Agathe Balayn

Curriculum Vitae

Koornmarkt 82A, 2611EJ, Delft, the Netherlands
08/01/1995, Paris, France
☎ *+33(0)6.99.55.72.23*
✉ *a.m.a.balayn@tudelft.nl*

Research Mission

While machine learning is increasingly recognized as a powerful tool for diverse applications, it can also cause or reinforce safety issues and societal harms. I am interested in understanding and characterizing current theory and practices for developing and evaluating machine learning models with regard to these issues, and in proposing new, supporting methods, workflows, and processes.

Education

04/2019 - now **PhD candidate in Computer Science.**, *Delft University of Technology*, The Netherlands, Department: Web Information Systems.

Topics of interest: debugging computer vision models for biases and feature failures using explainability; understanding and revisiting the practices for evaluating machine learning models on diverse algorithmic harms; using mixed methods, e.g. literature-based work, empirical, qualitative studies, development of new methods and workflows.

09/2016 - 09/2018 **MSc in Computer Science. Data Science and Technology track**, *Delft University of Technology*, The Netherlands.

- 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): *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?*

09/2014 - 09/2018 **MSc in Embedded Systems**, *ENSTA ParisTech (Top French Grande École, graduate school in Engineering)*, France, (Program leading to a "Diplôme d'ingénieur", equivalent of a masters).

- Strong component of Control, Informatics and Signal.
- Graduated first year 2nd of the class out of 144 students. GPA: 4.0/4.0.

2012-2014 **Preparatory class**, *Lycée du Parc*, France, 2 years of intensive undergraduate program (Mathematics, Physics, Engineering Sciences) preparing for the national competitive entrance examination to graduate schools in engineering.

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 machine learning pipelines for the inference of subjective labels.

- Experimentations and writing of scientific publications. Publications at the Human-Centered Machine Learning workshop of CHI 2019, and the Rigorous Evaluation of AI Systems workshop of HCOMP 2019, and a survey in the Transactions on Social Computing.

11/2017 - 09/2018 **Graduate Intern at IBM Benelux (Master's thesis)**, *Center for Advanced Studies, Amsterdam, the Netherlands*, Study of biases and fairness in crowdsourced data and machine learning models for the prediction of subjective properties, with the use-case of sentence toxicity prediction.

- Design and implementation of a new fairness metric. Experiments on crowdsourcing techniques for training dataset collection, and on machine and deep learning models for fairer predictions.
- Publication at the CrowdBias workshop of HCOMP2018.

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 (mainly RNN, LSTM neural networks) for Japanese sign language synthesis and recognition based on Kinect and motion capture data.
- Publication at RO-MAN 2018.

05/2016 - 07/2016 **Research Intern at the Research Institute for Cognition and Robotics (CoR-Lab)**, *Bielefeld University, 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), able to cope with unknown weights at the end-effector.
- Publication at SIMPAR 2016.
- 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 (canyoning, kayaking, basic ski lessons), set up air balloons, barista. Entirely Japanese-speaking team.

Publications

under review **Agathe Balayn**, Natasa Rikalo, Jie Yang, Alessandro Bozzon. **Bugged? Debugging Deep-Learning Computer Vision Models: A Study of Practices, Challenges, and Needs**.

under review Mireia Yurreta, **Agathe Balayn**, Dave Murray-Rust, Alessandro Bozzon. **Towards a multi-stakeholder value-based assessment framework for algorithmic systems**.

under review 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**.

CHI 2022 **Agathe Balayn**, Natasa Rikalo, Christoph Lofi, Jie Yang, Alessandro Bozzon. **How can Explainability Methods be Used to Support Bug Identification in Computer Vision Models?** *At the Conference on Human Factors in Computing Systems (CHI) 2022*.

WWW 2022 **Agathe Balayn**, Gaole He, Andrea Hu, Jie Yang, Ujwal Gadiraju. **Ready Player One! Eliciting Diverse Knowledge Using A Configurable Game.** *At the Web Conference (WWW) 2022*.

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

VLDBJ 2021 **Agathe Balayn**, Christoph Lofi, 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.** *At the Journal of Very Large Database Systems (VLDBJ) 2021*.

