Play 4. Establish policies and practices that enable responsible algorithm development.

Build practices that check for and actively mitigate bias into every stage of the algorithm development process. This involves equipping teams with ethical frameworks that allow them to prioritize equity while defining their algorithms' objectives, ensuring datasets used are responsibly developed and labeled, and ensuring variables do not disadvantage certain communities.

PLAYERS INVOLVED:

- AI governance & responsible AI leads
- Product managers
- Data scientists & engineers developing AI

BUSINESS BENEFITS:

- Mitigate risk
- Have a superior value proposition

Elements:

□ Collectively establish the algorithm's purpose, keeping in mind potential fairness tradeoffs and ethical considerations when defining its objective.

- Ensure that the algorithm's objective function prioritizes equity as well as efficiency.
- Facilitate conversations amongst development teams, product managers, and business leaders around the guiding principles the algorithm should abide by upon deployment.

□ Ensure that datasets and proxies chosen do not disadvantage certain identities.

- Ensure that the training and testing datasets are responsibly developed or sourced.
- Assess proxies to identify labels that directly identify protected classes (e.g. gender, race) as well as those that indirectly identify them (e.g. hormone levels, zip codes).
- Assess how historical biases could be reflected in features of the data or proxies selected (integrating assessments from social scientists / subject matter experts).
- Integrate tools and implement processes to review and reorganize feature importance in the model.

□ Document the provenance and development of the AI system – including data sources and variables.

- Record information about the team(s) developing the model, the inputs, parameters and constraints the model operates under, intended use, and metrics against which performance is evaluated.
- Release bias impact statements evaluating the algorithm's potential harmful effects as well as their severity.[1]
- Where relevant, include descriptions of how long personal information is stored, what information is available to consumers, etc.
- Encourage communication between teams developing the models and those selling them around appropriate use cases.

Tools:

- **Fairness Analytic** (Mulligan et al.)[6]
- **Ethics canvas**
- **EthicalOS Toolkit**, with a focus on Risk Zone 4: Machine Ethics and Algorithmic Biases

- See Play 3 for tools and resources around responsible dataset development and documentation.
- **Fairness Analytic** (Mulligan et al.)[7] with a focus on "Action" and "Consequences of unfairness"
- **AI Fairness 360 Toolkit** (IBM)
- **What-If Tool** (Google)
- **fairlean.py** (Microsoft)

- **Model Cards for Model Reporting** (Mitchell et al.)[8]
- **AI Factsheets** (IBM)
- **Algorithmic Impact Assessments** (The AI Now Institute)[9]
- **IBM Watson OpenScale** (IBM)

Elements:

□ Integrate human-in-the-loop processes.

- Have development teams evaluate when and how an AI system should seek human input during critical situations, and how a system controlled by AI should transfer information to a human in a meaningful, intelligible manner.
- Ensure transparency around where and how human assessments are integrated, and protect consumer privacy.
- Familiarize human operators on how the AI works, the data powering it, and validity of outputs.

□ Where possible, engage communities impacted by the AI systems in their development and deployment.

- Use community-based system dynamics (visual tools to tease out the complex and dynamic nature of AI systems) to provide stakeholders with a transparent view of the variables and theories at play.[2]
- Ask questions (such as those in the Fairness Analytic[3]) to stakeholders to gather their perspectives and concerns on the prospective AI system, particularly related to fairness, privacy, security, and accountability. Having diverse stakeholders discuss these topics in groups can enrich the conversation and surface nuances.
- View stakeholders as informants in the AI development process and incorporate revisions, critiques, and contestations around the impact of the products and decisions provided through their evaluations.[4]
- Where relevant, establish a robust feedback loop for users to report performance issues. In systems with no way to opt out, make an "appeal" process available for individuals to request human review.
- Where possible, engage communities impacted by the AI systems in their development and deployment.

□ Conduct internal and external audits on the AI system.

