How might companies mitigate gender bias in Machine Learning?

Pablo Nazé, 2019-20 GATE MBA Fellow
Institute for Gender and the Economy (GATE), Rotman School of Management, University of Toronto

April 2020

Acknowledgments

I would like to thank all of the project's interviewees for their openness and insights. Thanks to all of the several professionals and researchers who provided feedback and pointed to resources during this research, including multiple attendees at the ACM Conference on Fairness, Transparency, and Accountability (FAccT). Further, I'm extremely grateful for the support of The Institute for Gender and the Economy (GATE) and Bain & Company during the entire fellowship. Finally, my sincere compliments to my colleagues and MBA fellows at GATE and my dear friends and family who supported me through these intense final months of the MBA.

Table of Contents

Executive Summary

With Artificial Intelligence (AI) and Machine Learning (ML) becoming more widespread, companies need to face the challenges of mitigating gender bias in these systems.

With an explosion of resources available, companies need help to make sense of the latest insights provided by research. We provide a framework – understand fairness, engage stakeholders, and build fairness skills – to help companies navigate the complex and multidisciplinary field of gender bias in Machine Learning. For those interested in a quick overview of the field, the report points to a quick summary table with the main concepts and suggested action points.

This report draws insights from a vast body of studies, working papers, and conferences, as well as 8 semi-structured interviews with practitioners. We shed light on the complexities of fairness, introducing concepts from a myriad of disciplines, ranging from computer science to sociology.

Companies can consult this report as a handy reference guide to help them in their journey towards gender bias mitigation.

Introduction

The world is a biased place for women. According to a report published by the United Nations Development Programme, more than 90% of men and more than 85% of women exhibit some sort of bias against gender equality and women's empowerment [1]. Artificial Intelligence (AI) and Machine Learning (ML) systems[1] are not immune against those bias, which have been reported in multiple domains of this field of study.

A famous study called Gender Shades tested the accuracy of gender classification software across groups, only to discover that the algorithm worked worse for women, especially women of colour [2]. Research has also shown occurrence of gender bias in algorithms that learn to associate words with a specific gender and in methods of automatic language translation [3]. More recently, Apple suffered gender bias allegations during the launch of its new credit card, when heterosexual couples that filed taxes together received different amounts of credit [4].

Responding to several of these cases, companies started to deploy different strategies to deal with biased systems. Microsoft, Google, and IBM have released AI principles or ethics codes [5], and companies such as Accenture and Salesforce created positions [6] [7] for the development of responsible AI.

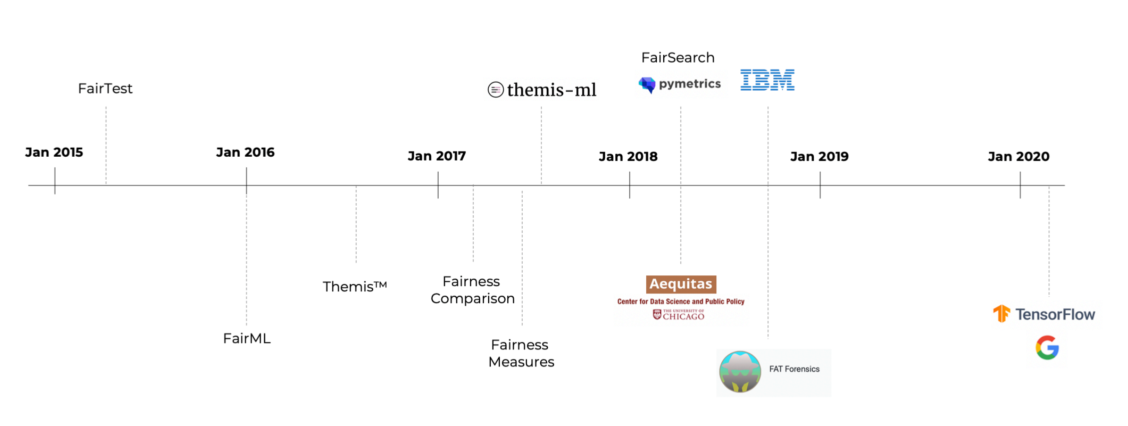
However, these advancements are still not widespread. The majority of existing technical toolkits are recent (Figure 1) and one of the biggest conferences in this subject, the Association for Computer Machinery conference on Fairness, Accountability, and Transparency (FAccT) had its first edition in 2018. Further,

[1] Although scholars may still debate precise definitions of these terms, we can think of Artificial Intelligence as "the broader concept of machines being able to carry out tasks in a way that we would consider 'smart'" and Machine Learning as a subset and current application of Artificial Intelligence, around the idea that machines can learn from data without being explicitly programmed [71]

research has shown that not all practitioners include fairness and bias in their checklists while developing AI/ML models [8]. During this project we have come across similar results, as we conducted 8 semi-structured interviews with practitioners in this field.

