Agathe Balayn

Curriculum Vitae

Residence: New York City, United States of America
☎ *+33.6.99.55.72.23*
✉ *agatheb.research@gmail.com*
🌐 *https://agathe-balayn.github.io/*

Research Mission

The grand goal of my work is to better understand and control the harmful impacts of Machine Learning (ML) systems. Particularly, my research focuses on characterizing theories and practices for developing and evaluating ML systems with regard to safety issues and societal impacts, on critically reflecting on the assumptions of current ML research lenses, and on proposing supporting methods, tools, and policies for ML practitioners and researchers.

Education and Academic Training

04/2019 - 04/2023 **PhD researcher in Computer Science**, *Human-Computer Interaction*, Delft University of Technology
- Topic: Supporting ML practitioners in developing safe and non-harmful models, via a mixed-method approach (empirical qualitative studies; literature reviews; workflow design; quantitative user-studies).
- PhD thesis *"On developers' practices for hazard diagnosis in machine learning systems."* Graduated cum laude. *(Advisors: Alessandro Bozzon, Geert-Jan Houben)*

09/2016 - 09/2018 **MSc in Computer Science**, *Data Science and Technology track*, Delft University of Technology
- GPA: 8.72/10. Focus on artificial intelligence, machine and deep learning, human-computer interaction
- Master thesis (9/10) entitled: *"On the fairness of crowdsourced training data and ML models for the prediction of subjective properties. The case of sentence toxicity." Graduated as an Honour student.*

11/2017 - 09/2018 **Graduate Intern at the IBM Center for Advanced Studies (Benelux)**, *the Netherlands*
Study of biases and fairness in crowdsourced data and ML models for the prediction of subjective properties, with the use-case of sentence toxicity prediction. *(Manager: Zoltan Szlavik)*

08/2017 - 10/2017 **Research Intern at the Honda Research Institute (HRI-JP)**, *Wako, Japan*
Creation of encoding schemes for sign language annotations. Design, implementation, and evaluation of deep learning models for sign language synthesis and recognition from motion capture data. *(Advisor: Heike Brock)*

09/2014 - 09/2018 **MSc in Systems & Control**, *ENSTA ParisTech Institut Polytechnique de Paris*, France
- Strong component of Control, Informatics, and Signal. (Program leading to a "Diplôme d'ingénieur")
- GPA: 4.0/4.0. Graduated first year 2nd of the class out of 144 students.

05/2016 - 07/2016 **Research Intern at the Research Institute for Cognition and Robotics (CoR-Lab)**, *Germany*
Design, implementation, and evaluation of an active-compliance control mode using ELM neural networks for an industrial lightweight robotic arm (Universal Robots UR5). *(Advisor: Jeffrey Queisser)*

Professional Experiences

10/2024 - now **Postdoctoral researcher**, *Fairness, Accountability, Transparency, Explainability (FATE) lab*, Microsoft Research NYC
Empirical studies of evaluation approaches for LLM systems. Conceptual studies of the gap between policies and responsible AI practices or transparent reporting of LLM usages. *(Manager: Hanna Wallach)*

03/2024 - 10/2024 **Postdoctoral researcher**, *Technology, Policy & Management*, Delft University of Technology
Empirical, qualitative study of the impact of ML in terms of political economy questions. In-depth investigation of the computer vision supply chain for pig farming. *(Advisor: Seda Guerses)*

09/2023 - 10/2024 **Guest and visiting researcher**, *University of Trento (Italy) & Delft University of Technology*
Collaborations with professors and PhD students on various research projects dealing with: studying the fairness perceptions of ML's decision-subjects; studying ML researchers' data practices related to ML fairness and robustness; and surveying the literature on ML robustness. *(Advisors: Fabio Casati, Jie Yang)*

08/2023 - 05/2024 **Visiting researcher**, *Machine Learning Trust & Governance team*, ServiceNow
Large-scale, empirical, qualitative, investigation of the ethical concerns and challenges (especially with regard to ML knowledge and trust dynamics) of the stakeholders in the ML supply chain (interviews with 72 participants). *(Advisor: Fabio Casati)*

04/2023 - 08/2023 **Postdoctoral researcher**, *Delft University of Technology (the Netherlands)*

- Activities conducted 40% in the Computer Science Faculty (EEMCS) and 60% in the Technology, Policy, Management Faculty (TPM). *(Advisors: Jie Yang, Seda Guerses)*
- Topic: Critically looking at ML practitioners' work using infrastructural and political economy lenses.

01/2021 - 08/2021 **Consultant for the non-governmental organisation EDRi**, *(European Digital Rights Organisation)*

Writing and presentation of a policy report about Computer Science notions of ML fairness, their conceptual and practical limitations, and the implications thereof for policy documents and regulations.

09/2018 - 03/2019 **Researcher at the IBM Center for Advanced Studies and at the TU Delft**, *the Netherlands*

Investigation of the fairness of ML pipelines for the inference of subjective labels; survey on hate speech detection adopting a critical, psychology, lens.