- Conduct regular internal audits that review input data and output decisions and identify remedial actions needed before deployment / after operationalization.
- Supplement with external auditor(s) reviewing the ML development process and documentation. Ensure proper safeguards are in place and check against predefined metrics that capture values cared about. Give auditors access to the documentation tools used for the data and algorithm.
- Incorporate considerations of use cases for the AI system. If open-sourcing, make sure to plot out worst-case scenarios, and where possible, build contractual terms for use.[5]

Tools:

- **Community Based System Dynamics**[10]
- Fairness Analytic (Mulligan et al.)[11]

- Aequitas (Center for Data Science and Public Policy at University of Chicago)

Case study on effective community engagement:

Predictive systems, using machine learning (ML), can help determine which pneumonia patients have a high risk of death based on individual characteristics and therefore may need more aggressive treatment.[12] One such system was built by Rich Caruana in the mid-90's (then a graduate student at Carnegie Mellon University, currently at Microsoft research).[13]

However, Caruana was skeptical of deploying this model in practice to use on patients. Another individual had developed a highly-interpretable model (using a technique known as rule-based learning) that was classifying patients with asthma as having a lower chance of dying from pneumonia, which seemed counterintuitive because people with asthma have a higher risk of developing pneumonia. This model seemed potentially inaccurate with huge implications for people's lives. It was conceivable that his neural net may have learned this relationship as well, with little-to-no way for him to be sure of at the time.

The team spoke with doctors – the domain experts in this case, with first-hand knowledge of how to assess a patient's risk of death. The team found the doctors believed the aforementioned model's findings couldn't be medically correct, but that it could be finding a real pattern in the data. Asthmatics may be more likely to seek healthcare sooner when presented with difficulties breathing, thereby allowing the pneumonia to be treated quicker. That is, whether or not a patient had asthma was acting as a proxy variable for time to care. Concluding that his neural net may have learnt this obtuse relationship too, Caruana advocated for using a different model altogether that could be better scrutinized. Through substantive discussions, Caruana was effectively able to interact with the community of interest (here, doctors) to ensure that the model was behaving reasonably.

Effective community engagement in this healthcare setting shed light on new scientific discoveries as well. Recently, Caruana approached the same prediction problem using a more interpretable ML approach, and discovered something that doctors found curious.[14] While doctors typically view a blood urea nitrogen (BUN) level of around 50 or higher as being worrisome for pneumonia patients, the model learned that the risk of death was actually starting to rise when BUN was around 40. This prompted doctors to re-evaluate the most appropriate BUN threshold that signals an increase in the likelihood of pneumonia-related death. This demonstrates how model developers can partner with stakeholders to ensure AI models work responsibly, while allowing impacted communities to leverage new insights learned from the models.

Background:

As our **bias in AI map** illustrates, bias can creep in at every stage of the development process. Teams working on algorithms and AI models must address how each action they take could lead to disparate impacts for different identities / communities.

It is imperative to carefully define the purpose of the algorithm before spending time developing any code. The problem that the algorithm tries to solve must be framed in terms of a mathematical objective function. Unfortunately, this often requires trade-offs between accuracy and fairness, and the typical objective of minimizing error rates on test data tends to overlook the following issues:

- ***The cost of all errors may not be identical.*** The effects of false positives and false negatives are not experienced homogeneously across members of society. For instance, when setting up fraud detection in credit card purchases, the impact of flagging a grocery purchase as fraudulent and halting the transaction may vary from individual to individual. The impact would depend on characteristics such as having additional means to make the purchase (e.g. having sufficient cash on hand, a debit card, or a different credit card) and the immediate need of the items, to name a few.
- ***If the inputs are biased, an algorithm (even if technically accurate) will make biased predictions.*** Solely specifying an objective function in terms of predictive error rates, for instance, may end up perpetuating existing biases, as seen in the COMPAS model.

Prior to developing an AI system, it is important for teams to envision the system and its role in society, while also scrutinizing potential fairness-related harms to different stakeholder groups. Reflective questions such as "what effect do we want the algorithm to have?" and "what are the side effects of working towards this particular objective?" are important first questions to ask before iterating and delving deeper into the nuances of unpacking desirable versus undesirable behavior. However, as meticulous as these considerations may be, they are not sufficient.