Given this scenario, **how might companies mitigate gender bias in Machine Learning?** This report investigates this question, drawing insights from practitioners in the industry and from the latest available research.

Figure 1 – Timeline of Programming Libraries for Fair Machine Learning



Source: Author's research, Dunkelau & Leusche (2019)

Mitigating bias is more than a technical challenge

Interviews with practitioners

During this project we conducted 8 semi-structured interviews with practitioners in the Data Science and Machine Learning (ML) space. The questions were open-ended, asking practitioners to recall past experiences with Machine Learning and in projects were gender was one of the variables. We also prompted interviewees with thought exercises on how to mitigate bias in models. Finally, we presented practitioners with a proposed checklist to help them mitigate bias in their projects, collecting their reactions and feedback. Although not statistically representative, these interviews revealed interest insights that illustrate and validate conclusion from other studies.

Technical Approaches

Interviewees were prompted with a thought exercise. If they were working in a system with gender bias allegations in the press, how would they go about this challenge? The goal of this question was to reveal any underlying mental model of how practitioners would approach bias in Machine Learning.

Most practitioners took an analytical approach, replying that they would try to identify the root causes of the bias. Some mentioned model debugging approaches, while also acknowledging the challenges of interpretability in more complex models.

Practitioners intuitively thought that datasets were the main cause of bias. Although this is a valid first guess, few practitioners acknowledged other possible sources of bias[2]. This result suggests that practitioners may not feel empowered to fix bias in

[2] Sources of bias and more technical aspects will be discussed further in the report

Machine Learning. A recurrent comment was that if datasets, or society as a whole, are biased, the model will be biased as well. Consequently, some practitioners believed that "there is nothing they could do about it".

Practitioners often proposed that removing gender as a variable would eliminate bias from systems. However, this approach of "fairness through unawareness" [9] is usually not effective, as models can infer gender through other variables, such as names. Although some interviewees recognized this risk, no practitioner spontaneously suggested to keep gender as a variable to help to measure and mitigate bias. The closest reaction was given by one practitioner, who was interested in using gender to try to predict customer behaviour.

Organizational Behaviour

Practitioners reacted well to the idea of having a "bias mitigation checklist"[3] to guide them through model development. This suggests that there is a lack of guidance and frameworks available to help practitioners, as also pointed by a bigger study [8] .

When asked what organizations needed to do to successfully implement such checklist, practitioners usually referred to the importance of company support. Some practitioners pointed out that senior leadership would have to be committed to the checklist, providing an example from the top. Another theme explored was pressure from customers, who may squeeze teams for faster results, even if that means skipping steps in the checklist or disregarding fairness altogether. Finally, a few practitioners pointed out that regulations may motivate teams to implement the checklist.

[3] An overview of checklists will be discusses further in the report

The presented checklist [10] covered more topics than just fairness, including user-consent and privacy. An interesting result was that practitioners often focused in these more regulated and known issues instead of fairness. This suggests a potential risk for fairness procedures to be overshadowed by other data governance aspects.

A study who surveyed over 200 practitioners found similar results [8]. Researchers found that only 30% of the surveyed practitioners included fairness and bias in their regular model development checklist. At first this may seem as a reasonable number, but when asked about including privacy concerns in their checklists, 64% of practitioners responded positively, more than the double for fairness issues.

Limitations and Discussion

As mentioned, these interviews had a small sample, resident in North America and recruited through the researcher's immediate network. Although these insights cannot be extrapolated to a larger set of practitioners, they point in the same direction as larger studies.

Perhaps the most interesting insight from the interviews is that mitigating gender bias in Machine Learning is more than a technical challenge. While practitioners still need to understand better where bias come from and how to address it properly, organizational aspects, such as leadership and change management, should also play a role in helping companies tackle this challenge.

A framework to help companies mitigate gender bias in Machine Learning

Incorporating organizational structure in approaches towards fairer Machine Learning is not new. Similar ideas were explored by a recent tutorial session in one of the main conferences in the field, using the Berkana Institute's Loop Theory of Change model [11] .

