The production and solidification of a ML dataset

My provocation is a tale about the strategies used by dataset creators to produce and solidify one version of several possible machine learning (ML) realities. I explore the work involved in assembling, disseminating, and entrenching an ML dataset within the broader ML ecosystem. This needs special attention since benchmark datasets, like ImageNet, display effects that go beyond simple reference points - they also impose norm-setting mechanisms (Denton et al., 2021; Koch et al. 2021). My descriptive-analytic provocation comes from an ethnographic case study of a benchmark dataset of first-person footage of *unscripted daily activities* for computer vision (CV). The provocation rests on the assumption that a ML dataset performs a specification of a conceptualization attributed to it[1]. The dataset under study produces a specification of *daily activities,* which as I will show is situated and partial (cf. Raji et al. 2021).

A growing area of social research on ML culture is the analysis of ML datasets. ML datasets are composed solely for the sake of a technological other: ML models in-the-making. Trained ML models are directly implicated by the composition of data it has been fed – serving as the "ground-truth" (Jaton, 2017). One important area of focus in social analyses of ML datasets concerns power in the social organization and exploitation of data collection (the harvest of online materials) and data annotation work (often crowdsourced and invisibilized) (Irani, 2013). Another major concern in this area, and rightly so, regards the use and operationalization of identity categories (e.g., Scheuerman et al. 2020). Post-structural insights and claims of recognition have been brought into contact with the binary relations common in the classification procedures of computation, hitherto contained by a culture shaped by masculine, white, capitalist, military-industrial, and solutionist assumptions and ideals.

In my view, a vital insight from this second focus area is that ML datasets can be subject to scrutiny on its own terms, according to its own standards. While I owe a debt of gratitude to the work of these scholars, I aim to contribute to the current body of useful ways to perform dataset analyses, such as critical historiographies (e.g., Denton et al. 2020, 2021), by using an analytical strategy that tends to an object still under construction: a form of *datasetting-in-action* (cf. Latour 1987) that may lead to complementary results, though the analysis is preliminary. We can therefore turn to an inquiry of the ways that specifications (as situated and partial collections of data) of conceptualizations (e.g., 'human faces' or 'everyday life') are achieved in practice.

Approaching the case at hand

The dataset under study is the result of a corporate-academic collaboration. A group of corporate engineers announced the dataset in a technical publication, a promotional video, and an online seminar during 2021. As a basic science project, the home of the dataset is in one of several technoscientific vectors – first-person computer vision – that requires wide collaboration to drive meaningful progress, and where materials can be gatekept and access to supporting ecologies limited. The corporation is the main organizing actor, while academic institutions were responsible for the recruitment of research participants, data collection, and associated ethics reviews conducted by independent university boards. The dataset includes footage recorded by almost 1000 research participants. A project document reveals that "two firms in Africa" were commissioned to annotate the dataset through close to 30 years of 24/7 annotation work.

The specification of 'daily activities'

Imagine this massive dataset as a visual cascade where individual data entries follow one after another as a reel. The observer sees first-person footage of an artist's hands shading a portrait; another scene shows another person reaching to refuel a machine; doing home improvements with a circular saw, a sewing machine being operated, and so on and so forth. The dataset

[1] I derive this formulation from how *ontology* has been defined in computer scientist circles. Moreover, and drawing on Ian Hacking (1983), a specification "does not represent but intervenes", and thus recomposes a version of the conceptualization that is 'captured' for computational manipulation.

creators have provided a high-level overview of the dataset's recorded activities: what it shows, which is useful to understanding their specification of daily life. About 50 % of the footage represents handicrafts, preparing food, carpentry, and domestic work such as washing clothes. The other 20 % involve activities such as shopping, eating, reading, and playing music. The remainder of the dataset represent fewer instances of activities such as, playing various sports and games, watching TV and other screens, walking indoors and outdoors, commuting and interacting with others.

Activities that may be ambiguous in character, ethically controversial, or considered as too private are missing from this specification. When interviewed[2] about their experience of recording their lives, one research participant said that "wearing cameras was a challenge at the beginning", but "got used to it as the days passed by". Another framed it as a "unique experience", and a third participant brought up a felt "performance anxiety". These reports destabilize the unscripted, 'in-the-wild' status attributed to this specification, suggesting that participants performed and stylized activities which they considered adequate from the perspective of others.

Text and symbols – annotated metadata for computational reference – have been superimposed on the footage. Ongoing action and object-centered annotations such as "C [camera wearer] folds pizza box" dominate the footage. Annotators are provided with standard-setting examples regarding how to proceed with describing the footage, "Fix the start time to the moment the person takes up the knife and apple and the end time to the moment they are done, then write "C is slicing an apple" into the text area". The camera wearers' own interpretations, intentions, goals, and affects are not of concern, according to these guidelines. The dataset *shows*, and as the annotation guidelines explicate, it has been made to *tell* in specific ways.

