Real Estate
Operations Manager
Assistant Engineering Teacher
AI Model
Electrician Registered Nurse
Assistant Customer Services
Delivery

3. Practice responsible dataset development

Ensure that dataset development is conducted responsibly, with standard checks and balances in place for creating new datasets as well as adapting existing ones. The creation and implementation of such practices requires businesses to be intentional about gathering inclusive data and asking important questions around who is benefiting from the data collected.

BUSINESS BENEFITS:

- Mitigate risk
- Have a superior value proposition

PLAYERS INVOLVED:

- Product managers
- Engineers & data scientists developing or sourcing datasets
- Employees & contractors labelling data

Elements:

□ When generating and collecting primary data, adopt an approach that integrates social science methods with more technical data science methods.

- Where possible, supplement quantitative data with qualitative insights to add depth to the data being collected. Collaborate with social scientists in relevant disciplines.
- Consider participatory data collection methods that empower communities whose data is being collected.
- Integrate non-binary gender data. Explore solutions around using string variables for fluidity of representation, measuring gender through opt-in or self-representation options, etc.
- Disaggregate data by gender, race, etc. Ensure that the data points are well balanced across communities, and that identities are accurately represented.

□ Place checks on labelling practices for employees and/or contractors.

- Provide clear, accurate annotation instructions to labelers at the beginning of each project. Ensure there are clear directions on appropriate classifiers and adjectives.
- Require that labellers undergo DEI / inclusive language and implicit bias trainings. This is particularly crucial for those working with images of people and words that are subject to interpretation—such as identifying offensive / hate speech.
- Implement inter-rater reliability as an added measure of validity.[1]
- Explore self-labelling practices, particularly for demographic / employment information.
- Consider having humans in the loop when relying on automated labeling.
- As is possible, engage a diverse team of labellers to annotate training data while ensuring fair working conditions and wages.

Tools:

- **Do Ask, Do Tell** (Stonewall Org) – guide making the case for and outlining best practices for integrating inclusive gender data
- See Play 6 for more information on data justice and participatory research methods

- **EFL Glossary of Key Terms** (EGAL)
- Use case studies such as **ImageNet Roulette** to illustrate importance.[6]

Elements:	Tools:
□ Document the provenance and creation of machine learning datasets, and make this information available to all users. Ensure that data was collected for a purpose that matches the intended use of the dataset. • Collect information about how, why, and from where data points were collected. Disaggregate by gender, ethnicity, race, education level, socioeconomic status, etc. to encourage the formation of a well-balanced dataset. • Maintain a record of how and by whom data points were annotated—including geographic, cultural, and other demographic backgrounds. • Include in documentation the possible over-/under-representation of identities in dataset.	• **Datasheets for datasets** • **Data Nutrition Label** • **Writing inclusive documentation** (Google)
□ If sourcing existing datasets, establish standards for sharing data across sectors and between multiple public and private entities.[2] • Explore collaborative structures like corporate data cooperatives / pooling and trusted data intermediaries.[3] • In identifying datasets, prioritize datasets built with equity and inclusion in mind. Also, assess existing datasets to check for over-/under-representation of certain identities, or underlying inequities that reflect reality but are ultimately problematic. • When working with image datasets or word embeddings for natural language processing, use technical tools and algorithms developed to tackle bias / stereotypes in data.[45] • Require documentation of the provenance and creation of the dataset being acquired. Ensure the data was collected for a purpose that matches the intended use of the dataset. • Establish a feedback loop, ensuring that errors and instances of bias are reported in real time.	• **Data Risk Checker wiki** (Responsible Data Forum)
□ Address privacy concerns that may arise when assessing an existing dataset for fitness of use. • Specify clear guidelines around what elements of the dataset are appropriate to analyse for fitness of use. Document and distribute these guidelines, and ensure that they are part of the training process for employees assessing datasets. • Combine technical mechanisms (such as audit logs) with corporate governance mechanisms to protect against unintended exploration.	• **A Risk-Based Approach to Privacy** (Centre for Information Policy Leadership) • **Privacy Analytic** (Mulligan et al.)
□ For all datasets used, maintain as living resources. • Audit datasets periodically to ensure they incorporate recent developments in DEI language, social classifications, and quality standards.	