Biases present in data and other essential inputs for the ML model can persist despite a well-defined objective. Play 3 outlines the importance and challenges of ensuring the datasets used to train and test the algorithm are responsibly constructed and/or sourced. The proxy variables used to capture abstract, unquantifiable features of data are just as important to assess, as they may advertently or inadvertently lead the algorithm to make undesirable correlations. For instance, when a team at Amazon developed an AI system that screened resumes to recommend top candidates, they found that a majority of the candidates being recommended were male. Trained on resumes of past candidates who had been successful (primarily white and male), the algorithm had interpreted gender as a measure of employment success (despite not being included explicitly as a feature), and gone on to downgrade female candidates' resumes.[15]

From a legal standpoint, algorithmic decision-making processes are subject to the same US anti-discrimination laws as human decision-making.[16] Variables that indicate protected classes such as race, national origin, and gender are assumed to automatically code bias into algorithms, and are illegal to use as proxies in industries like housing, employment, and credit. However, simply removing explicit race and gender labels is not an adequate solution – plenty of other variables may indicate protected attributes. Common examples include zip code as an indicator of race, hormone levels in a medical record as an indicator of gender, etc.[17] If feature importance is not reviewed to identify these implicit links, the algorithm could continue to deduce and amplify historical inequities. In fact, gender and race-blindness might make it harder to identify which variables are leading the algorithm to make problematic inferences. Some researchers argue that purposely introducing protected class characteristics into a model makes it easier for developers to track and mitigate any manifested biases linked to them.[18]

Ultimately, humans and algorithms must work together to mitigate bias. "Human in the loop" systems make recommendations or provide options that human operators double check or choose from. That is, the AI systems complement rather than replace human judgment.[19] Designers can evaluate when and how an AI system should seek human input.[20] Stakeholders can be made aware of where human assessments and deliberations are integrated in order to build transparency around how the system functions. There may be privacy implications – particularly with sensitive information – so consumer privacy principles that empower users over what data to share are needed.[21] Simply implementing these processes is insufficient – critical examination of both elements is needed. If bias is found in "human in the loop" systems, there can be a mistaken tendency to trust the outputs of the algorithm rather than the human decision maker.[22]

Inevitably, designers and programmers will be building systems for life experiences they haven't had, making it important for communities impacted by these information systems to be involved in their development. Pertinent stakeholders from the target communities can make up a diverse and expansive list that varies from case to case. Depending on the specifics of the setting, this can include (but is not be limited to) users, internal employees involved in the development process, other potentially affected or involved parties, advocacy groups, etc. In addition, the notion of community may evolve. The construction of this community of stakeholders can be an iterative, dynamic process. This raises interesting and potentially difficult questions to inevitably grapple with, such as "who gets to decide and construct the boundaries of what a 'relevant stakeholder' is?" and "how does this construction serve as a vector of power?"

Once identified, relevant stakeholders' feedback and evaluations should be incorporated in a meaningful way. To the extent possible, technical barriers to entry should be reduced, allowing stakeholders to focus on providing their input and expertise as opposed to worrying about the technical nuances of the process. This can be achieved by using simulation and visualization tools in a manner that invites participation and engagement, such as through community-based system dynamics.[23] The goal should be for the whole team to "develop deep insights about important data to collect and consider, as well as evaluate the impact of products and decisions."[24]

It is crucial to monitor not just the output of an AI system, but also the process of checks, control, and quality of the system itself through regular audits.[25] Algorithms can be audited with several goals in mind: to control an organization's risk, to enable society to verify that its expectations of public and private organizations are met, to offer a way to articulate "accountability" and "responsibility", and to provide evidence of performance to a standard.[26]

Both internal and third-party audits are important. Internal AI audits can point the way for remedial actions needed before sale and deployment as well as after operationalization.[27] AI continues to learn and change over time.[28] Therefore, it is critical for these evaluations to be ongoing not only through the development / R&D cycle, but also when the AI system is in the market and being updated (e.g., through retraining). Internal audits can be supplemented by external auditors inspecting an ML product development process to ensure proper safeguards are in place. These external audits check the model's performance against predefined and agreed upon metrics that capture the values cared about. Increasingly, organizations and consultancies are offering third-party algorithmic audits for companies.

Algorithmic auditing is a relatively new practice, and it is important for the process to extend beyond a mindless "check-the-box" activity with little power or purpose. Lessons can be drawn from other fields employing audits. Still, innovation specific to the field of AI is necessary, since certain industry aspects lead to the following challenges:

- ***'Black box' auditing –*** Given the lack of standards around algorithmic development, it is unclear what artefacts or documents an audit should demand, and against what predefined criteria. In some cases, the auditor (particularly third-party / external) may only have access to the system's inputs and outputs —deep neural networks in particular are not designed to be auditable. This is problematic, often preventing audits from being sufficiently thorough. External auditors should have as much information as possible, even on "black box" AI systems (See Box 1). Documenting this information throughout the development process can help shed light at the auditing stage.

