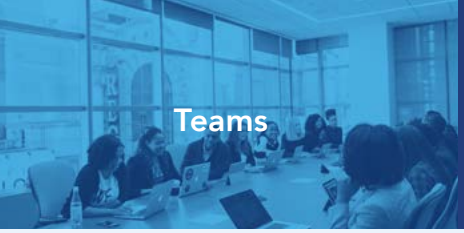

Play 2. Promote a culture of ethics and responsibility related to AI

Enable a culture that empowers and encourages employees to prioritize equity considerations at every step of the algorithm development process. In line with the understanding that completely de-biasing AI may not be feasible, organizations should uphold a standard of explainability around the workings of their models, as well as transparency around potential shortcomings / pitfalls.

PLAYERS INVOLVED:
- CEO & C-suite
- AI governance team(s)
- Product managers and teams designing and developing AI

BUSINESS BENEFITS:
- Mitigate risk
- Have a superior value proposition

Elements:

Tools:

Elements:	Tools:
☐ Have and communicate explicit principles for responsible and ethical AI.	• See Play 5 for specific guidance on developing and implementing responsible AI principles.
☐ Update individual performance review processes to include a component around responsible and ethical AI practices. • Incorporate responsibility and ethics as part of the criteria individuals are assessed against – from entry-level employees through management and senior leadership. These criteria can be tied to the company's ethical AI principles. • Ensure that the effects of AI systems on users and society more broadly are considered in conjunction with user growth and engagement. • Reward, recognize and publicly highlight those that practice strong responsibility and ethics related to AI.	• Quick Win! Update performance review process & OKRs (EGAL)
☐ Update Objectives & Key Results (OKRs) / Key Performance Indicators (KPIs) to integrate goals and metrics on the mitigation of bias in AI. • Ensure adequate resources are provided to address potential bias or fairness issues that arise.	• Quick Win! Update performance reviews process & OKRs (EGAL)
☐ Integrate ethics screening into hiring requirements for new employees as well as promotion processes. • Integrate questions (behavioural or situational) and/or case studies related to ethics and responsibility in the interview process. Example: "Tell me about a time when you uncovered a potentially unethical issue or were faced with an ethical conundrum professionally. What did you do and why?"	
☐ Embed education on ethics, bias, and fairness for employees developing, managing, and/or using AI systems. • Include technical and non-technical forms of bias. • Ensure that education materials are adapted to address industry-specific issues, concerns, and trends.	• Quick Win! Educate staff about bias in ML artificial intelligence (EGAL)

Elements:	Tools:
☐ Embed education on equitable and inclusive language for employees developing, managing, and/or using AI systems. • Educate employees labelling data or working with labelled data on challenges and pitfalls of language used to label data, particularly images. Incorporate trainings on unconscious biases / how those can be embedded in labels. • For employees developing AI systems for different business functions (e.g., HR and hiring), build in training on equitable and inclusive language as it relates to that topic (e.g., gendered language in job descriptions).	• **EFL Glossary of Key Terms** (EGAL) • **Writing inclusive documentation** (Google) • See Play 3 for specific training guidance for labelers
☐ Make explainability and transparency around shortcomings and pitfalls of AI systems the norm.	• See Play 4 for specific guidance on how to operationalize this throughout the development process for each AI system.

Note: See more on responsible AI governance and ethical AI principles in Play 5; and find specific practices to operationalize ethics in AI models are outlined in Play 4.

Examples & leaders:

Google: Google has several resources and initiatives to advance equitable and inclusive language internally among employees developing AI systems. In addition to internal guides with educational information around 'inclusive' and 'respectful' coding language, the company has also established **inclusive documentation guidelines for developers**. This includes primers on how ableist, gendered, or unnecessarily violent language can slip into documentation. It offers concrete examples of such language as well as alternatives, encouraging developers to be conscious about the unintended harms that word choices can cause. Importantly, instead of attempting to curate an exhaustive list of terms and examples, the guidelines outline best practices for developers to incorporate the preferences of specific communities.

Johnson & Johnson: Johnson & Johnson, which leverages AI for healthcare and drug discovery, includes ethics in executive 360-degree evaluations. The evaluation builds from the four components of the company's "**Credo**", the values on which their decision-making is based, and which centers responsibility. In a version of the evaluation, each executive was rated on items such as, "nurtures commitment to our Credo," "confronts actions that are, or border on, the unethical," and "establishes an environment in which uncompromising integrity is the norm."[1]

Background:

The priorities of business leaders matter, as they influence the organization and its employees.