Examples and leaders:

Samasource: Samasource employs workers from underserved areas and pays a living wage to classify training data for AI and ML applications. They work with over 25% of Fortune 50 companies. Inherent to its social mission, Samasource hires labelers with low socioeconomic backgrounds, tapping into communities in Africa that are typically excluded from the digital economy and ensuring that they have access to favorable, dignified working conditions, living wages, and upskilling opportunities to build their career. The majority of this workforce consists of Black Africans, and 50% are women. Samasource recognizes that using platforms such as MTurk can result in data that is labeled with little to no direction or understanding of bias. Samasource workers are not only equipped with basic background knowledge around AI and vertical-specific trainings, but also provided accurate, clear annotation instructions at the beginning of each project. For more technical forms of bias related to over-/under-representation of classes or identities in data, Samasource has developed an AI bias detection solution to gauge potential over-/under-representation of classes or identities. The findings are used to address the distribution of classes in their datasets.[7]

Gender Shades: In order to combat the lack of diversity (in terms of gender and skin type) widely used datasets, Joy Buolomwini and her team at MIT Media Lab developed a new benchmark image dataset. This dataset contains images of 1270 individuals' faces from three African countries and three European countries. The photos are of public figures whose images are available under non-restrictive licensing, intentionally selected to avoid privacy violations. The creators of the dataset carefully detail the process through which they chose their intersectional labels, acknowledging shortcomings and constraints under which they operated. A test of gender classification algorithms trained on this inclusive benchmark dataset allowed for an intersectional analysis of how said algorithms performed, increasing transparency.[8]

Background:

Data is a key determinant of an algorithm's output. Large amounts of data points are acquired, "cleaned",[9] and labelled to prepare training datasets for algorithms. Algorithms then learn from, test their learnings against, and perform their operations based on these datasets. Poor quality and/or low quantity of data is often cited as one of the top issues faced by the data science community, according to a survey conducted by Figure Eight.[10]

There are various ways that bias can enter a dataset and result in biased AI (where AI results in discriminatory or inaccurate predictions for certain subsets of the population). Bias can enter as a result of how data points are generated, collected, or labelled, as well as how a dataset is constructed. For a breakdown of how and why this happens—accompanied by real world examples—**see our Bias in AI Map and pages 23 to 29 of our Mitigating Bias in AI Playbook.**

While the datasets most fraught with bias are the ones containing demographic data or categorizing individuals or groups, bias in datasets is not limited to data explicitly about people. Datasets can still be used in contexts such that they result in inaccurate or discriminatory outputs for different subsets of users. For instance, researchers at Facebook found that datasets used to train widely used object recognition systems had a very skewed geographic distribution—almost all the photos came from Europe and North America.[11] As a result, these systems perform worse for users in Africa and Asia—and were notably found to perform ten to twenty percent worse for the wealthiest households than for the least wealthy households.

There is a clear need for businesses to be intentional about gathering inclusive data and asking important questions around who benefits from the data being collected. This is because dataset issues cannot always be resolved through testing. Although algorithm performance is evaluated on "test" datasets, these are usually random sub-samples of the original training datasets. The assumption is that the training data is an unbiased representation of society. But as Catherine D'Ignazio (Assistant Professor at MIT) puts it: "data is never this raw, truthful input and never neutral. It has been collected in certain ways by certain institutions for certain reasons." While the testing phase may be sufficient to flag and correct technical inaccuracies, it may not always serve as a solution to the problems presented by biased data.

While the elements listed previously are important to implement, there are certain pitfalls to be aware of. In particular, pitfalls can occur when implemented without an understanding of broader societal contexts. Here are some challenges, risks, and ethical dilemmas to be aware of.

1. Challenges and risks of collecting and using detailed gender data:
Collecting accurate gender information ensures the inclusion of gender minorities in the development of goals/ priorities, and helps evaluate the effectiveness of AI programs. Transgender / gender-nonconforming people typically represent statistical outliers in most datasets (0.3%, according to a UCLA study on data collection and gender identity).[12] However, if folks identify as LGBTQ and provide information on their gender and sexual orientation, this can be risky or dangerous—even if the data is anonymized, and depending on the context. In certain countries, e.g., Ethiopia, it is illegal to identify as LGBTQ.[13] The fact that these sample sizes are small amplifies risk of identification and potential harm.

