Discrimination and AI

Naoto Hieda

One with a naive understanding of artificial intelligence (AI) might think that decisions made by AI and computers are neutral compared to humans; therefore, financial advisers, face recognition, and criminal prediction used by police, to name a few, are more trustable when AI is introduced rather than human operation. However, in reality, AI is often as discriminatory and biased as humans since it is trained by data provided by humans, or in other words, AI reflects the history of discrimination. In fact, not only recent AI applications, any computer program reflects human subjectivity. Good Old-Fashioned Artificial Intelligence works based on an algorithm designed by computer scientists and engineers. Even current deep-learning neural networks are heavily biased by hand-tuned hyperparameters, which, for example, determine the learning rate, and most importantly, the neural networks are trained by datasets collected and labeled by humans. In this comprehensive report, real-world problems of biased AI and creative ideas against the discrimination are presented. At the end, I describe my work-in-progress piece that questions the diversity in machine learning datasets.

MNIST dataset mapped to a 2D space based on similarities (t-SNE algorithm) shows different styles of handwritten 7 (https://lvdmaaten.github.io/tsne/)

A dataset for machine learning can contain any data, e.g., texts, images, videos, waveforms and share prices. For supervised learning, it often contains labels or annotations made by humans. A classic example is the MNIST database (Modified National Institute of Standards and Technology database) by Yann LeCun, Corinna Cortes, and Christopher J.C. Burges [1]. An entry in this dataset consists of a 28x28 pixel grayscale image of a handwritten digit, and a digit associated with the image, which serves as the ground truth. Although at a glance, this dataset seems to be neutral and not affected by any bias, handwritten digits are heavily biased by the culture. For example, 7 is sometimes written with a slash in the middle to distinguish from 1;

nevertheless, this is only the case in continental Europe and a few other parts of the world, and therefore, it is not regarded as an international standard. Thus, blindly trusting this dataset to create a “universal” digit recognition system will end up discriminating people with different handwriting conventions as well as minority groups that do not use Arabic numerals.
[1] http://yann.lecun.com/exdb/mnist/

Datasets with higher dimensional data are more troublesome. ImageNet, one of the well-known image dataset, has become a benchmark for image classification since it was released in 2009 containing 14 million images labeled by Mechanical Turk [2]. To encourage computer scientists and engineers to enhance the accuracy of image recognition, ImageNet Large Scale Visual Recognition Challenge (ILSVRC) was founded in 2010. Soon in 2012, the top-5 error [3] dropped from the previous record 26% to 15% by deep convolutional neural network named AlexNet developed by Geoffrey Hinton, Ilya Sutskever, and Alex Krizhevsky [4]. In 2014, the ImageNet classification task by humans has been already outperformed by computers, reported by Andrej Karpathy [5]. However, this fact does not suggest that computers are smarter than humans at any given image classification task. Nevertheless, convolutional neural networks are tuned to solve a specific image recognition task: in this case, recognition of images provided in ImageNet, which contains biases. Karpathy suggests the following issues. First, when an image contains multiple objects, the label depends on the subjectivity of the AI or the human recognizing the image. Although Karpathy noted that humans are better in finding the most salient object, I must mention that it is dependent on individuality as well as neurodiversity. Second, the dataset in fact contains mislabeled data as they are annotated by humans through Mechanical Turk. Third, some of the labels are too fine-grained: for example, 120 out of 10,000 labels are dog species, which most of the humans are not trained to distinguish.
[2] http://www.image-net.org/
[3] The rate at which the predicted label does not match any of the top 5 labels
[4] ImageNet Classification with Deep Convolutional Neural Networks
[5] What I learned from competing against a ConvNet on ImageNet

These data biases, due to culture or to the specificity of data, become problems in real life when implemented as an application. One of the examples is a face tracker trained with a dataset consisting of Caucasian population. When a Nikon camera was directed to take a photo of a woman with a Taiwanese origin, the camera prompted that the person was blinking [6]. Although she was not blinking and simply smiling in front of the camera, the face detector falsely detected her thin eyes as blinking. In another example, a YouTube video “HP computers are racist” criticizes the webcam on an HP laptop, which is supposed to follow a face in front of the computer for automatic panning and zooming [6]. Nevertheless, the program was only following the face of a Caucasian woman and could not detect an Afro-American person’s face. These two cases show that lack of data and testing leads to discrimination. A case of Google Arts & Culture shows another type of cultural problems [7]. The company released an app that analyzes the user’s portrait and finds a painting that resembles the portrait. Although it was perhaps intended to encourage Internet users to be familiarized with art history, as Mashable reported, when a portrait of a Latin American is input to the app, the result showed either a

European or Asian man. This result is because of the eurocentric art history lacking data of Latin American artworks. An installation “Uncanny Mirror” by Mario Klingemann observes visitors’ faces and transforms a face in real-time based on what the installation has been observing [8]. For example, when it was exhibited at Media City Biennale Seoul (Korea, 2018), the installation was biasing itself to generate Korean faces as most of the visitors are local. Even though a non-Korean person is standing in front of the installation, the “mirror” morphed the person to have a Korean face. This installation can be seen as a criticism of bias in datasets by enhancing the bias although some visitors may feel discriminated to see the face completely transformed.