Awards, Honors, and Recognitions

Best paper awards

- Best paper awards: at the Conference on Human Factors in Computing Systems (CHI'23); at the Conference on AI, Ethics, and Society (AIES'23); at the AAAI Conference on Human Computation and Crowdsourcing (HCOMP'22)
- Honorable mention: at the Conference on Human Factors in Computing Systems (CHI'25); at the Web Conference (WWW'22)
- Best demo award: at the AAAI Conference on Human Computation and Crowdsourcing (HCOMP'21)

Honors

- Rising Talent prize from the UNESCO and Fondation L'Oreal *For Women in Science* initiative (honorable mention given to 2 out of 74 applicants, sole prize attributed for the STEM field) (2023)
- Conference award given by the Renmin University of China during the International Conference on Frontier and Innovation for Young Scholars (2023)
- Completion of the Honors Programme of the Delft University of Technology (additional 20 ECTS)
- Valedictorian for all the three high-school years; salutatorian during the MSc, PhD cum laude

Scholarship, *Obtained the Erasmus Plus scholarship based on merit for a research internship (2016)*

Professional Services

Reviewer, *CHI'21-25, CSCW'21-24, FAccT'24-25, IUI'20-21, HCOMP'20-22, WWW'20-22, AAAI'22, NeurIPS'22, CIKM'21-23, NAACL'21, UMUAI'21, IEEE Access'21, ChineseCHI'20, HyperText'20-22, ROMAN'20-21, reviews for various conference workshops*

Chair, *Program chair (EWAF'24); Area chair (EWAF'25); Workshop organizer (CSCW'24)*

Student volunteer, *International Conference on Management of Data (SIGMOD 2019)*

Presentations at local events

- At schools: *Keynote speaker at the NoBias Summer School (Pisa University, 2023); invited speaker at the spring school "Ethos+Tekhné: a new generation of AI researchers" (Pisa University, 2023)*
- At workshops: *Public Interest AI workshop (Humboldt Institute for Internet and Society, 2022); Lorentz workshop on fairness in automated decision-making systems (Lorentz workshop, 2022); Young Scholar International Conference (Renmin University, Beijing, China, 2023)*
- At local events: *discussion chairing (on the reviewing crisis in HCI) at CHI Netherlands post-CHI event (2023, 2022); ICT.Open (2022); HUMAINT Winter School on Fairness, Accountability and Transparency in Artificial Intelligence (2020); Symposium on Biases in Human Computation and Crowdsourcing (BHCC) (2019); Dutch-Belgian Database Day (2019)*
- PhD consortium: *FAccT PhD consortium (2020)*
- Talks in research groups, e.g., *ServiceNow Trust and Governance team (remote, 2023), Microsoft 2023, iHub Radboud University (Netherlands, 2022), Law School of SciencesPo (France, 2022), FU Berlin (Germany, 2022), Platform for the Ethics and Politics of Technology (PEPT) (Netherlands, 2021)*

Participation to panel discussions

- European Workshop on Algorithmic Fairness (EWAF) (06/2023); Workshop on Algorithmic Injustice (University of Amsterdam, 06/2023); Online panel series on "Taking back control of data in the UK" –Redesigning fairness: concepts, contexts and complexities (Ada Lovelace Institute, 10/2021)

Diversity & Inclusion in STEM research

- Participation to the ACM WomEncourage workshop (2023); Participation in the women in Computer Science reflection group at TU Delft (2022-2024); Participation in the women in STEM discussion group at Microsoft Research (2024-now); Member of the *Slow Reading* group on AI and gender inequality organized by Rotterdam's Royal Academy of Art (bringing a Computer Science perspective to the artist collective)

Public outreach

- Redaction of a report for the non-governmental organisation EDRi to advise on the European Union AI Act
- Comments for various digital newspapers around the report; presentation of the report's insights to law, social science, and computer science research groups and to interdisciplinary workshops
- Comments on news around ML for several digital newspapers, e.g., AlgorithmWatch, NextInpact, Further-Up-HR

PhD cover design and realization

For the following PhD candidates: Agathe Balayn (ML harms), Nirmal Roy (information retrieval), Siddharth Mehrotra (Human-AI interactions), Lijun Lyu (information retrieval), Christos Koutras (data management), Mireia Yurrita (Human-AI interactions)

Teaching and Mentorship

Teaching

- Material designer and teaching assistant for the ML fairness introduction within an inter-faculty ML course (TU Delft, 2022)
- Teacher for introductory lectures on AI ethics at the TU Delft CS faculty (2022, 2o23)
- Teaching assistant for the Crowd Computing course and the Web Information Systems seminar (2019-2023)

Mentorship

- Supervision of nine Bachelor students for their BSc thesis projects; and thirteen Master students for their MSc thesis projects; support of one BSc student for their Honours program; support of four PhD students in various research projects (2019-2024)
- Supervision of a group of five second year Bachelor students for a software engineering project; and four groups of 4 Master students for crowdsourcing+AI projects

Languages

French	**Native speaker.**	Mandarin	**Elementary proficiency.**	Japanese	**Notions.**
English	**Professional proficiency.**	German	**Elementary proficiency.**	Greek	**Notions.**

Technical Skills

Highly proficient	**Python** (TensorFlow, Keras, Scikit-learn, Pandas, Numpy, etc.), **version control (Git), LaTeX, common software suites (Office)**
Proficient	**C++, C, MATLAB, Maple, HTML, CSS, Linux, Bash**
Familiar	**Java, PHP, Javascript**, libraries (OROCOS and Gazebo environment for C++, D3 library)

Extra-Curricular Activities

2021-now	Participation in **art** classes and events, *Painting (watercolor, oil, acrylic), sketching, museum visits*
2017-2019	**Band member** of a traditional Chinese music band at TU Delft, *Flutist*
2015-2016	Member of the **robotics club** of ENSTA ParisTech (ENSTAR), *Treasurer, Arduino programmer*
2015-2016	Member of the organisation team of the **cultural festival** of ENSTA ParisTech (Arts en Scene), *Communication manager (social media manager, promotional illustration organizer)*
2014-2016	Member of the **Board of European Students of Technology** (BEST) at ENSTA ParisTech, *Manager of one international event, establisher of company relations for consulting workshops*
2014-2015	Volunteer for the **NGO** ZUPDeCo, *Tutoring support for middle school students in difficulty*
2009-2012	Volunteer for the **NGO** *Les Enfants du Mekong*, *Fund gathering via volunteering activities*