

Agathe Balayn

Residence: Delft, the Netherlands

+33.6.99.55.72.23

a.m.a.balayn@tudelft.nl

<https://agathe-balayn.github.io/>

Curriculum Vitae

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. Some of my recent work under review deals with: studying practices of ML practitioners concerning algorithmic harms beyond ML unfairness; studying ML researchers' data practices related to ML fairness and robustness; surveying the literature on ML robustness; and proposing a cost-efficient, concept-based ML model diagnosis method.

Education and Academic Experiences

- 04/2023 - **Postdoctoral researcher, Delft University of Technology (the Netherlands)**
- 08/2023
 - Activities conducted 40% in the Computer Science Faculty (EEMCS) and 60% in the Technology, Policy, Management Faculty (TPM).
 - Topic: Critically looking at ML practitioners' work using infrastructural and political economy lenses.
- 04/2019 - **PhD candidate in Computer Science (cum laude), HCI, Delft University of Technology**
- 04/2023
 - 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" (04/10/2023).
- 09/2016 - **MSc in Computer Science, Data Science and Technology track, Delft University of Technology**
- 09/2018
 - 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."
- 09/2014 - **MSc in Systems & Control, ENSTA ParisTech Institut Polytechnique de Paris, France**
- 09/2018
 - 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.

Professional Experiences

- 08/2023 - **Visiting researcher at ServiceNow, (Machine learning trust and governance team)**
- now Empirical, qualitative, investigation of the ethical concerns and challenges of ML stakeholders beyond ML developers, towards the ideation of new governance structures for organizational responsible AI.
- 01/2021 - **Consultant for the non-governmental organisation EDRI, (European Digital Rights Organisation)**
- 08/2021 Writing and presentation of a report about the ML fairness framework in computer science, its conceptual and practical limitations, and the implications of these limitations for policy documents and regulations.
- 09/2018 - **Researcher at the IBM Center for Advanced Studies and at the TU Delft, the Netherlands**
- 03/2019 Investigation of the fairness of ML pipelines for the inference of subjective labels; survey on hate speech detection adopting a critical, psychology, lens.
- 11/2017 - **Graduate Intern at the IBM Center for Advanced Studies (Benelux), the Netherlands**
- 09/2018 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.
- 08/2017 - **Research Intern at the Honda Research Institute (HRI-JP), Wako, Japan**
- 10/2017 Creation of encoding schemes for sign language annotations. Design, implementation, and evaluation of deep learning models for sign language synthesis and recognition based on motion capture data.
- 05/2016 - **Research Intern at the Research Institute for Cognition and Robotics (CoR-Lab), Germany**
- 07/2016 Design, implementation, and evaluation of an active-compliance control mode using ELM neural networks and model-space learners for an industrial lightweight robotic arm (Universal Robots UR5).
- 08/2015 **Summer trainee at the company Hakuba Lion Adventure, Hakuba, Japan**
- Accompanied groups of tourists to outdoor activities (e.g., canyoning, ski lessons). Japanese-speaking team.

Awards, Honors, and Recognitions

Best paper award

- Best student paper award: at the Conference on AI, Ethics, and Society (AIES'23)
- Best paper awards: at the Conference on Human Factors in Computing Systems (CHI'23); at the AAAI Conference on Human Computation and Crowdsourcing 2022 (HCOMP'22)
- Nomination for best paper award: at the Web Conference 2022 (WWW'22)
- Best demo award: at the AAAI Conference on Human Computation and Crowdsourcing 2021 (HCOMP'21)

Honors

- Rising Talent prize from the UNESCO and Fondation L'Oreal *For Women in Science* initiative (honourable mention given to 2 out of 74 applicants, sole prize attributed for the STEM field)
- Conference award given by the Renmin University of China during the international conference on frontier and innovation for young scholars
- 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*

Professional Services

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

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); CHI Netherlands post-CHI event (2022); ICT.Open (2022); 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), 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)

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 machine learning for several digital newspapers, e.g., AlgorithmWatch, NextImpact

Teaching and Mentorship



Teaching

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

Mentorship

- Supervision of nine Bachelor students for their BSc thesis projects; and thirteen Master students for their MSc thesis projects; support of four PhD students in various research projects
- 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



Human-centered Studies, Frameworks, and Tools


- AIES 2023  **Agathe Balayn**, Mireia Yurrita, Jie Yang, and Ujwal Gadiraju. “fairness toolkits, a checkbox culture?” on the factors that fragment developer practices in handling algorithmic harms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, 2023*, (best student paper award)
- CHI 2023 **Agathe Balayn**, Natasa Rikalo, Jie Yang, and Alessandro Bozzon. Faulty or ready? handling failures in deep-learning computer vision models until deployment: A study of practices, challenges, and needs. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 2023*
- CHI 2023  Mireia Yurrita, Tim Draws, **Agathe Balayn**, Dave Murray-Rust, Nava Tintarev, and Alessandro Bozzon. Disentangling fairness perceptions in algorithmic decision-making: the effects of explanations, human oversight, and contestability. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2023, (CHI best paper award)
- CHI 2022 **Agathe Balayn**, Natasa Rikalo, Christoph Lofi, Jie Yang, and Alessandro Bozzon. How can explainability methods be used to support bug identification in computer vision models? In *CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2022
- FAccT 2022 Mireia Yurrita, Dave Murray-Rust, **Agathe Balayn**, and Alessandro Bozzon. Towards a multi-stakeholder value-based assessment framework for algorithmic systems. In *FAccT Conference on Fairness, Accountability, and Transparency, 2022*
- CVPR (WS) 2021 **Agathe Balayn**, Bogdan Kulynych, and Seda Guerses. Exploring data pipelines through the process lens: a reference model for computer vision. *arXiv preprint arXiv:2107.01824 (CVPR 2021 workshop “Beyond Fairness”)*, 2021
-

