Rūta Binkytė

Curriculum vitae

Senior Researcher and Deputy Lead
Efficient and Explainable AI (E&E) group · DFKI · Saarbrücken, Germany
ruta.binkyte@dfki.de rutabinkyte.com LinkedIn

I use behavioural experiments and interpretability methods to study bias and safety in AI, from the way agents influence one another's decisions to how people and AI share credit for their work.

Experience

Current	**Senior Researcher and Deputy Lead, Efficient and Explainable AI (E&E) group** German Research Center for Artificial Intelligence (DFKI), Saarbrücken. Simon Ostermann is Group Lead. The group is part of the Multilinguality and Language Technology department.
2024–2026	**Postdoctoral Researcher, Machine Learning Ethics** CISPA Helmholtz Center for Information Security, Saarbrücken. Research on fairness, safety, and multi-agent interaction with Mario Fritz.
2022–2023	**Associate Professor in AI Ethics** Aivancity, Paris-Cachan. Co-designed and taught a graduate course on fair machine learning.
2020	**Research Engineer** Inria Saclay. Worked on causal discovery and privacy-preserving machine learning.
2018–2019	**Data Scientist** Sciences Po médialab, Paris. Studied digital art history using data from French museums.

Education

2020–2023	**PhD in Computer Science** Institut Polytechnique de Paris, École Polytechnique and Inria Saclay. Research on fairness, causality, and privacy, supervised by Catuscia Palamidessi.
2017–2020	**MSc with Distinction, Data Science, Technology and Innovation** University of Edinburgh.
2006–2008	**MA in History** Vilnius University.
2002–2006	**BA in Sociocultural History and Anthropology** Vilnius University.

Research grants

2026	**Coefficient Giving grant for AI Deception Dynamics** Principal investigator. We study deception in the behaviour and internal representations of language-model agents.
2025	**DFG International Collaboration grant** Principal investigator. With CSIRO's Data61 and Macquarie University in Australia, I study how to attribute contributions in human–AI co-creation.

Awards and honours

2025 **AI Publication of the Year nomination**
The Artificial Intelligence Association of Lithuania nominated my paper "Interactional Fairness in LLM Multi-Agent Systems: An Evaluation Framework".

2025 **100 Brilliant Women in AI Ethics**
Named to the annual list by Women in AI Ethics.

2020–2021 **Google Women Techmakers Scholar**
Scholarship for academic achievement and leadership in expanding women's participation in computing.

2020 **MSc with Distinction, University of Edinburgh**
Awarded for the MSc in Data Science, Technology and Innovation.

Teaching

July 2027 **Summer school, Durham University**
Invited to teach at the summer school in Durham.

Aug 2026 **Bias and Safety in AI**
Summer school lecture at the Probabilistic AI School in Vilnius. On 7 August I taught the final session on bias, fairness and safety in AI, with a hands-on notebook.

2022–2023 **Fair Machine Learning, Aivancity**
Co-designed and taught the graduate course at Aivancity School for Technology, Business & Society, Paris-Cachan.

2021–2022 **Python programming, École Polytechnique**
Teaching assistant for the introductory programming course in Paris.

School outreach

11 Oct 2025 **Kids Vibe Code Hackathon, I2SC**
Mentored children using AI to build their own coding projects at the Interdisciplinary Institute for Societal Computing, Saarland University.

2021–2022 **Data Science Mentor, Teachers Lead Tech**
Mentored primary school pupils through an award-winning data science programme in Lithuanian schools.

2018 **Big Small Screens**
Wrote learning materials on media and algorithms for secondary school pupils with Meno Avilys and the British Council in Lithuania.

Academic and community service

2027 **Programme Chair, European Conference on Algorithmic Fairness**
ECAF takes place in Vilnius on 25–27 August 2027.

2025 **Elected Postdoc Representative, CISPA**
Represented postdoctoral researchers at the Helmholtz Center for Information Security.

2025 **Co-founder, Women in AI Lithuania**
Helped establish the Lithuanian chapter of Women in AI and coordinated its academic partnerships.

2024 **Diversity and Inclusion Chair, ECML PKDD**
Served as the conference's Diversity and Inclusion Chair in Vilnius.

2024 **Postdoc Perspectives, Helmholtz Association Postdoc Network**
Member of the steering committee and working group.

2024 **Advisory Board Member, Savannah Digital Research Institute**
Advised on interdisciplinary research and research funding.

2021–2024 **Board Member, Lithuanian Professionals in Paris**
Served on the board of the professional network.

Event Organization

7 May 2026 **Towards Integrated Understanding of Trustworthy AI**
Co-organised the ELSA General Assembly workshop with Joonas Jälkö on how fairness, privacy, robustness and explainability interact. Keynote speakers included Catuscia Palamidessi, Plamen Angelov and Antoine Gautier.

18 Jun 2025 **Trade-offs in Trustworthy AI**
Organised and moderated an ELSA General Assembly panel in Brussels on how improving one property of an AI system can come at the expense of another. Panellists included Antti Honkela, Plamen Angelov, Battista Biggio and Antoine Gautier.

