

NICOLÒ PAGAN, Ph.D. ETH

AI Research Scientist

AI Research Scientist with a decade of experience in AI research, and three years in production software engineering. Mathematical background in game theory, dynamical systems, and multi-agent strategic interaction, applied to auditing economically consequential LLM decisions, from labor markets to social media recommender systems. Developed inverse optimization frameworks for agent preference inference; established the first formal taxonomy of feedback loops and biases in ML systems; designed scalable LLM evaluation and auditing frameworks, fine-tuning workflows, and RAG architectures across multiple models and platforms. Won a gold medal in the 2024 Kaggle ARC Challenge, one of AI's hardest benchmarks for abstract reasoning and few-shot generalization. Secured CHF 662,000 in competitive SNSF funding as Principal Investigator.

Research Focus: AI Evaluation — Automated Decision-Making — Multi-Agent Systems — Algorithmic Fidelity — Algorithmic Fairness — Recommender Systems — Responsible AI



20.12.1988, Swiss, Italian

LinkedIn

Google Scholar

GitHub

nicolo.pagan@outlook.com

nicolo-pagan.github.io

WORK EXPERIENCE

Senior Research Scientist

Sep 2026 - present

Post-doctoral Researcher

Jan 2025 - Aug 2026

University of Zurich — Social Computing Group — Zurich (CH)

LLM Decision-Making Evaluation: Building evaluation frameworks for multi-agent LLM systems deployed in labor-market and social media recommender pipelines

Algorithmic Fidelity: Designed multi-platform benchmarking suite combining BERT classification, cosine similarity, and NLP feature extraction to classify human- vs LLM-generated social media content at scale (*PNAS*, 2026)

Responsible AI: Formalizing bias risks in AI-assisted decision-making, in particular in the context of social media recommendation systems

Research Leadership: Supervising 4 PhD students and more than 20 Master's students across several projects

AI Research Scientist

Mar 2024 - Sep 2024

Giotto.AI — Lausanne (CH), remote

ARC Challenge (AGI) Evaluation Pipeline: Built advanced classifiers for ARC task types to reduce the solution search space; designed a test-time-training and candidate-filtering pipeline for abstract reasoning; team won gold medal in 2024 Kaggle ARC Prize

NCCR Automation Fellow

Jan 2022 - Mar 2024

Swiss National Science Foundation (SNSF) — CH

AI Evaluation: Developed first systematic taxonomy of feedback loop types in ML systems and their formal relation to algorithmic biases, with implications across labor-market decision-making, content moderation, and credit scoring

Fairness Auditing: Led 216'000 CHF SNSF project on algorithmic fairness in recommender systems and resource allocation problems (*Nature Communications*, 2021)

Algorithmic Fidelity: Pioneered LLM-based approaches for modeling and auditing human behavior in social systems

Doctoral Researcher

Jan 2016 - Apr 2021

ETH Zurich — Automatic Control Lab — Zurich (CH)

Inverse Optimization: Developed game-theoretic framework to infer agent objectives from large-scale behavioral observations; proved convexity guarantees enabling tractable recovery at network scale (*Nature Communications*, 2019)

Multi-agent Systems: Mathematical modeling of strategic interaction dynamics in directed social networks

Software Engineer and Research Consultant

Sep 2012 - Dec 2015

Ascomp AG — Zurich (CH)

R&D: Development of production-grade Computational Fluid-dynamics (CFD) solutions

DevOps: Full-stack software development, HPC pipelines, unit testing, CI/CD

KEY ACHIEVEMENTS

ARC AGI Challenge Gold Medal (2024)

SNSF Grants: 662'000 CHF in competitive funding for Responsible AI research

Nature Communications: Top 50 most-read Physics article (2019)

EDUCATION

Ph.D., ETH Zurich

Jan 2016 - Apr 2021

Automatic Control Laboratory

M.Sc., EPF Lausanne

Sep 2011 - Apr 2013

Computational Science and Engineering

B.Sc., Politecnico di Torino

Sep 2007 - Oct 2010

Mathematical Engineering, cum Laude

SKILLS

Languages & Frameworks: Python — C/C++ — Py-Torch — Hugging Face — Transformers — scikit-learn

LLM Engineering: Fine-tuning (LoRA, PEFT) — Evaluation pipelines — Automated decision-making audits — Agentic systems — RAG