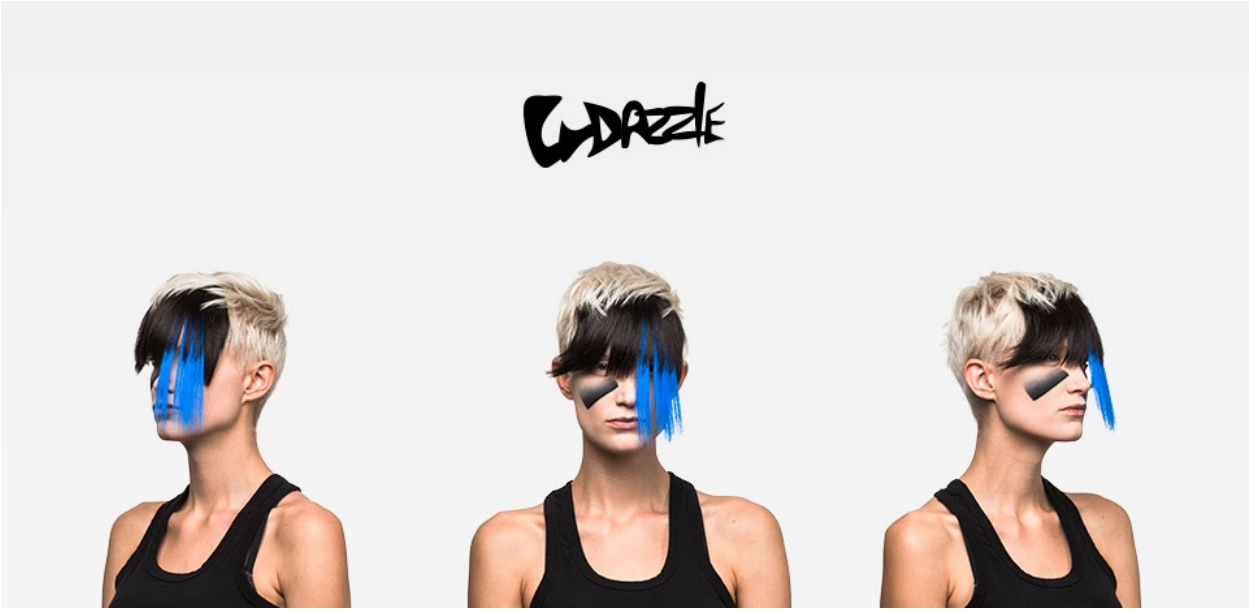
[6] http://content.time.com/time/business/article/0,8599,1954643,00.html
[7] https://mashable.com/2018/01/16/google-arts-culture-app-race-problem-racist/?europe=true
[8] https://www.dazeddigital.com/art-photography/gallery/26460/3/ai-more-than-human

Discrimination by AI not only makes people disgusted but also involve people in legal issues. From 2012 to 2018, New Orleans police had a partnership with Palantir Technologies, a data-mining company, providing the company police reports to create a predictive policing system [9, 10]. The partnership became controversial since the city and company collaborated without public notice, and what is worse, the provided data was contaminated. The United States Department of Justice revealed that the New Orleans police department violated constitutional and federal law, and the corrupted, biased records were used to train the predictive system. As a result, the system tends to target black residents, racial minorities, non-native English speakers, and LGBTQ individuals. In Germany, SCHUFA has been regarded as a reliable source to provide trust information of individuals; however, the company is privately owned. Thus, their rating system is closed and no one outside the company can access their database. OpenSCHUFA is an initiative to collect SCHUFA scores from volunteers to crack the algorithm behind SCHUFA [11]. OpenSCHUFA reported that the scores are biased; for example, if there is a person with the same first and last names, one’s SCHUFA scores are likely to be influenced by the other. In China, it is not a secret that the Chinese government is working with private tech startups to improve their surveillance technology by providing access to public surveillance cameras [12]. For example, when one breaks a law, one’s photo, name and public ID are shown on a public display as a punishment and enforcing the power of the communist party.