- ***Insufficiency of technical audits –*** Technical solutions can be designed to detect specific types of misbehaviors. But the fundamental concept of undecidability[29] in computing makes it impossible for technical solutions to detect and mitigate all instances of bias and social harms in algorithms.
- ***Navigating definitions of fairness –*** Fairness-related issues aren't often self-evident or exposed through discrete incidents. In addition, as explored in **EGAL's brief on fairness in machine learning** there is no industry-wide consensus on what and whose values get to decide what is fair.
- ***Stakeholder conflicts –*** Effective auditing requires slowing down to check concerns of various stakeholders and standards, which can be at odds with the culture of tech that has traditionally favored a rapid product development lifecycle.[30]
- ***Solutions serving specific functions –*** Audits do not guarantee safety, and can miss systemic risks – particularly as auditing AI systems for every form and source of societal bias is unrealistic.[31] While being unable to capture every malady, when leveraged appropriately appropriately audits can be used to find specific instances of bias. As such, audits in isolation should not be viewed as sufficient solutions, but rather as necessary safeguards that should be employed as part of a broader bias mitigation strategy.
- ***Third-party accountability –*** At present there are no regulations holding those selling algorithmic auditing capabilities accountable to a set of standards. There is also a lack of clarity around who should be held accountable to begin with.

BOX 1. UNDERSTANDING "BLACK BOX" VS "WHITE BOX" AI.

When it comes to auditing AI, auditors can be given varying levels of access to information about the algorithm. "Black box" auditing entails giving the auditors access to just the system's inputs and outputs. By contrast, "white box" auditing can be more thorough, with auditors knowing the internal workings and processes involved in its development.[32] However, these terms are likely to be subjective – even when access to the guts of a system is provided, implications of design decisions may be obscured.

This terminology can be applied to machine learning models themselves. Black box models refer to those AI systems for which users are able to observe inputs and outputs, but are unable to follow the exact decision making process. White box models satisfy two key criteria – their features are understandable, and their decision-making processes are clear and explainable.[33] To achieve this, white box models are likely to have less predictive capacity as compared to models that internalize massive amounts of data. Within this usage too, there is a degree of subjectivity. Explainability and what it means to "know" aspects of an algorithm differ based on the audience.

In light of these limitations, audits of AI systems should extend beyond the technical, to assess "fairness", robustness, security, etc., with technical audits being utilized as one component of a larger, more comprehensive auditing and monitoring strategy.

ADDITIONAL READING:

- **A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence** (IEEE)

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, Genevieve Smith, and Nitin Kohli.*

Berkeley Haas

egal

Endnotes

1 Turner-Lee, N., Resnick, P., & Barton, G. (2019, October 25). Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. Retrieved from https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

2 Martin Jr, Donald, Vinod Prabhakaran, Jill Kuhlberg, Andrew Smart, and William S. Isaac. “Participatory Problem Formulation for Fairer Machine Learning Through Community Based System Dynamics.” arXiv preprint arXiv:2005.07572 (2020).

3 Mulligan, Deirdre K., Joshua A. Kroll, Nitin Kohli, and Richmond Y. Wong. “This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology.” Proceedings of the ACM on Human-Computer Interaction 3, no. CSCW (2019): 1-36.

4 Martin Jr, Donald, Vinod Prabhakaran, Jill Kuhlberg, Andrew Smart, and William S. Isaac. “Participatory Problem Formulation for Fairer Machine Learning Through Community Based System Dynamics.” arXiv preprint arXiv:2005.07572 (2020).

5 Melgoza, A. [2020, June 18] Personal interview.

6 Mulligan, Deirdre K., Joshua A. Kroll, Nitin Kohli, and Richmond Y. Wong. “This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology.” Proceedings of the ACM on Human-Computer Interaction 3, no. CSCW (2019): 1-36.

7 Mulligan, Deirdre K., Joshua A. Kroll, Nitin Kohli, and Richmond Y. Wong. “This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology.” Proceedings of the ACM on Human-Computer Interaction 3, no. CSCW (2019): 1-36.

8 Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. “Model cards for model reporting.” In Proceedings of the conference on fairness, accountability, and transparency, pp. 220-229. 2019.