16–21 Mar 2025 **ELSA Workshop on Privacy-Preserving Machine Learning**
Co-organised the Bertinoro workshop on privacy-preserving machine learning, with a focus on federated learning and the trade-offs between privacy, accuracy and efficiency.

Sept 2024 **Diversity and Inclusion Panel, ECML PKDD**
Organised and moderated a panel in Vilnius with Karolina Griciūnė, Ana Peleteiro Ramallo and Gintare Karolina Dziugaite on barriers to participation in AI research and how research communities can address them.

11 Sept 2024 **Women's Networking Night and Food Art Performance, ECML PKDD**
Organised an evening with food artist Giedrė Selenytė at the National Gallery of Art in Vilnius. Erika Kuročkina, then Lithuania's Vice-Minister of the Economy and Innovation, gave the opening remarks.

23–24 Nov 2023 **Second Comète Workshop on Ethical AI**
Co-organised the Inria workshop in Palaiseau on how fairness, privacy, explainability and robustness can come into conflict.

Sept 2022 **Comète Workshop on Ethical AI**
Co-organised the first edition with the Comète team at Inria Saclay. The programme included talks on bias in algorithms and fair ranking.

2021 **Workshop on AI justice, equity and algorithmic discrimination**
Co-organised an interdisciplinary workshop bringing together researchers in computer science, law and social science in Paris.

Peer review and programme committees

2026 **AIES**
Reviewer for the AAAI/ACM Conference on AI, Ethics, and Society.

2026 **IASEAI**
Reviewer for the International Association for Safe and Ethical AI conference.

2025 **EU Digital Technologies and Policies (EUDTP)**
Programme committee member.

2024 **ACM FAccT**
Reviewer for the ACM Conference on Fairness, Accountability, and Transparency.

2023 **SDS and the European Workshop on Algorithmic Fairness**
Programme committee member for SDS and EWAF.

Invited talks and panels

Summer 2026 **Panel discussion at Vinted**
Vinted · Invited panellist

Mar 2026	**Rethinking Fairness for LLMs and Agentic AI** Lovelace–Turing Club · University of Basel, Switzerland · Invited seminar
Mar 2026	**From Prediction to Interaction: Rethinking Fairness for LLMs and Agentic AI** The Future of Trustworthy AI: Frugality, Fairness & Trust · ELIAS, ELLIOT and ENFIELD, Paris · Keynote
Jan 2026	**From Humans to Machines and Back: Fairness, Causality, and the Role of Social Science** Interdisciplinary Institute for Societal Computing · Invited lecture
Nov 2025	**Interactional Fairness in LLM Multi-Agent Systems** Keio University, Tokyo · Invited workshop talk
Nov 2025	**Franco-Japanese Cybersecurity Workshop** Tokyo, Japan · Invited workshop talk
Oct 2025	**Stereotypes, Fairness, and Oppression in LLM-based Social Interactions** AIES 2025 · IE University, Madrid · Invited panellist
Aug 2025	**A Causal Perspective on Balancing Competing Goals in Trustworthy Machine Learning** CAUSALI-T-AI annual workshop · Institut Henri Poincaré, Paris · Invited workshop talk
Jun 2025	**Women Shaping AI in Lithuania** Women in AI Lithuania chapter launch · Vinted, Vilnius · Invited talk
May 2025	**A Causal Perspective on Balancing Competing Goals in Trustworthy Machine Learning** Saarland Informatics Campus · Guest lecture
24 Mar 2025	**Causality in Trustworthy AI** France–Japan workshop on fairness in machine learning · Keio University, Mita Campus, Tokyo · Invited workshop talk
Nov 2024	**AI in Business. Present and Future** DI versle. Dabartis ir ateitis · Online · Invited talk
Nov 2023	**AIvolution: Transforming the European Union's Economy and Society** European Parliament, Brussels · Invited speaker
Jun 2023	**Identifying and mitigating bias in machine learning** Women in Machine Learning and Data Science · Paris · Invited talk
Nov 2022	**MALOTEC seminar** LORIA · Université de Lorraine, Nancy · Invited seminar
Oct 2022	**Causality for Fairness** 7th Franco-Japanese Cybersecurity Workshop · Keio University, Tokyo · Invited workshop talk

Selected publications

Interactional Fairness in LLM Multi-Agent Systems: An Evaluation Framework
Ruta Binkyte. AIES · Oral presentation, 2025.

Position: Trustworthy AI Suffers from Invariance Conflicts and Causality is The Solution
Ruta Binkyte, Ivaxi Sheth, Zhijing Jin, Mohammad Havaei, Bernhard Schölkopf, Mario Fritz. ICML, 2026.

Inspectable AI for Science: A Research Object Approach to Generative AI Governance
Ruta Binkyte, Sharif Abuaddba, Chamikara Mahawaga, Ming Ding, Natasha Fernandes, Mario Fritz. MetaCRiSP · IEEE Security and Privacy Workshops, 2026.

Justice in Judgment: Unveiling (Hidden) Bias in LLM-assisted Peer Reviews
Sai Suresh Macharla Vasu, Ivaxi Sheth, Hui-Po Wang, Ruta Binkyte, Mario Fritz. ACL Findings, 2026.

Publication archive and research links