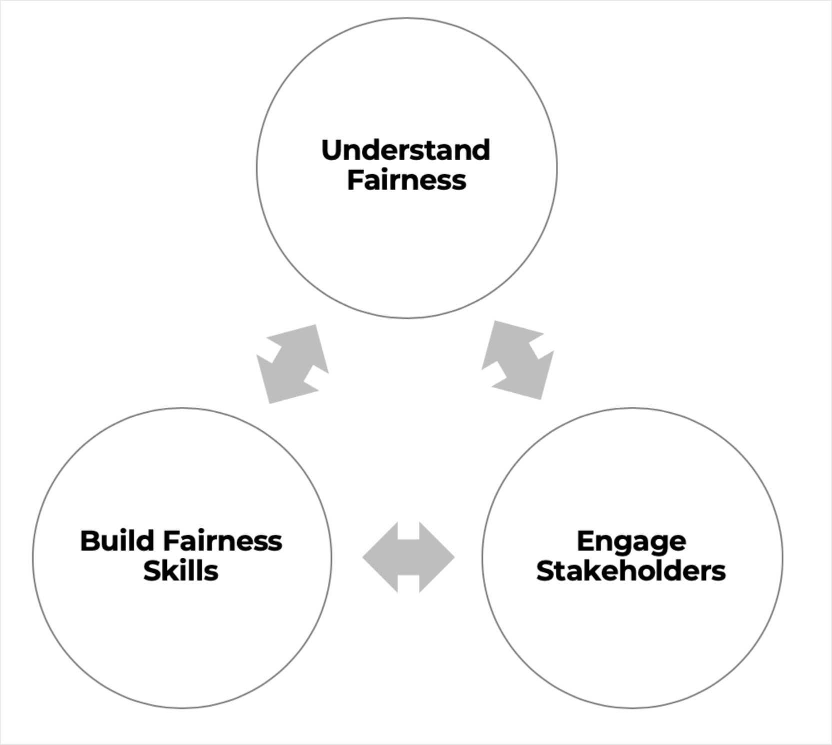
The framework proposed in this report builds on similar concepts but leverages John P. Kotter's classic change management framework [12]. Familiar to any business administration student, Kotter's framework has eight steps and a specific order that leaders need to follow while introducing organizational change. Among those steps, creating a vision, building a coalition, and institutionalizing change are relevant to our scenario.

Our framework has three stages, who will be discussed in detail through the report. Analogous to some of the Kotter's steps, these stages are *understanding fairness*, *engaging stakeholders*, and *building fairness skills*.

First, companies need to understand fairness with all its nuances and complexities. This is will help them to treat this matter more objectively and to make more informed decisions. Second, companies should engage stakeholders. Fairness and bias impact different groups in different ways, and companies should take their feedback into account. Third, companies need to build fairness skills. Without them, execution will not happen and mitigating bias will become an empty speech.

We also argue that these steps should be taken simultaneously, as they continuously influence each other (Figure 2).

Figure 2 – A framework to mitigate gender bias in Machine Learning



This report expands on each of these steps, providing many supporting arguments through the latest research and evidence available. Future work could expand on this framework by applying it to use cases or organizing experiments in active companies.

Understanding Fairness

We can think about mitigating bias as being similar to optimizing for fairness, an approach followed my most of the literature published on this subject. This brings the question of what fairness is and how one can approach this subject. In other words, how might companies **Understand Fairness**, the first step in the framework.

Fairness is a hard, contextual, and multidisciplinary concept, which have been explored through centuries by philosophers and thinkers. More recently, scholars proposed formalizing fairness concepts in mathematical terms, through the so-called "fairness metrics". The advantage of introducing fairness metrics is to provide a measurable way to measure a nuanced concept, making computational treatment and interventions easier. An example is useful to understand them.

Imagine a school selecting students to attend university for the first time. This school receives many applications, both from men and women-identified individuals. For simplicity, assume applicants only identify with these two genders[4], how this school should approach fairness and which metric should it choose?

Perhaps the school believes that approving the same percentage of students in both groups is fair. This would be called *demographic parity,* when the outcome has no correlation with the *protected attribute (*a variable that marks certain characteristic, such as gender, age, or race) [9]. The school may decide this approach as to avoid the following argument: "Selecting 60% of male students and only 20% of female students is discrimination".