There are also challenges to collecting information on gender. Some are due to algorithmic constraints, as outlined in **the playbook**. Historical studies of non-binary gender reveal a host of varying vocabulary used to identify types of gender, making even elaborate drop down menus of choices inadequate.[14] Even when accuracy is ensured, algorithmically inferring an individual's gender or social orientation from available data can result in discrimination (sometimes violent) or even imprisonment, based on the individual's social context / community.[15]

Given the complexities of recording and using gender / sexual orientation data, decisions on whether or not to do so should be made on a case-by-case basis. Dataset creators should consider why gender classification is being used—what benefits it brings to the system being built, etc. Where unnecessary to the system, embracing gender ambiguity might be the right move instead. Where necessary, ensuring that information is collected responsibly and in an inclusive manner is critical.

2. Potential ethical trade-offs of collecting more data to improve representation in datasets:
No doubt, collecting more data from underrepresented communities improves the ability of algorithms to produce accurate results for said communities. However, it also brings up complex issues of data ownership and power dynamics. For instance, when a group of researchers were attempting to improve the accuracy of their facial recognition AI, they recognized the lack of images documenting faces on transgender individuals before and after HRT.[16] They chose to rectify this by compiling data from YouTube videos of individuals documenting their transition process—without their consent. Even though the dataset was never shared for commercial purposes, this sort of biometric collection from a community that is already marginalized and discriminated against has major ethical implications.

When drawing data en masse from search engines, data collectors often don't account for the representational harms of classifying images of people without their consent or participation.[17] The above example also illustrates the fact that individuals from certain underrepresented communities may consciously choose not to provide complete / accurate demographic data about themselves owing to safety concerns. In her book Algorithms of Oppression, Safiya Umoje Noble anecdotally highlights a data gap on the business review site, Yelp: Black people are less likely to 'check in' and let people know where they are as they are already used to being over-tracked and overpoliced.[18] In the U.S., immigrant communities—particularly undocumented immigrants—and indigenous communities have also been at the forefront of resisting harms caused by data abstraction. The right of a nation to govern the collection, ownership, and application of data about its citizens through census and population counts is contested—and these concerns are evident worldwide.

The probability of consenting to data collection varies across groups, and is linked to specific historical, cultural, and political contexts that the tech industry often overlooks.[19] While collecting more data remains the most mainstream solution for improving datasets, alternative models do exist. These include opt-in methods of collecting data as opposed to opt-out, and even password protecting datasets so potential users / teams have to provide information to the owner regarding what they're looking to use it for, to build in an element of accountability. However, little research exists around these models, and they are likely to be very resource intensive.[20]

3. Lack of ethical standards for sharing data:
As the digital economy and the number of players in it continue to expand, the sheer amount of data available through existing datasets has made data sharing an essential business practice.[21] Repurposing existing datasets allows more entities to draw deeper insights from data, and allows smaller groups to innovate with data they may not otherwise have the means to acquire. Where necessary, data cleaning and re-labeling

techniques are used to tailor these datasets to specific uses, but the sharing process itself remains unchecked, raising several concerns.

- **Privacy concerns:** Even though identifiable personal details may be removed before datasets are given to a third party, algorithms with access to auxiliary datasets can still draw unwelcome conclusions. For instance, analysts using the anonymized dataset of the NYC Taxi Commission's trip records were able to determine the likely religion of certain cab drivers.[22] They could also de-anonymize the names of drivers using variables like medallion numbers, which could then be correlated with private details like income.[23]
- **Safety / security concerns:** The potential harm caused by making large amounts of data public is not always evident from the offset. In 2017, fitness app Strava released a detailed data visualization map showing all the activity tracked by its users. However, this also revealed sensitive information about the location and staffing of military bases and spy outposts. The heatmap could have had unintended consequences on a subset of Strava's users across the globe: military personnel on active service.[24]
- **Ethical concerns:** These processes complicate the obtainment and maintenance of consent. Consent—whether "informed" in a medical context or end-user license agreements for internet service—typically occurs within the context in which data is initially collected. When used outside these contexts, the gap between data analytics and the tools used to protect subjects widens. For instance, when the UK National Health Service (NHS) planned to centralize and sell citizens' medical data to health management companies and other private partners, it provided subjects with only a short opt-out window.[25] This was dubbed a "smash and grab" scheme, not an example of well-executed open data policy.

ADDITIONAL READING:

- Data Genesis Program (AI Now)

This is part of **Mitigating Bias in AI: An Equity Fluent Leadership Playbook** *of the Berkeley Haas Center for Equity, Gender and Leadership. It was written by Ishita Rustagi and Genevieve Smith, with input from Nitin Kohli.*

Endnotes

1 “Inter-rater reliability is the extent to which two or more raters (or observers, coders, examiners) agree. It addresses the issue of consistency of the implementation of a rating system.” Source: https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-79948-3_1203#:~:text=Definition,a%20number%20of%20different%20statistics.