TSC 2021 **Agathe Balayn**, Jie Yang, Zoltan Szlavik, Alessandro Bozzon. **Automatic Identification of Harmful, Aggressive, Abusive, and Offensive Language on the Web: A Survey of Technical Biases Informed by Psychology Literature.** *At the ACM Transactions on Social Computing (TSC)*.

CVPR (WS) 2021 **Agathe Balayn**, Bogdan Kulynych, Seda Guerses. **Exploring Data Pipelines through the Process Lens: a Reference Model for Computer Vision.** *At the Beyond Fairness: Towards a Just, Equitable, and Accountable Computer Vision workshop co-located with CVPR 2021*.

HCOMP demo 2021 **Agathe Balayn**, Gaole He, Andrea Hu, Jie Yang, Ujwal Gadiraju. **FindItOut: A Multiplayer GWAP for Collecting Plural Knowledge.** *Best demo award at the AAAI Conference on Human Computation and Crowdsourcing (HCOMP) 2021*.

WWW 2021 **Agathe Balayn**, Panagiotis Soilis, Christoph Lofi, Jie Yang, Alessandro Bozzon. **What do You Mean? Interpreting Image Classification with Crowdsourced Concept Extraction and Analysis.** *At the Web Conference (WWW) 2021*.

HCOMP (WS) 2019 **Agathe Balayn**, Alessandro Bozzon. **Designing Evaluations of Machine Learning Models for Subjective Inference. The Case of Sentence Toxicity.** *At the Rigorous Evaluation of AI Systems workshop co-located with HCOMP 2019*.

CHI (WS) 2019 **Agathe Balayn**, Zoltan Szlavik, Alessandro Bozzon. **Unfairness towards subjective opinions in Machine Learning.** *At the Human-Centered Machine Learning Perspectives workshop co-located with CHI 2019*.

RO-MAN 2018 **Agathe Balayn**, Heike Brock, Kazuhiro Nakadai. **Data-driven development of Virtual Sign Language Communication Agents.** *At the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN) 2018*.

MSc thesis **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 thesis, TU Delft repository, 2018.

HCOMP (WS) 2018 **Agathe Balayn**, Panagiotis Mavridis, Alessandro Bozzon, Benjamin Timmermans, Zoltan Szlavik. **Characterising and Mitigating Aggregation-Bias in Crowdsourced Toxicity Annotations.** *At the CrowdBias workshop co-located with HCOMP 2018*.

SIMPAR 2016 **Agathe Balayn**, Jeffrey Frederic Queißer, Michael Wojtynek, Sebastian Wrede. **Adaptive handling assistance for industrial lightweight robots in simulation.** *At the IEEE International Conference on Simulation, Modeling, and Programming for Autonomous Robots (SIMPAR) 2016*.

Professional Service

Reviewer, *HCOMP, WWW, IUI, HyperText, ROMAN, CIKM, NAACL, UMUAI, IEEE Access, ChineseCHI.*

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

Presentations at various local events, *the first symposium on Biases in Human Computation and Crowdsourcing (BHCC), FAccT PhD consortium, the Dutch-Belgian Database Day (DB-DBD), ICT.Open.*

Teaching and Mentorship

Teaching.

- Teaching assistant for the Crowd Computing course at the TU Delft Computer Science faculty.
- Teaching assistant for the seminar on Web Information Systems at the TU Delft Computer Science faculty.
- Teaching assistant and material creator for the introduction to machine learning fairness session within the inter-faculty Machine Learning course at TU Delft.
- Teaching introductory lectures on AI ethics at the TU Delft Computer Science faculty within the Web Science Engineering course.

Mentorship.

- Supervision of five Bachelor students for their BSc thesis projects.
- Supervision of nine Master students for their MSc thesis projects.
- Supervision of five second year Bachelor students for a software engineering project.
- Supervision of four groups of 4 Master students for crowdsourcing+AI projects.

Technical Skills

Programming languages *Most experienced with*: **Python** (TensorFlow, Keras, Scikit-learn, etc.), **MATLAB, C++** (OROCOS, Gazebo simulation environment).

Some experience with: **C, Java, Maple. HTML, CSS, PHP, Javascript** (D3 library).

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.**