[4] Later in the report we will address gender binarism

However, suppose that for some reason the *abilities* of the students are indeed different. In this case, selecting the same percentage of students would be undesirable, as non-qualified students may be admitted to meet *demographic parity*. Assume that men are *less qualified* (instead of equally qualified) than women, i.e.: they would be less successful students than women. In this case, approving the same percentage in each group would imply selecting non-qualified students, an undesirable outcome. The school then can decide to achieve *predictive equality*, or equal false positive rates, to ensure that the percentage of accepted unqualified men is the same percentage of accepted unqualified women [13].

There is no lack of different fairness metrics. A recent literature review points to more than 20 fairness metrics in seven different categories [13], while IBM's technical toolkit mentions over 70 fairness metrics [14]. Although new metrics can also be proposed at any point by researchers, a study from 2019 identified that many metrics are correlated with each other, recommending that "a new measure for fairness should only be introduced if it behaves fundamentally different from existing metrics" [15].

Choosing which metrics to pursue is not straightforward, as discussed by many studies [13]. Some of these metrics are mutually exclusive, implying that satisfying all metrics and achieving "complete fairness" is impossible [16] [13]. Moreover, this choice may be dependent on a company's available techniques and datasets characteristics [15].

There are some existing efforts to help practitioners decide on metrics, such as a "fairness tree" provided by Aequitas, an "open source bias audit toolkit" made by The University of Chicago [17]. This decision tree helps user navigate some of the trade-offs across metrics, such as being fair based on representation or errors.

Nevertheless, perhaps companies should choose a fairness metric also considering social aspects. For example, companies may be tempted to choose the "most popular" definition of fairness. However, delegating this choice to the public may not be effective. A recent study asked 502 participants to choose which fairness definition they would prefer. The results showed no overwhelming preference, implying that "relying on society" to make this decision is also not simple [18].

Conclusion

At this point, consequently, there are no easy technical or social rule of thumb to choose one metric over the other. Use cases are highly contextual and social implications may vary across industries. Therefore, companies may face challenges while understanding what fairness means to them. To tackle these challenges, we argue that comprehending the different approaches and possibilities is better than delegating or ignoring this choice. Understanding fairness and its complexities is essential to help companies make more informed decisions on this subject.

Action Points

- Commit to understand fairness and its nuances; internalize that there are no easy answers
- Embrace the many possibilities of fairness while refining your company's comprehension of the subject
- Familiarize yourself with concepts such as bias and fairness metrics

Consulting Stakeholders

Companies should engage different perspectives while refining their comprehension of fairness. A multidisciplinary and nuances concept, fairness is better understood when taking the opinions and views of multiple stakeholders.

Engaging stakeholders, the second step of the framework, offers a quick way to consider the opinions of key stakeholders in fairness issues. The proposed list of agents can be adjusted across companies, industries, and products, but the four agents treated here may be a useful starting point for all organizations.

Investors – The business case for fairness

In a corporate culture of "moving fast and breaking things", executives may not be keen in investing resources to address fairness issues. A clear obstacle are the requests for returns on any investment. When faced with this issue, practitioners may ask themselves if there could be a business case for having more fair systems. Adjacent markets may provide evidence that fairness in Machine Learning is good for business.

Recently, companies have invested heavily in sustainable products. Starbucks reports to have invested a hundred million dollars in coffee communities, achieving a milestone of having 99% of ethically sourced coffee [19]. Additionally, Nike is increasingly using more sustainable materials in its production line, having diverting "more than 7.5 billion plastic bottles from landfills and waterways" since 2010 [20].

Customers are also responding well to these sustainable products. A comprehensive study of more than 36 categories in customer packaged goods (CPG) identified that "50% of CPG growth from 2013 to 2018 came from sustainability-marketed products", with 114 billion dollars in sales coming from "products that had a sustainability claim on-pack" [21].

In the tech industry, data breaches provide an interesting case. With recurring reports[5] well documenting the cost of a data breach, which can cost millions of dollars and span for several years [22], business leaders are often confident in making a clear business case for investing in safer stored data. Venture capitalists are also concerned with trust and cybersecurity, having published whitepapers around the subject [23] and pouring over 5 billion dollars into cybersecurity ventures in 2018 [24].

In terms of gender equality, much has been published about the business case for diversity[6]. Reports correlating gender diversity with profitability have been published [25] and Pamela Newkirk's book, *Diversity, Inc.: The Failed Promise of a Billion-Dollar Business,* points to billions of dollars spent in the "diversity industry", invested by companies trying to tap into the benefits of diversity [26].