2 Data Sharing Within Cross Sector Collaborations. (2018, August 14). Retrieved from https://buildhealthchallenge.org/resources/report-data-sharing-within-cross-sector-collaborations/

3 Ethics of Data Sharing – Accenture. (n.d.). Retrieved from https://www.accenture.com/t20161110t001618z__w__/us-en/_acnmedia/pdf-35/accenture-the-ethics-of-data-sharing.pdf

4 Amini, A., Soleimany, A. P., Schwarting, W., Bhatia, S. N., & Rus, D. (2019). Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure. Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. doi:10.1145/3306618.3314243

5 Bolukbasi, T., Chang, K., Zou, J.Y., Saligrama, V., & Kalai, A. (2016). Quantifying and Reducing Stereotypes in Word Embeddings. ArXiv, abs/1606.06121.

6 Excavating AI. (n.d.). Retrieved from https://www.excavating.ai/

7 Gadonniex, H. (August 31, 2020). Personal interview.

8 Joy Buolamwini | Timnit Gebru /. (n.d.). Gender Shades. Retrieved from http://gendershades.org/

9 Data cleaning identifies and removes incomplete, inaccurate, or irrelevant data. This involves ensuring that a large number of unique data points are included to give an algorithm enough information to learn from and ensuring that each data point contains information (“features”) that an algorithm can parse through and use to support decision-making. https://blog.aimultiple.com/data-cleaning/

10 Data Scientist Report 2018. (n.d.). Retrieved from https://www.datasciencetech.institute/wp-content/uploads/2018/08/Data-Scientist-Report.pdf

11 A new way to assess AI bias in object-recognition systems. (n.d.). Retrieved from https://ai.facebook.com/blog/new-way-to-assess-ai-bias-in-object-recognition-systems/

12 Data Collection Methods for Sexual Orientation and Gender Identity. (2020, April 14). Retrieved from https://williamsinstitute.law.ucla.edu/publications/data-collection-sogi/

13 Map of countries that criminalize LGBT people. Human Dignity Trust. Retrieved on August 19, 2020 from https://www.humandignitytrust.org/lgbt-the-law/map-of-criminalisation/.

14 Kanarinka, & Kanarinka. (2016, June 03). A Primer on Non-Binary Gender and Big Data. Retrieved from https://civic.mit.edu/2016/06/03/a-primer-on-non-binary-gender-and-big-data/

15 View of Gaydar: Facebook friendships expose sexual orientation: First Monday. (n.d.). Retrieved from https://firstmonday.org/ojs/index.php/fm/article/view/2611/2302

16 Vincent, J. (2017, August 22). Transgender YouTubers had their videos grabbed to train facial recognition software. Retrieved from https://www.theverge.com/2017/8/22/16180080/transgender-youtubers-ai-facial-recognition-dataset

17 Excavating AI. (n.d.). Retrieved from https://www.excavating.ai/

18 Umoja Noble, S. Algorithms of Oppression.

19 (n.d.). Retrieved from https://ainowinstitute.org/AI_Now_2019_Report.html

20 West, S. (2020, February 12) Personal Interview.

21 Ethics of Data Sharing – Accenture. (n.d.). Retrieved from https://www.accenture.com/t20161110t001618z__w__/us-en/_acnmedia/pdf-35/accenture-the-ethics-of-data-sharing.pdf

22 Franceschi-Bicchierai L (2015) Finding Muslim NYC Cabbies in Trip Data. Mashable. Available from: http://mashable.com/2015/01/28/redditor-muslim-cab-drivers/#0_uMsT8dnPqP

23 Hern, A. (2014, June 27). New York taxi details can be extracted from anonymised data, researchers say. Retrieved from https://www.theguardian.com/technology/2014/jun/27/new-york-taxi-details-anonymised-data-researchers-warn

24 Hern, A. (2018, January 28). Fitness tracking app Strava gives away location of secret US army bases. Retrieved from https://www.theguardian.com/world/2018/jan/28/fitness-tracking-app-gives-away-location-of-secret-us-army-bases

25 Anderson, R., Says:, S., Says:, F. R., Says:, R. C., Says:, R. W., Says:, J. W., . . . Says:, K. T. (2014, January 08). Retrieved from https://www.lightbluetouchpaper.org/2014/01/08/opting-out-of-the-latest-nhs-data-grab/