Solidifying the specification

The dataset is tightly linked to a set of benchmark challenges, promoting distinct work packages. Recently, the corporation organized a competition with standardized evaluative metrics to reward those teams with the best technical performance on specific tasks with a cash prize, and subsequent contests will follow. The challenges revolved around what the ML community calls episodic memory, forecasting, hand and object manipulation, and so on. To date the leaderboards are populated by almost 100 contestants distributed in different team constellations.

In one of my interviews with a benchmark contestant, working at a separate Big Tech company, he describes his uptake of the dataset accordingly: "There was a lot of circling around it internally. My colleagues were emailing around about it. I was working on a related task before this dataset was released, and I figured that I could make use of this additional benchmark to evaluate my prior work….". As suggested by my interlocutor, the benchmark becomes an arena for testing model performance. As new actors come into play, it is also a moment in which the dataset picks up pace on the way to becoming entrenched in the ML ecosystem.

We might thus say that a central device for solidifying the specification is the 'benchmark challenge', which orchestrates temporary work efforts between diverse actors in the community (cf. Raji et al. 2021). A benchmark that has been certified by the relevant actors reconfigures weak horizontal networks into temporary pressure zones, in which the specification of a ML dataset becomes increasingly solidified. The dataset is in an early stage at the time of writing, although around 60 publications have cited the technical publication of the dataset. Through following additional 'traces', such as the production logs at Github, and the project forum, there

[2] The questions include the reasons for their participation, their recorded activities, if they watched it in retrospect, and their opinions on first-person cameras and big research projects.

is an abundance of work-in-the-making, not least troubleshooting-related concerns. These data suggest a continuous uptake of this specification within the ML community, and make up a vantage that may be uniquely leveraged by *datasetting-in-action*.

Conclusion

Datasetting-in-action, the mode of analysis used here, seem to nudge a dynamic and lengthy chain of human actors and technical elements into view. We may conclude that a dataset is situated "as culture" as a produced artifact as well as an entity that holds a situated, partial specification of a conceptualization – in this case of *daily life* – in place (Seaver, 2017). It does so not solely by its own accord, but through the intricate mingling between the 'social' and 'technical' and the iterative expansion of those tangled interrelations. Most strikingly, the production and solidification occur through strategies such as eliciting a) research participants to generate data; b) crowdsourcing annotators to sprout computer-legibility; c) outsider scientists' conduction of model work; and d) new models enveloping the specification within the ML ecosystem. To better comprehend the scattered dataset origins of currently proliferating pre-trained ML models, it is crucial to pay attention to the specifications of conceptualizations to which ML datasets are committed.

References:

Denton, Emily L. Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole and Morgan Klaus Scheuerman (2020) "Bringing the People Back In: Contesting Benchmark Machine Learning Datasets." CML 2020 Workshop: Participatory Approaches to Machine Learning (PAML)

Denton, Emily L. Alex Hanna, Razvan Amironesei, Andrew Smart and Hilary Nicole (2021) "On the genealogy of machine learning datasets: A critical history of ImageNet." *Big Data & Society*, 8(2). https://doi.org/10.1177/20539517211035955

Hacking, Ian (1983) *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge: Cambridge University Press

Irani, Lilly (2013) "The Cultural Work of Microwork." *New Media and Society*, 17(5), 720-739

Jaton, Florian (2017) "We get the algorithms of our ground truths: Designing referential databases in digital image processing." *Social Studies of Science* 46(7): 811-40.

Koch, Bernard, Emily L. Denton, Alex Hanna and Jacob G. Foster (2021) "Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research." NeurIPS 2021, arXiv:2112.01716.

Latour, Bruno (1987) *Science in Action: How to Follow Scientists and Engineers through Society*. Cambridge, MA: Harvard University Press

Raji, Inioluwa Deborah and Bender, Emily M. and Paullada, Amandalynne Denton, Emily L. and Hanna, Alex (2021) "AI and the Everything in the Whole Wide World Benchmark." NeurIPS 2021, arXiv:2111.15366.

Scheuerman, Morgan Klaus, Kandrea Wade, Caitlin Lustig, and Jed R. Brubaker (2020) "How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis." Proc. ACM Hum.-Comput. Interact. 4, CSCW1, Article 58. https://doi.org/10.1145/3392866

Seaver, Nick (2017) "Algorithms as culture: Some tactics for the ethnography of algorithmic systems." *Big Data & Society*, *4*(2).
https://doi.org/10.1177/2053951717738104