Although to the best of our knowledge there are no specific studies quantifying the cost of gender bias in Machine Learning, we could predict that they will eventually be created. A current alternative for companies is to create and track metrics for their bias mitigation projects, such as number of potential customers impacted. Establishing these metrics help senior leaders to understand more concretely the impacts of initiatives in this space.

Some fields of inquiry are pointing towards more dollar-measure of gender bias in Machine Learning, such as bias in digital advertising delivery. Although advertisers

[5] The Ponemon Institute annually publishes its "Cost of a Data Breach Report", sponsored by IBM Security

[6] For an interesting discussion on the limitations of the business case for diversity, see [47].

can often target and exclude groups while creating digital ads, there is evidence that these systems may be biased. This happens during the delivery of ads in "ways that the advertisers do not intend", in parts because how the platforms predict ad relevance [27]. As a result, it may be possible to quantify how much companies may lose in business, in the form of poorly targeted advertising, due to bias.

Conclusion

We can expect that mitigating gender bias in Machine Learning is good for business. There is evidence of business results in adjacent markets, such as sustainable products, preventing data breaches, and having more diversity in the workforce. Although, to the best of our knowledge, there are no studies quantifying financial impacts of gender bias in Machine Learning, recent research may point a path towards more concrete numbers.

Action Points

- Use adjacent industries to make a business argument for mitigating gender bias in Machine Learning
- Develop metrics and key performance indicators (KPIs) to track initiatives in your company

Regulators and legal teams – The legal case for fairness

Regulation may be a more compelling argument for companies to act on bias mitigation. There are two main stakeholders concerned with the legal case for fairness. Regulators, who will develop specific rules regarding Machine Learning (ML) and Artificial Intelligence (AI), and legal teams inside companies, who will make sure that their organizations comply with these rules. We suggest approaching legal considerations threefold, understanding the current landscape, keeping up with ongoing efforts, and paying close attention to industry-specific regulation.

The AI regulatory landscape[7]

There have been many attempts to propose regulatory frameworks for Artificial Intelligence, both by private companies and by the public sector. A recent study mapped the landscape and identified 84 of documents "containing ethical principles or guidelines for AI". Out of these 84 documents, 19 have been produced by private companies whereas another eight documents have been authored by "intergovernmental or supranational organizations" [5].

After reviewing these 84 guidelines, researchers identified five converging ethical principles. This convergence may help to guide companies wondering which documents they need to pay attention to. The identified principles are "transparency, justice and fairness, non-maleficence, responsibility, and privacy" [5]. Although a global convergence is a positive sign of alignment, the authors also identified that these documents diverge on specificities of each principle, including their exact interpretation and how to implement them. This convergence and divergence across the documents showcase how AI regulation is important, yet not fully developed.

[7] Some researchers may question the applicability of the current regulatory framework towards AI. For an overview of the discussion and recommendations, refer to Clark and Hadfield, 2019 [56]

Keeping up with ongoing efforts

Companies can keep up to date with regulatory efforts through several means. A interesting source is the Organisation for Economic Co-operation and Development (OECD), who launched its AI Policy Observatory in February 2020. Defined as a "platform to share and shape public policies for responsible, trustworthy and beneficial AI", the observatory consolidates initiatives, trends, and data concerning the responsible usage of AI [28].

Besides multinational efforts, companies may want to pay attention to regional guidance and existing regulation. Europe is a good example for that. The European Union (EU), who has made tremendous impact in 2018 with the implementation of its General Data Protection Regulation (GDPR), is working towards clearer AI rules. The European Commission created a High-Level Expert Group and published its "Ethics Guidelines for Trustworthy AI" in 2018 [29]. These guidelines are still being tested and revised, with the goal of finalizing this work by June 2020 [30]. Nevertheless, the European Commission stressed in a recent whitepaper that "Developers and deployers of AI are already subject to European legislation", including fundamental rights, consumer protection, and product safety [30].

Canada is also working towards an AI regulation. The Government of Canada published a Directive on Automated Decision-Making, ensuring that its systems "are deployed in a manner that reduces risks to Canadians and federal institutions" [31]. The government also released its Algorithmic Impact Assessment (AIA), an open-source questionnaire designed to help "assess and mitigate the risks associated with deploying an automated decision system" [32]. Additionally, the Office of the Privacy Commissioner of Canada (OPC) opened a consultation on its proposals for "ensuring appropriate regulation of artificial intelligence" [33].

Industry-specific regulation

Some industries may already have clear anti-discrimination rules that companies should comply to, such as the hiring lending industries.