Software Engineering: HPC — CI/CD — Git — Docker

Methods: Optimization — Control Theory — Game Theory — Statistical Inference — NLP

SELECTED PUBLICATIONS

Pagan N., Törnberg P., Bail C.A., and Barrie C., *AI-Generated Language Remains Systematically Distinguishable from Human Communication*, *PNAS* (2026)

Pagan N., Barrie C., Bail C.A., and Törnberg P., *LLMs as Content Curators: A Large-Scale Audit of Recommendation Bias Across Providers and Platforms*, *COLM* (2026)

Bouleimen A., Pagan N., and Hannák A., *Illusory Truth or Mere Exposure? Model-Dependent Repetition Effects in LLM-Based Social Media Simulations*, *COLM* (2026)

Pagan N., Baumann J., Elokda E., De Pasquale G., Bolognani S., and Hannák A., *A Classification of Feedback Loops and Their Relation to Biases in Automated Decision-Making Systems*, *EAAMO* (2023)

Pagan N., Mei W., Li C., and Dörfler F., *A meritocratic network formation model for the rise of social media influencers*, *Nature Communications* (2021)

Pagan N. and Dörfler F., *Game theoretical inference of human behavior in social networks*, *Nature Communications* (2019)

FULL LIST OF PUBLICATIONS

Journal Articles

- [J6] **Pagan, N.**, Törnberg, P., Bail, C.A., and Barrie, C. "AI-Generated Language Remains Systematically Distinguishable from Human Communication." *PNAS* 123, e2531748123 (2026). doi: [10.1073/pnas.2531748123](https://doi.org/10.1073/pnas.2531748123).
- [J5] Bouleimen, A., De Marzo, G., Kim, T., **Pagan, N.**, Metzler, H., Giordano, S., Hannák, A., and Garcia, D. "The Collective Turing Test: Large Language Models Can Generate Realistic Multi-User Discussions." *Scientific Reports* (2026). doi: [10.1038/s41598-026-62286-9](https://doi.org/10.1038/s41598-026-62286-9).
- [J4] Manivannan, A. et al., **Pagan, N.**, et al. "Generative AI for climate governance and acceptability-constrained policy design." *npj Climate Action* 5(1), 37 (2026). doi: [10.1038/s44168-026-00362-6](https://doi.org/10.1038/s44168-026-00362-6).
- [J3] Ionescu, S., Hannák, A., and **Pagan, N.** "The role of luck in the success of social media influencers." *Applied Network Science* 8(1), 46 (2023). doi: [10.1007/s41109-023-00573-4](https://doi.org/10.1007/s41109-023-00573-4).
- [J2] **Pagan, N.**, Mei, W., Li, C., and Dörfler, F. "A meritocratic network formation model for the rise of social media influencers." *Nature Communications* 12, 6865 (2021). doi: [10.1038/s41467-021-27089-8](https://doi.org/10.1038/s41467-021-27089-8).
- [J1] **Pagan, N.** and Dörfler, F. "Game theoretical inference of human behavior in social networks." *Nature Communications* 10, 5507 (2019). doi: [10.1038/s41467-019-13148-8](https://doi.org/10.1038/s41467-019-13148-8).

Journal Articles — Under Review

- [J7] Törnberg, P., Heseltine, M., **Pagan, N.**, Bail, C.A., Schimmel, M., and Barrie, C. "Political Sorting Can Drive AI Models Apart Through User Feedback." *Under review, PNAS* (2026).

Journal Articles — In Preparation

- [J10] **Pagan, N.**, Ebling, M.A., Ionescu, S., and Hannák, A. "AI evaluators favor job applications written by certain AI models, independent of candidate quality." *In preparation for Nature Machine Intelligence* (2026).
- [J9] Nae, T. and **Pagan, N.** "Literature Review on Social Machine Behavior." *In preparation for Artificial Intelligence Review* (2026).
- [J8] **Pagan, N.**, Törnberg, P., Bail, C.A., and Barrie, C. "LLM Recommenders Produce Demographic Inequality Through Content Preferences." *In preparation for Science Advances* (2026).

