GERMANS SAVCISENS, Ph.D.

[Ghe-r-man Saf-chi-shen]
Postdoctoral Research Associate
Boston, US | germans@savcisens.com | LinkedIn | Personal Page

PROFILE

I am a postdoctoral researcher studying machine learning methods for auditing, explaining, and improving the reliability of AI systems, with particular focus on how large language models form beliefs. My research investigates how human and AI beliefs influence one another, and how this process shapes collective knowledge and decision-making. I am especially interested in epistemic risks and opportunities of large language models: how they encode uncertainty, represent contested knowledge, and affect human belief formation at scale. I am motivated to pursue an academic research career focused on trustworthy AI and human-centered machine learning.

The areas of interest: Interpretable machine learning, belief and opinion dynamics, human–AI interaction, network science, behavioral and social trajectory modelling.

EDUCATION

Technical University of Denmark | Copenhagen, Denmark **2020 – 2024**
PhD in Applied Mathematics and Computer Science

- Dissertation: "Life Trajectories as Symbolic Language: Exploring Human Behaviour with Language Models"
- Committee: Søren Hauberg (chair), Mads Nielsen, and Arkadiusz Stopczynski.

Aalto University | Helsinki, Finland **2021**
Visiting Student for Digital Ethics and Statistical Natural Language Processing

Technical University of Denmark | Copenhagen, Denmark **2018 – 2020**
MSc in Human-Centred Artificial Intelligence

- Thesis (Collaboration with Novo Nordisk): "Analyzing Health Data Records Using Neural Networks"
- Thesis Advisors: Sune Lehmann (DTU) and Adam Lenart (Novo Nordisk)

Aalborg University | Copenhagen, Denmark **2013 – 2016**
BSc in Medialogy (Human-Computer Interaction)

RESEARCH EXPERIENCE

Postdoctoral Research Associate | Northeastern University **May 2024 — now**
Advisor: Tina Eliassi-Rad

- Developed statistical auditing frameworks to study the behavior of large language models, including belief representation, uncertainty, and bias.
- Designed multiclass probing methodologies for analyzing LLM beliefs and implicit knowledge.
- Supervised graduate researchers in interpretable machine learning and social data modeling.
- Submitted three manuscripts for peer review at leading venues.
- Presented research at NeurIPSMechanistic Interpretability Workshop, IC2S2, and NEMI Workshop.

Research Assistant | Technical University of Denmark & University of Copenhagen **Dec 2023 — Apr 2024**

- Conducted research on transformer-based models
- Supervised projects on modeling life trajectories and smartphone use.

PhD Student | Technical University of Denmark & University of Copenhagen **Sep 2020 — Dec 2023**
Advisors: Sune Lehmann and Lars Kai Hansen.

- Led population-scale modeling of nationwide socioeconomic and health trajectories using transformer-based models.
- Developed and evaluated life-outcome prediction models with an emphasis on interpretability and uncertainty analysis.
- Contributed to academic committees and departmental activities, including research coordination.
- Actively disseminated research through invited talks at MIT, ODISSEI, University of Oxford, NordicAI, and SODAS.
- Taught and developed course materials for a graduate-level course in Social Network Science.

Guest Researcher | Denmark Statistics (Data Science Lab) **Sep 2020 — Dec 2023**

- Collaborated with economists and social scientists to analyze demographic and socioeconomic patterns.
- Led the preprocessing and modeling of large-scale longitudinal datasets.

Visiting PhD Student | Network Science Institute (Northeastern University) **Sep 2022 — Feb 2023**

- Analyzed ethical and societal risks of deep learning applied to socioeconomic data.
- Investigated methods for enhancing transparency, accountability, and fairness in algorithmic systems

Global Data Science Researcher | Novo Nordisk **Sep 2019 — Apr 2020**

- Conducted predictive modeling research on medical adherence using large-scale Electronic Health Records.
- Developed hierarchical attention-based neural models for health outcome prediction.
- Led preprocessing and integration of behavioral and clinical data sources.

PUBLICATIONS

Savcisens, G., & Eliassi-Rad, T.
Trilemma of Truth in Large Language Models. **Mechanistic Interpretability Workshop at NeurIPS**, 2025.

Dies, S., Maynard, C., **Savcisens, G.**, & Eliassi-Rad, T.
Representational Stability of Truth in Large Language Models. arXiv preprint**, arXiv:2511.19166 [cs.LG]**, 2025.

