Case Studies: Empowerment 2

Proposed Answers

Case 1: Modeling bad actors in the context of Empowerment

VoiceX is an advanced AI tool designed to create realistic audio deepfakes, with the inclusive goal of empowering individuals with communication disabilities. It enables users who struggle with vocal communication to convey clear and authentic messages, either with their own voice by uploading a sample, or with a pre-recorded voice of their choice available in the tool.

Apply the **Bad Actors Modeling Strategy** to identify and analyze potential harmful actions or negative consequences that could arise on VoiceX, using the five motivation categories (Money, Politics, Entertainment, Ideas, Self-interest).

Consider the following questions:

- ➔ What harmful actions can be taken in each category?
- ➔ How might these actions impact users and the platform?

Proposed answer:

Money

- **Scam:** Scammers could use VoiceX to create realistic audio deepfakes that mimic the voices of trusted individuals (family members, bank advisors) to deceive victims into transferring money or sharing sensitive information.
- **Exploitation of famous voices:** Scammers could create audio deepfakes of known personalities and post it on social media platforms to get advertisement revenue or to sell fake products (e.g. see this news article).

Politics:

- **Political manipulation:** VoiceX could be exploited to create audio deepfakes of political figures making false statements, endorsing false ideologies, or participating in non-existent events. These deepfakes could be used to spread misinformation and manipulate public opinion during elections.
- **Fake news:** Political activists or adversaries could use VoiceX to generate audio content that spreads propaganda or undermines political opponents, contributing to political polarization.

Entertainment:

- **Exploitation for mocking content:** VoiceX could be misused to create humorous or viral content, for instance at the expense of specific individuals (e.g. celebrities or political figures). This could involve creating deepfakes that mock or parody users, turning their attempts at communication into a form of entertainment for others.
- **Trolling:** VoiceX could be used to create harmful or embarrassing audio content of celebrities, influencers, or individuals (e.g. victims of bullying), which could be shared widely for entertainment, leading to reputational damage.

Ideas:

- **Spread of misinformation:** VoiceX could be used to generate deepfakes that spread harmful misinformation, e.g. diffused through podcast or streaming platforms. Branded as educational content, these fake resources could reach a large audience not sensitised to this type of issue, e.g. young publics.

Self interest:

- **Manipulation of personal relationships:** Individuals might use VoiceX to create deepfakes that manipulate personal relationships, such as creating fake audio of loved ones or caregivers. This could be used to exploit trust, manipulate emotions, or gain personal advantage.

To learn more: Deepfakes and the crisis of knowing | UNESCO (https://www.unesco.org/en/articles/deepfakes-and-crisis-knowing)

Case 2: Ethical Speculation

Task:

Imagine an episode of "Escape the Mirror" (our version of "Black Mirror") where a character is disempowered because of software (e.g. deceived, manipulated, left without recourse…).

You can get inspiration from one of the following topics, or choose any other topic related to Empowerment questions:

1. Trust and automation bias in algorithmic decision-making
2. Chatbots (manipulation, dependency, and privacy risks)
3. Deepfakes
4. Algorithms used by governments (e.g. social security…) or education institutions
5. Privacy and surveillance

Feel free to search the web and read news articles for inspiration.
Here are some:

- Instagram and Threads moderation is out of control - The Verge
- SocialAI offers a Twitter-like diary where AI bots respond to your posts | TechCrunch
- Researchers say AI transcription tool used in hospitals invents things no one ever said | AP News
- Someone Put Facial Recognition Tech onto Meta's Smart Glasses to Instantly Dox Strangers
- FTC Announces Crackdown on Deceptive AI Claims and Schemes | Federal Trade Commission
- Google Serving AI-Generated Images of Mushrooms Could Have 'Devastating Consequences'

Remember from the Introduction module, there are two steps to ethical speculation:

1. The first part is to create a dark dystopian episode that emphasizes an ethical issue related to software. Imagine a medium or distant future where technology causes significant impacts on society. Focus on a fictional person, or small group, whose story demonstrates this dystopian scenario. Your episode pitch must include: a title, a fifty to one hundred word summary of the episode, and an image to represent your dystopian tale.
2. You will then design the happy ending for your episode. Use the template to outline the ethical problem, its immediate and future consequences. Imagine how to resolve the ethical issues from part 1, think about solutions that lead to a better future and explain their positive outcomes .