[9] Palantir has secretly been using New Orleans to test its predictive policing technology
[10] Police across the US are training crime-predicting AIs on falsified data
[11] https://www.startnext.com/openschufa
[12] Inside China’s Dystopian Dreams: A.I., Shame and Lots of Cameras

Finally, the threat of AI is already reaching our hands through smartphones and social media. Most of the social media switched from the chronological order to the curated order, which is an opportunity for the companies to manipulate what contents to show. In January 2012, Facebook intentionally modified the curation algorithm of around 700,000 users; half of them see positive posts more often and the others see negative posts more [13]. By examining their status updates, Facebook revealed that users’ mood can be manipulated by what contents to show.

Despite this misconduct, later Facebook and Cambridge Analytica caused a huge scandal using personal data for targeted advertisement for political campaigns.
[13] 9 answers about Facebook's creepy emotional-manipulation experiment



CV Dazzle by Adam Harvey (https://ahprojects.com/cvdazzle/)

Given these examples of discriminatory AIs, how can we challenge these problems? Search engines such as DuckDuckGo claim that the company does not store personal data. Mozilla is not only developing Firefox browser but also raising awareness on online security through blog articles. Beyond protecting private spheres by blocking data collection, artists and researchers suggest creative hacks to resist AI. Adversarial attacks are ways to mislead machine learning algorithms to generate “undesirable” outputs. CV Dazzle by Adam Harvey is an experimental project to avoid face tracking by hair styling and makeup [14]. At a glance the makeups appear eccentric; however, the project carefully designed to reverse-engineer Viola-Jones face detection algorithm by adding color contrast on the face. Since Viola-Jones algorithm is used as the first step of other common face tracking algorithms, this makeup is effective for higher-level face tracking such as landmark detection. One pixel attack is a project that uses an iterative algorithm to crack a deep convolutional neural network trained on Cifar10 dataset by modifying a single pixel in the input image [15]. Anish Athalye, Logan Engstrom, Andrew Ilyas and Kevin Kwok created a real-life example, namely, a turtle toy with a special pattern that is recognized as a rifle [16]. Another approach against AI is to feed false data to the companies collecting personal data so that the dataset becomes unusable. Privacy Possum is a browser extension that does not block information trackers but rather sends false data, making personal information less profitable [17].
[14] https://ahprojects.com/cvdazzle/
[15] https://github.com/Hyperparticle/one-pixel-attack-keras
[16] https://www.labsix.org/physical-objects-that-fool-neural-nets/

[17] https://github.com/cowlicks/privacypossum

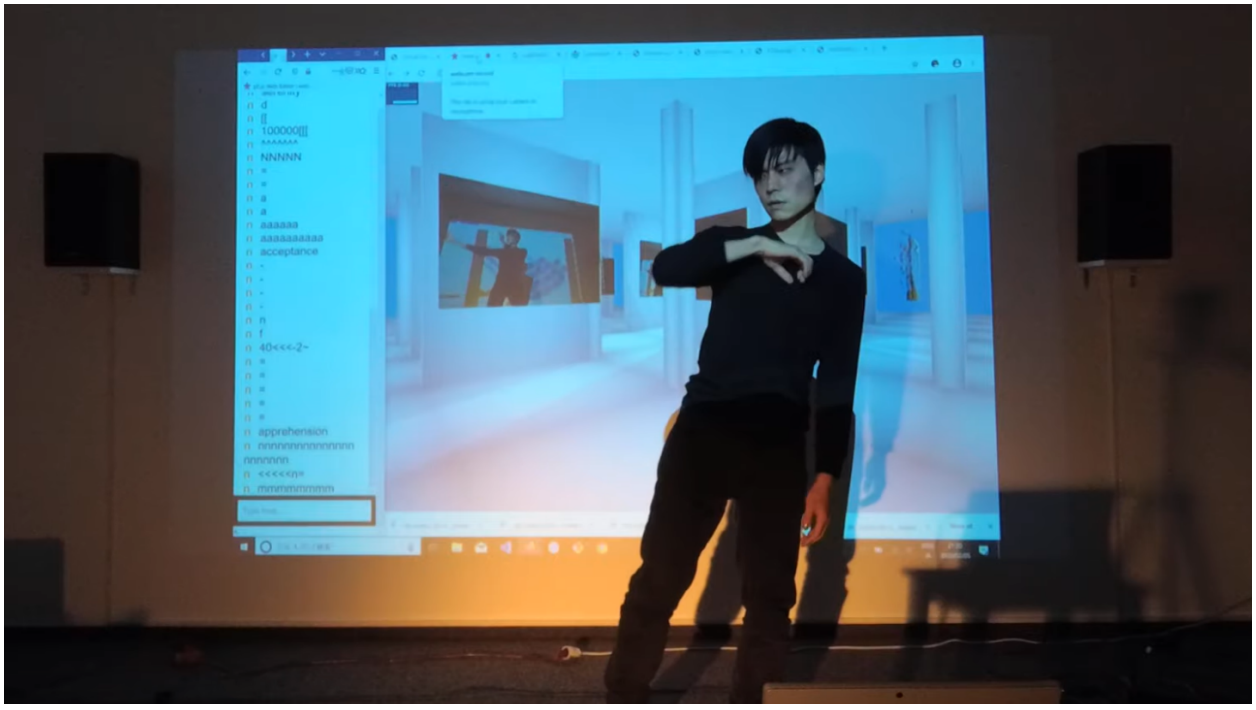
Caroline Sinders' Feminist Data Set project questions the process of data collection through a series of workshops [18]. The project does not necessarily collect data to train a specific machine learning algorithm; instead of creating a big data ready for analysis, people are asked to carefully gather information about feminism and minority groups. Through the act of collecting texts, ideas, poems and lyrics, she poses questions against sexist AI and online harassment and gives a chance for the participants to rethink data literacy.
[18] https://carolinesinders.com/work#/feminist-data-set/

