 2023 Workshop on EXplainable and TRAnsparent AI and Multi-Agent Systems (EXTRAAMAS)*, 2023;
- **B. Gyevnar**, C. Wang, C.G. Lucas, S.B. Cohen, S.V. Albrecht. "Causal Explanations for Stochastic Sequential Multi-Agent Decision-Making", *IJCAI 2023 Workshop on Explainable Artificial Intelligence*, 2023;
- **B. Gyevnar**, M. Tamborski, C. Wang, C.G. Lucas, S.B. Cohen, S.V. Albrecht. "A Human-Centric Method for Generating Causal Explanations in Natural Language for Autonomous Vehicle Motion Planning", Runner-up for best paper, *IJCAI 2022 Workshop on Artificial Intelligence for Autonomous Driving*, 2022;