Proposed answer:

Proposition 1:

ETHICAL ISSUE
Algorithmic bias, lack of transparency, educational inequality

IMMEDIATE AND FUTURE CONSEQUENCES
- **Immediate:** Students experience confusion and distress as they are unfairly graded, with no transparency into how or why the system determined their results. Trust in the educational institution diminishes as students feel the system cannot recognize their efforts.
- **Future:** Widespread reliance on biased, opaque grading algorithms leads to systemic inequality, favoring certain demographics over others due to inherent biases in training data. In the long term, students who are misjudged by the algorithm struggle to access higher education and job opportunities, creating a generation of disempowered youth and exacerbating social inequality.

SOLUTION AND/OR POSITIVE OUTCOME
After discovering that many other students faced the same issue, Sarah takes the initiative to lead a software engineering student association to investigate the source of the problem and develop an effective strategy for communicating with the university's remaining human administration. Through their efforts, the association uncovers that algorithmic bias, not the quality of their work, was responsible for the poor grades. They successfully convince the educational institution to collaborate with AI ethics experts to improve the AI system and reestablish a balanced human/AI workforce, with humans making final decisions. This change strengthens the relationship between students and the administration, fostering a renewed sense of trust and transparency.

CREATORS: Noa

Proposition 2:

TITLE
The monitored mind

SUMMARY
In 2042, corporations have widely adopted "MoodTrack," a wearable device mandated for all employees, which monitors mental and emotional states through biometric and neural feedback. Julia, a dedicated software engineer, notices her weekly performance review suddenly dropping. MoodTrack flags her as "emotionally unstable" due to stress at home, and she's placed under performance surveillance. As Julia struggles to appear "emotionally optimal" to retain her job, she realizes there's no way to escape the constant scrutiny of MoodTrack, leaving her feeling trapped and dehumanized by a system that punishes her for natural emotions.

CREATOR: Noa

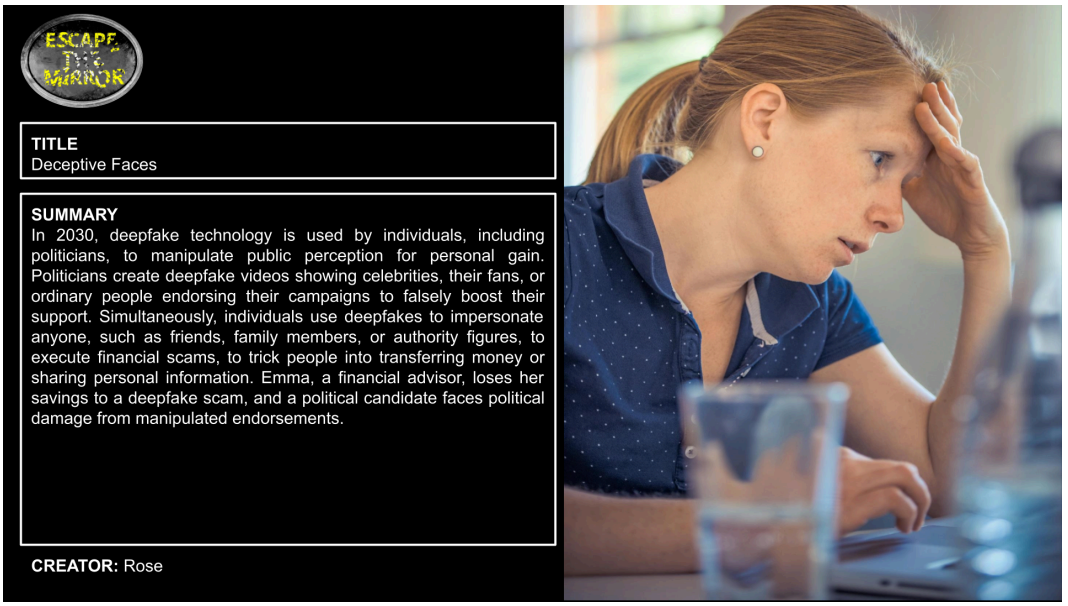
ETHICAL ISSUE
Privacy invasion, emotional manipulation, workplace exploitation