Systematic Literature Reviews

- FAccT 2023 Luca Nannini, **Agathe Balayn**, and Adam Leon Smith. Explainability in ai policies: A critical review of communications, reports, regulations, and standards in the eu, us, and uk. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1198–1212, 2023
- VLDBJ 2021 **Agathe Balayn**, Christoph Lofi, and Geert-Jan Houben. Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems. *The VLDB Journal*, 30(5):739–768, 2021
- TSC 2021 **Agathe Balayn**, Jie Yang, Zoltan Szlavik, and Alessandro Bozzon. Automatic identification of harmful, aggressive, abusive, and offensive language on the web: A survey of technical biases informed by psychology literature. *ACM Transactions on Social Computing (TSC)*, 4(3):1–56, 2021
- Technical report for EDRi **Agathe Balayn** and Seda Gürses. Beyond debiasing: Regulating ai and its inequalities. *Report for the European Digital Rights organisation (EDRi)*. https://edri.org/wp-content/uploads/2021/09/EDRi_Beyond-Debiasing-Report_Online.pdf, 2021
-

Technical Methods and Systems

- IUI 2023 Shahin Sharifi Noorian, Sihang Qiu, Burcu Sayin, **Agathe Balayn**, Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. Perspective: Leveraging human understanding for identifying and characterizing image atypicality. In *Proceedings of the International Conference on Intelligent User Interfaces, 2023*
- WWW 2022  **Agathe Balayn**, Gaole He, Andrea Hu, Jie Yang, and Ujwal Gadiraju. Ready player one! eliciting diverse knowledge using a configurable game. In *Proceedings of the Web Conference 2022*, pages 1709–1719, 2022, (WWW best paper nomination)
- HCOMP 2022  Gaole He, **Agathe Balayn**, Stefan Buijsman, Jie Yang, and Ujwal Gadiraju. It is like finding a polar bear in the savannah! concept-level ai explanations with analogical inference from commonsense knowledge. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, pages 89–101, 2022, (HCOMP best paper award)

- HCOMP demo 2021 **Agathe Balayn**, Gaole He, Andrea Hu, Jie Yang, and Ujwal Gadiraju. Finditout: A multiplayer gwap for collecting plural knowledge. In *Vol. 9 (2021): Proceedings of the Ninth AAAI Conference on Human Computation and Crowdsourcing*, 2021, (HCOMP best demo award) 
- WWW 2021 **Agathe Balayn**, Panagiotis Soilis, Christoph Lofi, Jie Yang, and Alessandro Bozzon. What do you mean? interpreting image classification with crowdsourced concept extraction and analysis. In *Proceedings of the Web Conference 2021*, pages 1937–1948, 2021
- MSc thesis 2018 **Agathe Balayn**. On the fairness of crowdsourced training data and Machine Learning models for the prediction of subjective properties. The case of sentence toxicity: To be or not to be #\$\$%! toxic? To be or not to be fair? Master's thesis, Delft University of Technology, the Netherlands, 2018
- HCOMP (WS) 2018 **Agathe Balayn**, Panagiotis Mavridis, Alessandro Bozzon, Benjamin Timmermans, and Zoltán Szlávik. Characterising and mitigating aggregation-bias in crowdsourced toxicity annotations. In *Proceedings of the 1st Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing co-located the 6th AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2018)*, pages 67–71, 2018
- RO-MAN 2018 **Agathe Balayn**, Heike Brock, and Kazuhiro Nakadai. Data-driven development of virtual sign language communication agents. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 370–377. IEEE, 2018
- SIMPAR 2016 **Agathe Balayn**, Jeffrey Frederic Queißer, Michael Wojtynek, and Sebastian Wrede. Adaptive handling assistance for industrial lightweight robots in simulation. In *2016 IEEE International Conference on Simulation, Modeling, and Programming for Autonomous Robots (SIMPAR)*, pages 1–8. IEEE, 2016

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)

Languages

- | | | | |
|---------|--|----------|--------------------------------|
| French | Native speaker. | Mandarin | Elementary proficiency. |
| English | Professional working proficiency. | German | Elementary proficiency. |

Extra-Curricular Activities

- 2021-2023 Participation in various **art activities** within the university, *Painting (watercolor, oil, acrylic), sketching*
- 2021-2022 Member of the **Slow Reading** group on AI and gender inequality, *Computer Science perspective to the artist collective*
- 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 collection via volunteering activities*