Conference Articles

- [C12] Jaoua, S., **Pagan, N.**, and Ionescu, S. "Ordered Pairwise Comparison: Structured Exploration for Improving Provider Fairness in Recommender Systems." *Beyond Workshop @ RecSys* (2026).
- [C11] **Pagan, N.**, Barrie, C., Bail, C.A., and Törnberg, P. "LLMs as Content Curators: A Large-Scale Audit of Recommendation Bias Across Providers and Platforms." *COLM* (2026).
- [C10] Bouleimen, A., **Pagan, N.**, and Hannák, A. "Illusory Truth or Mere Exposure? Model-Dependent Repetition Effects in LLM-Based Social Media Simulations." *COLM* (2026).
- [C9] Urman, A., Lichtenegger, E., Jaoua, S., Bouleimen, A., Forsberg, R., Hertweck, C., Ionescu, S., **Pagan, N.**, Hannák, A., and Baumann, J. "Prompt Revision as a Source of Cultural Bias in Text-to-Image Systems." *EMNLP* (2026).
- [C8] **Pagan, N.**, Philippou, A., and De Pasquale, G. "Learning to Control Misinformation: a Closed-loop Approach for Misinformation Mitigation over Social Networks." *L4DC* (2026). doi: [10.48550/arXiv.2511.12393](https://doi.org/10.48550/arXiv.2511.12393).
- [C7] **Pagan, N.**, Törnberg, P., Bail, C.A., Hannák, A., and Barrie, C. "Can LLMs Imitate Social Media Dialogue? Techniques for calibration and BERT-based Turing-Test" *COLM* (2025).
- [C6] **Pagan, N.***, Baumann, J.*, Elokda, E., De Pasquale, G., Bolognani, S., and Hannák, A. "A Classification of Feedback Loops and Their Relation to Biases in Automated Decision-Making Systems." *EAAMO* (2023). doi: [10.1145/3617694.3623227](https://doi.org/10.1145/3617694.3623227).
- [C5] Ionescu, S., Hannák, A., and **Pagan, N.** "Group fairness for content creators: the role of human and algorithmic biases under popularity-based recommendations." *RecSys* (2023). doi: [10.1145/3604915.3608841](https://doi.org/10.1145/3604915.3608841).
- [C4] Lanzetti, N., Dörfler, F., and **Pagan, N.** "The Impact of Recommendation Systems on Opinion Dynamics: Microscopic Versus Macroscopic Effects." *IEEE CDC* (2023). doi: [10.1109/CDC49753.2023.10383957](https://doi.org/10.1109/CDC49753.2023.10383957).
- [C3] Maddux, A.M., **Pagan, N.**, Belgioioso, G., and Dörfler, F. "Data-driven behaviour estimation in parametric games." *IFAC-PapersOnLine* 56(2) (2023). doi: [10.1016/j.ifacol.2023.10.220](https://doi.org/10.1016/j.ifacol.2023.10.220).
- [C2] Bouleimen, A., **Pagan, N.**, Cresci, S., Urman, A., and Giordano, S. "Dynamics of toxic behavior in the Covid-19 vaccination debate." *Complex Networks* (2023). doi: [10.1007/978-3-031-53503-1_26](https://doi.org/10.1007/978-3-031-53503-1_26).
- [C1] Ionescu, S., **Pagan, N.**, and Hannák, A. "Individual fairness for social media influencers." *Complex Networks* (2022). doi: [10.1007/978-3-031-21127-0_14](https://doi.org/10.1007/978-3-031-21127-0_14).

Conference Articles — Under Review

- [C14] **Pagan, N.** "Risk-Seeking Strategic Robustness in Games: A Substitutes-Complements Rule, and Where It Breaks." *Under review, ICLR* (2027).
- [C13] Joos, T.C., Bouleimen, A., Hannák, A., and **Pagan, N.** "Perceived Trust and Transparency in AI-Altered vs. Human-Written LinkedIn Posts." *Under review, ICWSM* (2027).

GRANTS

- SNSF Fellowship, "A Cross-National Study to Promote Democratic Dialogue on Social Media Using Generative Artificial Intelligence," **CHF 230,552** (2025–2027). Role: **Lead Researcher**.
- SNSF Fellowship (NCCR Automation, Phase II), "Definitions of Fairness in Control," **CHF 216,000** (2025–2027). Role: **Funding acquisition**.
- SNSF Fellowship (NCCR Automation, Phase I), "Ethical Auditing of Control Systems Methods," **CHF 216,000** (2022–2023). Role: **Funding acquisition, Lead Researcher**.