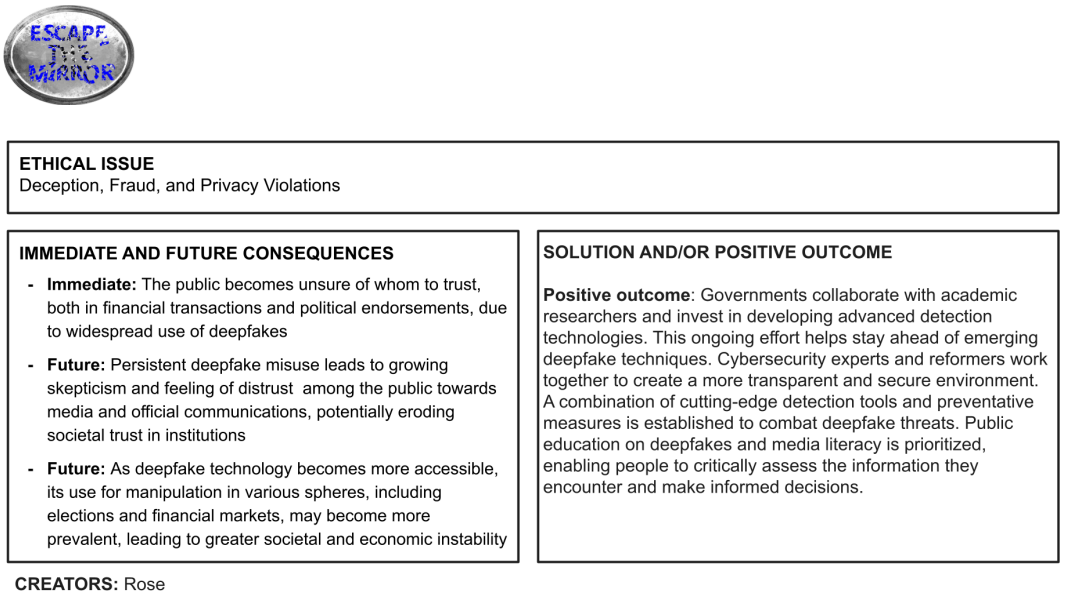
IMMEDIATE AND FUTURE CONSEQUENCES
- **Immediate:** Employees like Julia face pressure to suppress natural emotional responses to avoid being penalized by MoodTrack, leading to heightened stress, mental health issues, and a toxic work environment. The trust between employer and employee erodes as individuals feel constantly surveilled and judged by their moods rather than their work performance.
- **Future :** As MoodTrack technology proliferates, a culture of emotional suppression spreads across workplaces, causing long-term psychological damage to employees. The normalization of emotional surveillance leads to widespread mental health deterioration, increased burnout, and loss of individuality. Companies prioritize "emotional optimization" over genuine well-being, and people's lives become driven by the need to meet emotion-based performance metrics.

SOLUTION AND/OR POSITIVE OUTCOME
Recognizing a sharp rise in mental health issues and the increasing isolation of emotionally vulnerable individuals from the workforce, the government steps in, enacting strict regulations on biometric and emotional data collection in workplaces. Companies are now legally obligated to implement employee wellness programs that prioritize genuine mental health support over invasive monitoring. An independent oversight committee is established to audit mood-tracking technologies, ensuring they are used only with employees' informed consent and exclusively for enhancing well-being. These reforms promote a healthier, more supportive work environment, shifting the focus from surveillance to genuine care for employees' mental health.

CREATORS: Noa

Proposition 3:





Case 3: Filling a datasheet

Context

You are tasked to train a Machine Learning model that will serve as a layer of identification in an application: the model should, from a small sample of images from a person, be able to recognize it.