Hierarchies and the chain of command have important impacts on how mistakes, failures, and problems are recorded.[2] If ethics isn't explicitly embedded in organizational culture, lower level staff who can/do spot biases in AI may not feel empowered to speak up for fear of backlash. Mid-level management (including project managers) may not prioritize mitigating bias either. To overcome these challenges, it is important to understand and reassess organizational priorities and associated incentives, as well as power structures that can perpetuate harmful biased AI systems[3] (See Play 5 for more on corporate governance and leadership priorities).

It is important that employees are not only able to speak up about concerns around bias in AI, but expected to do so. This expectation can be created by building responsibility around development and management of AI into performance reviews and criteria for promotions, KPIs, OKRs, and hiring.[4] These incentive structures are instrumental in building clarity, comfort, and trust when it comes to flagging and addressing ethical issues.[5] They also signal the importance to the company of considering and addressing issues of bias in AI. There is not always a right or wrong answer when it comes to building and managing "fair" and "ethical" AI systems. However, when employees are able and expected to highlight and discuss potential areas of bias in their AI systems, it supports risk mitigation and enables documentation to understand the rationale for choices made.[6]

Employees must understand how bias can be embedded in AI systems and what to do about it. Integrating workshops around bias in AI into onboarding and ongoing trainings for employees is important. Internal training should focus on technical and non-technical forms of bias in AI systems, as well as fairness-related tradeoffs. These educational workshops should highlight how mitigating bias connects to the company's goals, values and priorities.

Bias trainings should be supplemented with workshops focused on using inclusive language. Developers should understand the implications and potential harms that can be caused by gendered, racialized, ableist, or otherwise biased labels in their data. Trainings can address how labelling data and using language in AI systems can be done in a way that does not perpetuate biases or discrimination.

Incorporating transparency and explainability of AI models is necessary – both internally and externally. The natural question that arises when algorithms produce unintended or undesirable consequences is: how did this happen?

Internally, teams developing AI systems should be able to explain why and how their AI system arrived at a particular decision—as well as reference information about training data and the algorithms. In many cases, there are several siloed teams working on different parts of an AI system. Also, ML systems are often "black box" models whereby it is difficult or impossible for even the developers to know why the ML system made a particular prediction. It is key to document what goes into an ML system (e.g., data, proxies, guiding frameworks for decisions made) and enhance communication between teams. This communication should extend to non-technical staff, including teams promoting and selling the model.

Externally, full transparency on a model is not realistic or necessarily sensible. There can be privacy and exploitation concerns. Complete transparency on the guts of a model can result in "model inversion", whereby bad actors can potentially exploit AI systems.[7] Even when models are not "black box", companies want to protect their IP. That said, it is still important to have a certain level of transparency and explainability for users and end targets whose lives are impacted by the model. It is important to ascertain what in particular should be made transparent and explainable – and for whom this information is curated.[8] See **Play 4** for specific processes that can help operationalize transparency and explainability for those building, sourcing, and/or impacted by AI systems.

ADDITIONAL READING:

- How businesses can create an ethical culture in the age of tech (World Economic Forum)

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 Genevieve Smith and Ishita Rustagi. Input was provided by Nitin Kohli.*

Endnotes

1 Epley, N. & Kumar, A. (2019). How to design an ethical organization. Harvard Business Review.
2 Neff, G. [2020, February 19] Personal interview.
3 Benjamin, R. (2019). Race after technology. Polity Press.
4 Krieger, Z. (n.d.). A Practical Guide for Building Ethical Tech. Retrieved from https://www.wired.com/story/opinion-a-practical-guide-for-building-ethical-tech/
5 Written by Ann Skeet, D. O. (n.d.). How businesses can create an ethical culture in the age of tech. Retrieved from https://www.weforum.org/agenda/2020/01/how-businesses-can-create-an-ethical-culture-in-the-age-of-tech/
6 Canca, C. [2020, February 5] Personal interview.
7 AFOG. https://afog.berkeley.edu/files/2019/07/AFOG_workshop2018_report_all_web.pdf
8 Kohli, Nitin, Renata Barreto, and Joshua A. Kroll. "Translation tutorial: a shared lexicon for research and practice in human-centered software systems." In 1st Conference on Fairness, Accountability, and Transparency. New York, NY, USA, vol. 7. 2018.