Title: Exploring Data Pipelines through the Process Lens: a Reference Model for Computer Vision and its Implications for Understanding Algorithmic Harms

Speaker: Agathe Balayn, Delft University of Technology

Abstract:

Researchers have identified datasets used for training computer vision models as an important source of hazardous outcomes, and continue to examine popular computer vision datasets to expose their harms. These works tend to treat datasets as objects, or focus on particular steps in data production pipelines. We argue here that we could further systematize our analysis of harms by examining computer vision data pipelines through a process-oriented lens that captures the creation, the evolution and use of these datasets.
As a step towards cultivating a process-oriented lens, we embarked on an empirical study of computer vision data pipelines informed by the field of method engineering. We performed a systematic survey of computer vision data pipelines (and of the ways they are made transparent) as reported in publications detailing popular datasets. Based on this, we proposed an initial reference model of computer vision data pipelines.
Besides exploring the questions that this endeavor raises, we will discuss how the process lens could support researchers in discovering understudied issues, and could help practitioners in making their processes more transparent. Especially, we will mention two ongoing works stemming from this reference model: 1) an empirical study of the brittleness of output fairness in relation to small parameter changes in the data pipelines, and 2) an empirical exploration of the limitations of fairness toolkits in relation to data pipelines and practical deployment.

References:
This presentation is based on a) the speaker's paper *Exploring Data Pipelines through the Process Lens: a Reference Model for Computer Vision* presented at the CVPR21 workshop "beyond fair computer vision", b) ongoing work to be submitted to CSCW2022 and FAccT23.