You have found the **dataset MS-Celeb-1M**, which is exactly what you need to pretrain your model! As a responsible software developer, you want to ensure that this dataset is safe to use. To do that, you remember that **datasheets** can help you identify potential ethical issues with a dataset. Unfortunately, there is no datasheet for this dataset. Therefore you decide that you will do it. To help you in this task, we provide you below **with a summary from the original paper describing the dataset** that should be **sufficient to fill the datasheet.**

Note: this summary is for pedagogical purposes only, the source article is the only reference to consider regarding this dataset and the associated research.

Original paper for reference:
Guo, Y., Zhang, L., Hu, Y., He, X., & Gao, J. (2016). MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), Computer Vision – ECCV 2016 (pp. 87–102). Springer International Publishing.
https://doi.org/10.1007/978-3-319-46487-9_6

Summary

The goal of the development work described in the paper is to train Machine Learning models to identify people in images.
For this purpose, the study describes 3 sets of data:

- A celebrity list: it includes 1 million celebrities (information only, no image).
- A dataset to be used for benchmarking (i.e. evaluating the performance of ML models): it includes 1500 celebrities from the list as well as 30K images, mixing manually verified images (2 per celebrity) with randomly selected images as distractors.
- A dataset to be used for model training: it includes only the top 100K celebrities and 100 images per celebrity.

The celebrity list is selected from a knowledge graph called Freebase [Note: see this wikipedia article]. Freebase is a graph made of nodes and links that establish relationships between the nodes. Nodes correspond to topics, and Freebase covers several millions of topics about real-world entities like people, places and things. Each entity is identified by a unique key and associated with rich properties. Celebrities have been selected from Freebase as entities which represent real persons. These entities have been ranked based on the frequency of their occurrence on the web. Only the top one million entities have been kept. These include people with varied professions, nationality, age, and gender: the list includes 2000 different professions, 200 distinct countries/regions, a range of ethnicities and age.

The benchmarking dataset contains Freebase entities and associated images. The celebrities are sampled from the celebrity list such that the dataset mainly focuses on top celebrities (ranked among the top in the occurrence frequency list) while 25% of the celebrities come from tail of the list celebrities (celebrities not mentioned frequently on the web, e.g., from 1 to 10 times in total) to guarantee the measurement coverage over the one-million list. The images have been scraped from the web using multiple variations of a search query used for each celebrity to capture diverse images which are truly about the given celebrity. Around 30 images have been scraped per celebrity. The authors of the study have manually evaluated the images and each celebrity has been associated with 2 images: one selected randomly, the other chosen so that it is the most different from all the other images for this celebrity. Then, these images have been blended with images from other celebrities or ordinary people, resulting in a dataset of 30K images.

The dataset provided by the authors to train Machine Learning models includes both Freebase entities and associated images. It contains the top 100K celebrities from the one-million celebrity list in terms of their web appearance frequency. For each celebrity, around 100 images have been retrieved from the web using popular search engines. This dataset contains around 75% of the celebrities from the benchmarking dataset.

Exercise

1. Fill the datasheet with information from the text provided above: cells with a gray background have already been filled out, you need to complete the cells with a white background.
2. Highlight 2 ethical problems with this dataset

Proposed answer:

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?	*Create a benchmark for face identification, based on images of celebrities.*
Who created this dataset and on behalf of which entity?	*Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, Jianfeng Gao, for Microsoft.*
Who funded the creation of the dataset?	*Microsoft.*
Composition	
What do the instances that comprise the dataset represent?	*Instances represent celebrities, i.e. people known from the general public.*
How many instances are there in total?	*1 million*
Does the dataset contain all possible instances or is it a sample 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? If so, please describe how this representativeness was validated/verified.	*The celebrity list contains only a subset of Freebase relating to real persons, only the top one million entities when ranked by frequency of appearance on the web.* *The benchmarking dataset contains a subset of the celebrity list (1500 celebrities), with a mix of manually selected images and automatically selected distractors.* *The training dataset contains also a subset of the list (100K celebrities), with 100 images automatically retrieved from the web for each celebrity.* *The representativeness of the datasets can be questioned in the following ways:* - *"celebrity" is defined as frequency of appearance on the web, which may not represent all types of celebrity, may misrepresent celebrity (e.g. journalists are present on the web but not celebrities) and under-represent populations for whom web presence is not developed* - *the datasets contain only "celebrities", therefore it is not representative of non-celebrities (their attributes and the type of images available online may differ greatly from the general population)* - *the celebrities in the benchmarking and training datasets have been sampled with strategies focusing on the celebrity ranking but we don't know if the representativeness in terms of attributes (e.g. gender, profession, etc.) has been ensured*
What data does each instance consist of? "Raw" data or features?	*Each instance has two pieces of data:* - *entity key from Freebase linking to personal information including profession, nationality, age, and gender* - *images of faces*
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.	*No.*
Are relationships between individual instances made explicit? If so, please describe how these relationships are made explicit.	*Yes, by their entity links. If two images share the same entity link, it means that the person in the image is the same.*
Is the dataset self-contained, or does it link to or otherwise rely on external resources?	*It links to Freebase entities.*

Does the dataset contain data that might be considered confidential?	*As we don't know where the images come from exactly, nor what information about the entity is present in Freebase, it is possible that data can be considered confidential.*
Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.	*It can, if images of the celebrities are unsafe (e.g. images of intimate life or violence taken by paparazzi).*
Does the dataset relate to people? If not, you may skip the remaining questions in this section.	*Yes.*
Does the dataset identify any subpopulations?	*Yes, subgroups based on profession, nationality, gender and date of birth.*
Is it possible to identify individuals , either directly or indirectly from the dataset?	*Yes, explicitly.*
Does the dataset contain data that might be considered sensitive in any way?	*Yes.* *Faces can be considered biometric data, and they make possible the identification of people (which is the goal of the research described in the paper).* *In addition, some images may have been made publicly available against the will of their authors.*
Collection Process	
How was the data associated with each instance acquired? Was the data directly observable, reported by subjects, or indirectly inferred/derived from other data?	*For the celebrities, data from Freebase has been reused but how it was originally obtained is not described.* *For the images, they have been obtained from web scraping.* *The link between the photos and the Freebase entities was manually labeled.*
What mechanisms or procedures were used to collect the data? If the dataset is a sample from a larger set, what was the sampling strategy?	*Celebrity list: selection of entities which correspond to real persons, ranked by their overall presence on the web as an indicator of celebrity.* *Benchmarking dataset:* - *only the top celebrities with 25% of the celebrities coming from the bottom 90% of the one-million list, 1500 in total* - *2 images per celebrity: one randomly selected, the other by maximizing difference with other images* *Training dataset:* - *only the top 100K celebrities* - *100 images from web scraping*
Who was involved in the data collection process and how were they compensated?	*Unknown, probably the authors of the paper.*
Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances?	*Unknown.*
Were any ethical review processes conducted?	*Not indicated (probably no).*
Does the dataset relate to people? If not, you may skip the remainder of the questions in this section	*Yes.*
Was the collection of the data from the individuals in question directly, or obtained it via third parties or other sources?	*Obtained via third parties (Freebase and web scraping).*
Were the individuals in question notified about the data collection?	*Not indicated (probably no).*
Did the individuals in question consent to the collection and use of their data?	*Not indicated (probably no). It is even possible that the individuals were not aware that the data existed in the first place.*

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?	*Not indicated (probably no).*

2. **Highlight 2 ethical problems with this dataset:**

- Empowerment & Privacy: Consent from the individuals has not been obtained, some people could have their faces in this dataset without them wanting to.
- Fairness: The dataset's focus on public figures, primarily celebrities, risks creating models that perform poorly on diverse populations and may reinforce biases.
- Safety: The dataset can be exploited for harmful applications such as mass surveillance, the tracking of individual persons (e.g. journalists) or to infer sensitive traits for discriminatory or oppressive uses.

Complementary information: the original dataset has been taken down because of the number of ethical concerns it raised. You can find more information here: https://exposing.ai/msceleb/