Shafi, Z., **Savcisens, G.**, & Eliassi-Rad, T.
REGE: A Method for Incorporating Uncertainty in Graph Embeddings. In Proceedings of the **SIAM SDM**, pp. 376–385, 2025

Savcisens, G.
Large Language Models Act as If They Are Part of a Group. **Nature Computational Science** (News & Views), 5(1), 9–10, 2025.

Savcisens, G., Eliassi-Rad, T., Hansen, L. K., Mortensen, L. H., Lilleholt, L., Rogers, A., Zettler, I., & Lehmann, S.
Using Sequences of Life-Events to Predict Human Lives. **Nature Computational Science**, 4(1), 43–56, 2024.

Denove, E., Michelet, E., **Savcisens, G.**, & Fernández Fernández, E.
An Industrial West? A Mixed-Methods Analysis of Newspaper Discourses about Technology over One Hundred and Ten Years.
Journal of Data Mining & Digital Humanities, 2024.

Fernández Fernández, E., & **Savcisens, G**.
A Sustainable West? Analyzing Clusters of Public Opinion in Sustainability Western Discourses in Multilingual Newspapers.
Proceedings of the **Digital Humanities in the Nordic and Baltic Countries 2023**, pp. 165–187, 2023.

TEACHING EXPERIENCE

Guest Lecturer | Northeastern University
Machine Learning with Graphs — **Autumn 2024 & 2025**

Principal Lecturer & Examiner | IT University of Copenhagen
Algorithmic Fairness, Accountability, and Ethics — **Spring 2023**

Assistant Lecturer | IT University of Copenhagen
Algorithmic Fairness, Accountability, and Ethics — **Spring 2022**

Teaching Assistant | Technical University of Denmark
Advanced Machine Learning — **Spring 2021**
Social Data Analysis and Visualizations — **Spring 2020 & 2021**
Social Networks and Interactions — **Autumn 2019 & 2020**

Supervision
MSc Student Projects: (1) Rule Detection in Unstructured Documents, (2) Forecasting Smartphone App Usage, (3) Automating the Identification of Compliance Rules from Investment Mandates, (4) Predictive Modelling of Mobile App Usage

SELECTED TALKS & PRESENTATIONS

Trilemma of Truth in LLMs | Mechanistic Interpretation Workshop at NeurIPS. San Diego, USA (Poster) — **Dec 7, 2025**
Foundation Models for Registry Data |Complexity Science Hub. Vienna, Austria (Invited Speaker) — **Nov 17, 2025**
Trilemma of Truth in LLMs | Gore Laboratory, MIT. Boston, USA (Invited Speaker) — **Dec 7, 2025**
From Life-Courses to Representations | SWECOV Workshop. Stockholm, Sweden (Keynote Speaker) — **Sep 1, 2025**
Improving Probes that Track Veracity in LLMs | IC2S2. Norrköping, Sweden (Poster) — **Jul 25, 2025**
Life Trajectories in High-Dimensional Spaces | University of Helsinki. Helsinki, Finland (Invited Speaker) — **Feb 4, 2025**
Life Trajectories in High-Dimensional Spaces | Max Planck Institute. Rostok, Germany (Invited Speaker) — **Jan 29, 2025**
Life Trajectories in High-Dimensional Spaces | Chinese University of Hong Kong (Invited Speaker) — **Nov 14, 2024**
Using Life-sequences to Predict Human Lives | ODISSEI Lecture. Amsterdam, Netherlands (Invited Speaker) — **Sep 23, 2023**

GRANTS

Travel Grants: Otto Mønsted Fond, STIBO Foundation, William Demant Fonden (Denmark) — **2022**
Scholarship: "Stability" Scholarship from UPB A/S (Liepaja, Latvia) — **2013-2016**

PROFESSIONAL SERVICES

Member, Research and Continuous Learning Steering Committee | Technical University of Denmark (Compute) — **2022-2023**
Mentor, Student Counseling (Studenterrådgivningen) | Technical University of Denmark — **2020-2021**

Manuscript Reviews

Journals

- Nature Computational Science
- npj Complexity
- Science Advances
- Scientific Reports
- EPJ Data Science

Conferences and Workshops

- Conference on Complex Networks and Their Applications (2025)
- Digital Humanities in the Nordic and Baltic Countries (2025)
- EMNLP (2024)
- IC2S2 (2025)
- NeurIPS (2023, 2024, 2025)
- New England Interpretability Workshop (2025)