In the United States (U.S.), the Equal Employment Opportunity Commission (EEOC) established the 4/5ths rule in its Uniform Guidelines on Employee Selection Procedures [34]. An example is helpful to understand it. Assume that company is hiring for a position and that, historically, 60% of the men applicants were selected. According to the 4/5ths rule, at least 48% of the women applicants should be selected, which is 4/5ths (80% times 60%) of the men's approval rate. Such rule impacts directly the deployment of AI. A recent studied identified that companies using AI to develop hiring software may develop specific techniques to comply to the 4/5ths rule in the U.S., which may be an issue when the same software is used in countries with different legislation [35].

Similarly, the U.S. Equal Credit Opportunity Act (ECOA) prohibits credit discrimination on the basis of sex, race, and many other attributes. Although companies may ask and collect for those specific data points, they may not use them when to make a decision on credit [36]. If companies are forbidden of using these *protected attributes* during model development, it might be impossible to accurately measure bias to then mitigate it. In fact, excluding these variables from a model may be even worse, as this "fairness by unawareness" approach [9] may introduce bias through variables that are initially "potentially non-discriminatory" [37]. For example, if women are under-represented in certain occupations, removing the "gender" from your model but keeping "occupation" may lead to bias.

Conclusion

As noted, there are many issues to consider for the legal case for fairness. Justice and fairness are emerging as a global principle during the development of ethical AI. Moreover, although specific AI legislation is still in development, companies may already be affected by current regulation, such as multiple anti-discrimination rules.

Action Points

- Make sure your legal team is aware of the latest Artificial Intelligence and industry-specific regulations
- Document any legal roadblocks that may prevent your company to implement fairness measures
- Consider developing a set of Responsible Artificial Intelligence guidelines, acknowledging the convergence of global principles

End-Users – The user-centric case for fairness

Companies value customer centricity more than ever. CEOs believe that data about customers preferences and needs is the most critical for making decisions about the long-term success of business [38]. Additionally, customer-centric companies are more profitable than their counterparts, display a more engaged workforce and generate more revenue [39].

The benefits of being user-centric are illustrated by Amazon, famous for treating customer obsession as one of its leadership principles [40] and one of the most valuable brands in the world [41]. As a result, we should expect that tech companies will consider their **end-user expectations** while developing more fair Machine Learning systems.

How users may perceive fairness

On one hand, there is evidence that customers would prefer more ethical companies. Ethical Artificial Intelligence may be a "critical differentiator" for business, as a large majority of customers seem to report being "more loyal to ethical companies" [42]. Additionally, customers may pay particular attention to how companies deal with gender issues, engaging in boycotts against firms connected to sexual harassment accusations [43].

On the other hand, however, some customers may have reservations to the best ways to address fairness in Machine Learning (ML). A study with ML practitioners cites how end-users may interpret fairness intervention as "manipulative" and "unethical" [8]. For example, a web search engine may decide to exhibit gender-balanced results for the "CEO" term, even if is not the reality of real-life companies.

Additionally, your users' may form a heterogenous group, with different perspectives on fairness. For companies operating as marketplace platforms, such as Uber and Airbnb, there may be additional conflict between the needs of the supply side and the demand side of the market. This conflict is also illustrated by the already mentioned study that surveyed participants on which fairness metric they would prefer, not identify a clear preference [18].

Consequences

Companies may also face unintended consequences of addressing fairness. For example, companies may have high costs of collecting granular demographic information to measure bias [8], both financially and in terms of user experience friction.

For example, some may argue that collecting data about gender may require another field in a registration form. Although a valid point, some companies may decide to collect the data while minimizing the friction in experience. Amazon and Twitter, for example, infer gender through other variables, creating a smoother user sign up process [44]. Nevertheless, this approach has problems on its own, as users should consent to these terms and understand how they are being classified [44].

Conclusion

Companies must listen to the aspiration of end-users, balancing different opinions and dealing with unintended consequences of addressing fairness in Machine Learning systems. A best practice would be incorporate User Experience (UX) designers in the Machine Learning development process, which is still an uncommon practice with challenges on its own [45].

Action Points

- Engage User Experience designers during the Machine Learning pipeline, ask for their inputs and insight
- Innovate and develop techniques for prototyping, testing models, and deploying ethical interventions in front of end users.
- Be aware of trade-offs between addressing customers’ demands and increased development costs

Civil Society – The social justice case for fairness

Companies may be inclined to pursue fairness in Machine Learning because it is the right thing to do, sometimes appealing to Corporate Social Responsibility (CSR) arguments. With many successful initiatives, especially in the environmental front, companies can understand that seeking social justice is possible in many contexts.

