Case Studies: Fairness 2

Case 1: Analyzing a datasheet to spot ethical issues

This case is extracted and adapted with permission from K. Boyd: Boyd, K. L. (2021). Datasheets for Datasets help ML Engineers Notice and Understand Ethical Issues in Training Data. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 438:1-438:27. https://doi.org/10.1145/3479582

Using the datasheet provided in appendix 1.1, answer the following questions:

1. Thinking about a **range of stakeholders**, find at least one ethical issue related to **safety** *(Note: remember that we take the general meaning of "safety" as negative impact from the system on its environment).*
2. Find one ethical issue related to **fairness**
3. Based on your analysis in the previous questions, if you were to use this dataset for training a machine learning model able to identify faces, which type of ethical issue(s) could manifest in the model?

Case 2: Humanizing COMPAS data

We have previously seen (in the videos and in the notebook) various ethical issues with the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS). In this exercise, we will use it again to practice the "people behind the data" strategy, as it opens a large door to creativity, imagination, and critical thinking.

Stage 01: Search for the dataset documentation, source materials, source data.
We provide you with the following documents - to download from courseware:

- The source questionnaire that defendants fill out for a COMPAS assessment
- A random extract of the dataset provided by ProPublica (which is otherwise quite large for this exercise) and the description of the columns in the dataset.

One issue we have with this source data is that the values collected from the questions in the questionnaire are not present in the dataset. We don't know either how the answers from the questions are used to compute the COMPAS aggregated score. On the other hand, the dataset contains a range of demographic information, judicial information and the aggregated scores from the COMPAS. Therefore you will combine the two sources of information in your stories.

Stage 02: Select a few inspiring questions/variables/columns from the dataset or its documentation.
👉 In the source questionnaire, **select around 5 questions** that:

- Can help you imagine characteristics of people represented in the data
- Can lead to odd, extreme or inconsistent values in the dataset

Stage 03: Select a few rows from the dataset, read the data, imagine the people behind the data, their profile and their stories.
👉 In the random extract of the dataset that we provide, **select 1 row based on specific characteristics** (e.g. gender, race, score…).
For this row, write the story of the person represented by the data:

- What could explain the value they get on this attribute, or the score they obtain for this scale?
- What is their character?
- What is their past?
- How about their family?

Stage04: Write down your conclusions in terms of ethical impacts/risks.
👉 Answer these questions:

- What have you learned about the data based on your exploration?
- Which potential harmful impacts could using this data generate?
- What would be your next steps: would you use these data? What other possibilities would you have?

Sources:

- Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). *Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica.* https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
- Questionnaire "Sample-Risk-Assessment-COMPAS-CORE" provided by Propublica: https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE/
- Github repository provided by Propublica: https://github.com/propublica/compas-analysis

Case 3: Harms modeling

Scenario:

An advanced Generative AI tool called "PAthfInder", is designed to provide users with personalized career advice and job recommendations. PAthfInder has been trained on over twenty years of labor market data including job postings, résumés, career trajectories, and global economic trends. The system is continuously updated with real-time data from job boards, professional networks, and government employment statistics to stay aligned with emerging roles. It analyzes an individual's skills, experiences, and even soft skills inferred from their digital footprint to suggest tailored career paths, reskilling opportunities, and specific companies where they are most likely to thrive. Rapidly, ImagineX becomes the tool of predilection for everyone to find a new job adapted for their career and skills.

Exercise:

Find **one type of harm for each category** in the simplified harms modeling table below:

Category	Type of harm	Description of harms in the scenario
Humans	*Physical injury*	*PAthfInder could propose jobs that are not adapted for a person's health situation. For example, recommending a career as a lumberjack to someone with an already weakened back.*

Allocation of resources		
Human Rights		
Social Systems		

Appendix

1.1 Datasheet

Motivation	
For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.	This dataset was created to provide images that can be used to study face detection in an unconstrained setting where image characteristics (such as pose, illumination, focus), subject demographic makeup (such as age, gender, race) or appearance (such as hairstyle, makeup, clothing) cannot be controlled.
Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?	The initial version of the dataset was created by researchers at the imaginary BBB corporation.
Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.	The construction of the original dataset was funded by BBB corporation.
Composition	
What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.	The dataset consists of just over 65,000 high-quality PNG images at 1024×1024 resolution. Each instance includes at least one human face. Images were crawled from Photobucket to increase the likelihood that it has good coverage of accessories, including glasses, sunglasses, make-up, hair accessories, hats, etc.

How many instances are there in total (of each type, if appropriate)?	65 104
Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable)	A full-resolution sample of the data is available for download. The sample is randomly selected, and so expected to be representative of the larger dataset in terms of image characteristics.
What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.	The data consists of unprocessed images of faces.
Is there a label or target associated with each instance? If so, please provide a description.	There is no label associated with each instance.
Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.	Instances are not missing information, but metadata was stripped from the original images to preserve the privacy of Photobucket users. They do not contain labels of any kind.
Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.	There are no links.
Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?	The dataset is self-contained.
Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.	This data comprises communication that was intended to be public, but publishers (individual Photobucket users) may not have anticipated that it would be used in this way. Further, publishers may have made images available of people other than themselves without permission.
Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.	Images were not thoroughly checked for offensive material. If you find anything that you believe should be removed, please email the creators and let us know. We will consider whether to drop the image and whether to report the original image to Photobucket.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.	Yes
Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.	The dataset does not identify subpopulations.
Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.	It is possible to indirectly identify publishers and subjects using reverse image search
Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.	Images may contain information that allows people to make inferences about race, ethnicity, sexual orientation, religious beliefs, political opinions, memberships, locations, health information, or criminal history. However, because these images were shared publicly, we assume that that information is not considered too private to be shared.
Collection Process	
How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.	The images were crawled from Photobucket and automatically aligned and cropped using dlib. The individual images were published in Photobucket by their respective authors under either Creative Commons BY 2.0, Creative Commons BY-NC 2.0, Public Domain Mark 1.0, Public Domain CC0 1.0, or U.S. Government Works license. All of these licenses allow free use, redistribution, and adaptation for non-commercial purposes.
What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated? • If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?	Images were collected using a custom crawler to limit data scraped to those including permissive Creative Commons Licences. Sample data available for download was sampled randomly with a visual check for offensive content and basic demographic representativeness.
Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?	No humans were involved in data collection and the data is not labeled. Humans involved in developing, testing, and executing the script and preparing it for publication were full time, paid

	employees of BBB.
Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.	Data was collected in 2018. Some data has been deleted since then, none has been added.
Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.	No separate ethical review process was conducted.
Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.	Yes
Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?	Via a third party (Photobucket)
Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.	Individuals were not notified of the data collection. They were aware that the images were public.
Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.	No
If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).	Consent was not provided, but individuals who are in the dataset can petition to have their images removed by contacting BBB.
Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis)been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation	No