At last, an ironic example of hacking AI is from Justin Bieber [19]. When he released a single "Yummy," he asked the fans on Instagram an absurd request to improve the sales. The request was to keep his single looping on Spotify at low volume while sleeping in order to augment the play count. I assume that Spotify has an algorithm that rejects such play counts as outliers; however, what can be learned from this example is that hacking AI is not only a matter for artists and nerds but every online user can be interested and be involved without noticing [20]. Also, this case poses another interesting question: are we producing contents for humans or for machines and AIs?
[19] https://www.rollingstone.com/music/music-news/justin-bieber-spotify-yummy-936194/
[20] In fact, this type of hack exists since search engine optimization (SEO) is invented.

My proposal against discriminatory AI is to create an autistic dataset; inspired by Caroline Sinders' project but instead of a workshop format, I question the modality of AI through live-coding and performativity of the body. While common datasets are based on texts and images, my research questions is this choice of modality that potentially discriminates people; for example, blind and visually impaired people may value visual information in a different manner, and people with synesthesia or autism who have strong visual thinking see the world in a specific way beyond texts, images and emotions. As an Asperger's, I am working on a hybrid live-code-dance performance that travels through modalities such as images, emotions as words, sounds and movements, simultaneously exhibited on a custom online platform. I have a practice with which I associate my emotion evoked through mediation to sketching and improvisational dance. Based on this idea, I went through an emotion vocabulary on the stage (boredom, annoyance, interest, serenity, acceptance, apprehension, distraction, and pensiveness, taken from Plutchik's Wheel of Emotions [21]), recording my movements with a webcam in webm format. To relate the movement to a figurative representation, I use Google image search to find an object and projected it as a backdrop. Despite technically a simple way to pick an image, it is effective since the audience can easily relate it to their experience of using a common search engine. The emotion terms and explanations of the performance are typed into a custom-built chat interface to be shown to the audience instead of a spoken language. Although I am performing in the same space, the words are uploaded to the Internet, shown on the projection as if the stage is isolated from the audience to create a personal sphere. Simultaneously, each byte of an input text is translated into sound in a live-coding manner based on p-code syntax developed by Yosuke Hayashi [22]. Therefore, the chat interface is

used not only to communicate with the audience but also to create sounds on the fly to help myself evoking emotions. Every recording of the movements are uploaded on a virtual space made with three.js; the platform resembles a white gallery space, and the exhibited artworks are the recordings. The concept of the virtual gallery is to blur the boundary between archival and exhibition. As every data in the dataset becomes visible as exhibited, the importance and authenticity of the data is revealed. At the same time, the ephemerality of digital data and physical movements makes people wonder if the value exists in the act or in the code (in other words, the algorithm) itself. Through the multi-layered performance, I express the struggle of autism as well as how we can bring diversity in machine learning, artificial intelligence and digital culture as a whole. The full video and virtual exhibition platform can be accessed from the link below [23].

[21] https://en.wikipedia.org/wiki/Emotion_classification#Plutchik's_wheel_of_emotions
[22] https://github.com/p-code-magazine/p-code
[23] https://naotohieda.com/blog/virtual-exhibition-003/



Performance by the author.

Acknowledgements

The references in this article is a compilation of talks by Caroline Sinders, Janus Rose (Machine Learning Literacy workshop at the School for Poetic Computation, 2018), Katharine Jarmul (Transmediale, 2020), and the author (Chaos Communication Congress, 2019). Technical backgrounds are cited from “Machine Learning for Artists” by Gene Kogan (Japanese translation by kynd and the author) [24].

[24] https://ml4a.github.io/ml4a/