Although a welcomed argument, companies should be aware of CST limitations. Sometimes these initiatives are co-opted by financial interests, evidenced by pressures in demanding business results from these initiatives [46]. For gender equality and diversity, specifically, companies should also understand the latest evidence that the "business case for diversity is not working" [47]. This does not mean stopping to pursue these initiatives, but finding alternatives to successfully deploy them.

In the case for gender bias in Machine Learning (ML), civil society and other organized groups will hold companies accountable for their development and deployment of ML systems. These groups may pay attention to two specific concepts, ethics washing and technosolutionism.

Ethics washing refers to the "growing instrumentalization of ethical language by tech companies" [48]. In the context of ML, this means companies promoting ethical initiatives without properly acting on them. Such initiatives may be the creation of ethical boards to oversee ML development or specific roles to aid in the development of more fair systems. The problem is not having those mechanisms in place, but rather depriving them from the power to make necessary changes, or simply using them as a "façade that justifies deregulation, self-regulation or market driven governance" [48].

Technosolutionism is the idea "that technology can unilaterally solve difficult social problems" . In the case of mitigating gender bias in ML, technosolutionism implies that companies can fix a broader social problem – gender inequality and sexism – only through a technical approach. Although the development of technical approaches and toolkits are fundamental to advance towards more fair ML systems, it is important to note that "these methods and toolkits often rely on simplified, quantitative definitions of complex, nuances concepts" [51].

Consequently, even with the best of intentions, companies need to be aware of possible consequences of launching ethical initiatives. Google faced backlash in 2019 after announcing a multi-disciplinary board to oversee its Artificial Intelligence efforts. The company dissolved the group one week after the announcements, in parts due to the repercussions of having included a conservative member on the board [48] [49].

There are organizations keeping industry and companies accountable for responsible Artificial Intelligence. Founded by Joy Buolamwini, The Algorithmic Justice League (AJL) is an organization that, among several initiatives, raises "awareness about the impacts of AI" and "build the voice and choice of most impacted communities" [52]. Another relevant organization in this space is The Algorithm Watch, which evaluates and sheds light on "algorithmic decision-making processes that have a social relevance" [53].

Conclusion

Mitigating gender bias in Machine Learning because it is the right thing to do is a much-welcomed complement to the other cases for fairness. Nevertheless, companies should be aware of how they will be hold accountable for seeking social justice. In our context, ethics washing and technosolutionism are possible traps that companies may fall into. To avoid them, companies should understand them and take action to avoid them, such as truly committing to responsible AI or partnering with civil society organizations.

Actions Points

- Listen to civil society and organizations during your Machine Learning development and deployment
- Consider putting as much emphasis in the social justice case for fairness as the business, legal, and user-centricity cases
- Fully commit to fairness, avoiding ethics washing and technosolutionism

Building fairness skills

Once a company is comfortable with its fairness approach, both understanding the concept and dialoging with stakeholders, the next logical step is to organize to execute an established vision. We identified four pillars of the fairness skills that companies need to effectively mitigate gender bias in Machine Learning.

People

Hiring and Training

The field of bias in Machine Learning is recent. There has been a ten-fold increase in technical publications with 'gender bias' as a main topic since 2015 [3] and one of the main conferences in this field, the ACM Conference on Fairness, Accountability, and Transparency (FAccT) had its third edition in 2020.

Consequently, companies should not expect to encounter a workforce who is familiar with these new concepts and techniques. To mitigate this issue, companies may consider training employees or compensating them for taking external courses. Nevertheless, assessing for knowledge about fairness during recruiting may still be desirable, screening candidates that may not even acknowledge the subject.

Team composition, Diversity and Inclusion

In its report *A Call to Action for Businesses Using AI,* The IEEE Standards Association points to several recommendations on how to create a culture of ethics. The report calls for the identification, recruitment, and training of multidisciplinary and diverse individuals[8] that "may be already doing ethical AI work" [54]. Given the nuances and complexities of fairness in Machine Learning, these multiple

[8] Some organizations are working to improve diversity representation in AI, such as AI4ALL, Black in AI, QueersInAI, Inclusive AI.

perspectives help companies to tackle this issue better. Besides developers and engineers, companies should draw from disciplines such as law and social sciences, including moral philosophy, sociology, history, and gender studies.