9 Reisman, Dillon, Jason Schultz, Kate Crawford, and Meredith Whittaker. “Algorithmic impact assessments: A practical framework for public agency accountability.” AI Now Institute (2018): 1-22.

12 See Rich Caruana’s talk “Friends Don’t Let Friends Deploy Black-Box Models” at the UC Berkeley School of Information for the Algorithmic Fairness and Opacity Group (AFOG) here: https://www.youtube.com/watch?v=TPY16CSIrwY&t=1581s

13 Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. “Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission.” In Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1721-1730. 2015.

10 Martin Jr, Donald, Vinod Prabhakaran, Jill Kuhlberg, Andrew Smart, and William S. Isaac. “Participatory Problem Formulation for Fairer Machine Learning Through Community Based System Dynamics.” arXiv preprint arXiv:2005.07572 (2020).

11 Mulligan, Deirdre K., Joshua A. Kroll, Nitin Kohli, and Richmond Y. Wong. “This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology.” Proceedings of the ACM on Human-Computer Interaction 3, no. CSCW (2019): 1-36.

14 See 58:40 of Rich Caruana’s talk “Friends Don’t Let Friends Deploy Black-Box Models” at the UC Berkeley School of Information for the Algorithmic Fairness and Opacity Group (AFOG) here: https://www.youtube.com/watch?v=TPY16CSIrwY&t=1581s

15 Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

16 Bartlett, P, Morse, A., Wallace, N., & Stanton, R. (2019). Algorithmic Accountability: A Legal and Economic Framework (Working Paper). Retrieved from: http://faculty.haas.berkeley.edu/morse/research/papers/AlgorithmicAccountability_BartlettMorseStantonWallace.pdf

17 Haniyeh Mahmoudian, P. (2020, April 10). How to Tackle AI Bias. Retrieved from https://towardsdatascience.com/how-to-tackle-ai-bias-ec39313ccacf

18 Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (n.d.). Algorithmic Fairness. DOI: 10.1257/pandp.20181018

19 James Manyika, S. (2019, October 25). What Do We Do About the Biases in AI? Retrieved from https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai

20 Dillon, S., & Collett, C. (2019). AI and Gender: Four Proposals for Future Research. https://doi.org/10.17863/CAM.41459

21 Turner-Lee, N., Resnick, P., & Barton, G. (2019, October 25). Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. Retrieved from https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

22 Elish, M. (2019). Moral crumple zones: Cautionary tales in human-robot interaction. Engaging Science, Technology, and Society, 5: 40-60.

23 Martin Jr, Donald, Vinod Prabhakaran, Jill Kuhlberg, Andrew Smart, and William S. Isaac. “Participatory Problem Formulation for Fairer Machine Learning Through Community Based System Dynamics.” arXiv preprint arXiv:2005.07572 (2020).

24 Martin Jr, Donald, Vinod Prabhakaran, Jill Kuhlberg, Andrew Smart, and William S. Isaac. “Participatory Problem Formulation for Fairer Machine Learning Through Community Based System Dynamics.” arXiv preprint arXiv:2005.07572 (2020).

25 AFOG. https://afog.berkeley.edu/files/2019/07/AFOG_workshop2018_report_all_web.pdf

26 AFOG. https://afog.berkeley.edu/files/2019/07/AFOG_workshop2018_report_all_web.pdf

27 IBM. https://www.ibm.com/blogs/policy/ai-precision-regulation/

28 HBR. https://hbr.org/2018/02/can-we-keep-our-biases-from-creeping-into-ai?referral=03758&cm_vc=rr_item_page.top_right

29 Desai, Deven R., and Joshua A. Kroll. “Trust but verify: A guide to algorithms and the law.” Harv. JL & Tech. 31 (2017): 1.

30 AFOG. https://afog.berkeley.edu/files/2019/07/AFOG_workshop2018_report_all_web.pdf

31 Neff, G. [2020, February 19]. Personal interview.

32 AFOG. https://afog.berkeley.edu/files/2019/07/AFOG_workshop2018_report_all_web.pdf

33 Sciforce. (2020, January 31). Introduction to the WhiteBox AI: The Concept of Interpretability. Retrieved from https://medium.com/sciforce/introduction-tothe-white-box-ai-the-concept-of-interpretability5a31e1058611