Agathe Balayn

Curriculum Vitae

Koornmarkt 82A, 2611EJ, Delft, the Netherlands
08/01/1995, Paris, France
☎ *+33.6.99.55.72.23*
✉ *a.m.a.balayn@tudelft.nl*
🖱 *https://agathe-balayn.github.io/*

Research Mission

I am interested in characterizing theories and practices for developing and evaluating machine learning (ML) models with regard to safety issues and societal harms, and in proposing supporting methods and workflows.

Education

04/2019 - now **PhD candidate in Computer Science**, *HCI*, Delft University of Technology (the Netherlands).
- 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).

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
- Completion of the Honours Programme of the university (additional 20 ECTS)
- 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 - 09/2018 **MSc in Control Systems**, *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.

Professional Experiences

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.

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.

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 based on motion capture data.

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 and model-space learners for an industrial lightweight robotic arm (Universal Robots UR5).
- Obtained one individual scholarship based on merit (Erasmus Plus).

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.

Selected publications and projects

Human-centered studies, frameworks, and tools.

CHI 2022 **A. Balayn**, N. Rikalo, C. Lofi, J. Yang, A. Bozzon. **How can Explainability Methods be Used to Support Bug Identification in Computer Vision Models?**.

Under review **A. Balayn**, J. Yang, U. Gadiraju. **Beyond Fairness?! A Study of the Algorithmic Harms Envisioned by Machine Learning Practitioners who Rely on Fairness Toolkits**.

Under review **A. Balayn**, N. Rikalo, J. Yang, A. Bozzon. **Faulty or Ready? Handling Failures in Deep-Learning Computer Vision Models: Practices, Challenges, and Needs**.

CVPR (WS) 2021 **A. Balayn**, B. Kulynych, S. Guerses. **Exploring Data Pipelines through the Process Lens: a Reference Model for Computer Vision.** *At the Beyond Fairness workshop*.

Under review M. Yurrita, T. Draws, **A. Balayn**, A. Bozzon. **Disentangling Fairness Perceptions in Algorithmic Decision-Making: the Effect of Explanations, Human Oversight, and Contestability**.

Methods and systems.

WWW 2021 **A. Balayn**, P. Soilis, C. Lofi, J. Yang, A. Bozzon. **What do You Mean? Interpreting Image Classification with Crowdsourced Concept Extraction and Analysis**.

WWW 2022 **A. Balayn**, G. He, A. Hu, J. Yang, U. Gadiraju. **Ready Player One! Eliciting Diverse Knowledge Using A Configurable Game**.

HCOMP 2022 G. He, **A. Balayn**, J. Yang, U. Gadiraju. **It's like Finding a Polar Bear in the Savannah! Concept-level AI Explanations with Analogical Inference from Commonsense Knowledge**.

Ongoing **A. Balayn**, J. Yang. **ARCH: a Framework to Optimize Concept-based Failure Diagnosis in Computer Vision Models**.

Reviews of the literature.

VLDBJ 2021 **A. Balayn**, C. Lofi, G-J. Houben. **Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches**.

FAccT 2022 M. Yurreta, D. Murray-Rust, **A. Balayn**, A. Bozzon. **Towards a Multi-Stakeholder Value-based Assessment Framework for Algorithmic Systems**.

Technical report **A. Balayn**, S. Guerses. **Beyond Debiasing: Regulating AI and its Inequalities.** *Technical report written for the European Digital Rights (EDRi) Organization*.

Under review **A. Balayn**, A. Tocchetti, L. Corti, M. Yurrita, P. Lippman, M. Brambilla, J. Yang. **A.I. Robustness: a Human-Centered Perspective on Technological Challenges and Opportunities**.

Professional Services

Reviewer, *CHI'21-22, CSCW'21-22, IUI'20-21, HCOMP'20-21, WWW'20-22, AAAI'22, NeurIps'22, HyperText'20-22, ROMAN'20-21, CIKM'21, NAACL'21, ChineseCHI'20*.

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

Presentations at local events, *the first symposium on Biases in Human Computation and Crowdsourcing (BHCC), FAccT PhD consortium, the Dutch-Belgian Database Day (DBDBD), ICT.Open, Public Interest AI workshop*.

Teaching and Mentorship

Teaching, material designer and teaching assistant for the introduction to ML fairness 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 nine Master students for their MSc thesis projects; a group of five second year Bachelor students for a software engineering project; and four groups of 4 Master students for crowdsourcing+AI projects..

Technical Skills

Programming languages *Most experienced*: **Python** (TensorFlow, Keras, Scikit-learn, etc.), **MATLAB, C++** (OROCOS, Gazebo environment). *Some experience*: **C, Java, Maple, HTML, CSS, PHP, Javascript**.

Others **Working knowledge of Linux, Git, common software suites (Office), LaTeX**.

Languages

French	**Native speaker.**	Mandarin	**Elementary proficiency.**
English	**Professional working proficiency.**	German	**Elementary proficiency.**