The same report also points to several specific “AI ethics skills” that teams working with these concepts should have. Besides understanding several key concepts in this field, the report points to a lot of soft skills, including communication, negotiation, and relationship-building abilities across different functions in a enterprise [54]. A key challenge for companies will be develop a shared language that facilitates communication and collaboration. IEEE’s first edition glossary for ethically aligned design has 90 pages of related terms, with each term encompassing six definitions that spans across disciplines [55] .

Another challenge is specifically in regards to gender. Companies should pay close attention to the female perception in their teams while addressing gender bias in their systems. In Artificial Intelligence this may prove to be even more challenging, as women are still underrepresented in the field. In its 2019 Global AI talent report, Element AI encountered that women are only 18% of the authors in leading AI conferences [55] .

Leadership support is an important topic in the AI Ethics discussion.

IEEE’s report, *A Call to Action for Businesses Using AI* , offers also an “AI Ethics Readiness Framework”. This framework provides companies with a roadmap to assess their maturity in dealing with ethical issues in AI. One of the dimensions of this framework is leadership buy-in. Organizations could range from a place where “leadership recognizes but does not prioritize AI ethics” to a leadership that “champions AI ethic efforts” [54]. As already discussed, our interviews with practitioners also revealed the importance of leadership support.

Individual practitioners may feel constrained by their organizations.

Market dynamics create incentives for companies to "build AI faster than their competitors" [56]. This pressure is felt by practitioners, who experience direct consequences from it.

A study with 48 practitioners identified that "passionate individual advocates" frequently are the ones that raise questions about ethics in the AI development process, often causing them social costs [51]. A tutorial session during the latest ACM FAccT exposed similar realities. Based on 25 ethnographic interviews with practitioners, the tutorial shed light on how individuals working with AI ethics can develop stress and burnout, especially when they are doing this work voluntarily, without a specific role or proper organizational support [11].

Action Points

- If your company decides to have an in-house team tackling AI Ethics, ensure they are multidisciplinary, diverse and inclusive.
- Develop a shared and common language around fairness issues throughout your organization
- Consider assessing for fairness during your recruiting process, while training employees not familiar with the subject
- Ensure senior leadership is committed to fairness and to providing the adequate organizational resources to tackle this issue

Processes

Another important pillar for companies to develop skills is around processes. Establishing processes helps to create standards across the organization, institutionalizing approaches and making sure that fairness interventions are still deployed through time.

Companies have seen an attempt to develop several processes that can help mitigate bias in Machine Learning. Among those processes, checklists are one of the most popular formats available, with at least nine checklists available for practitioners to use [51]. Additionally, researchers paid close attention to documentation practices, suggesting ways that developers could document the developments of their model, helping in internal algorithmic audits.

Although there are still no standard industry processes to mitigate bias in Machine Learning, companies can leverage existing resources to create processes that are tailored to their context and industries.

Checklists

Checklists are a popular artifact to ensure process quality. In 2009, Atul Gawande wrote *The Checklist Manifesto*. The book shares several lessons from Gawande's experience in introducing the Surgical Safety Checklist, instrumental in decreasing surgery complications and deaths around the world. One of the key lessons from the book is that checklists are most powerful when they are not just a mechanism to tick boxes, but rather an artifact to embrace "a culture of teamwork and discipline" [57].

Therefore, In the context of mitigating gender bias in Machine Learning, checklists should be used to trigger the same type of teamwork and discipline proposed by

Gawande. Checklists should be used by teams to spark tough conversations around how a model can be more fair and how to deal with gender bias.

Checklists may have questions such as “Have we tested our training data to ensure it is fair and representative?” and “Have we studied and understood possible sources of bias in our data?” [10]. Teams need to be aware of falling into a trap of merely answering ‘yes’ or ‘no’ to such questions, instead of actually reflecting and revising their practices.

Another recommendation is to involve developers in the confection of the checklists, similar to how Gawande involved doctors and nurses in their process. A recent study co-designed an Artificial Intelligence fairness checklist with 48 practitioners, identifying the potential for providing “organizational infrastructure” around this subject [51].

Documentation & Auditing

Researchers and companies also invested in the development of documentation practices to address bias issues. Researchers created “Datasheets for Datasets”, a document that describe a database “motivation, composition, collection process, recommended uses, and so on” [58]. Google created “Model Cards”, a short document describing a model’s performance characteristics across several groups, helping to avoid possible unintended uses [59]. IBM introduced “FactSheets”, a document to increase Trust in Artificial Intelligence systems through the voluntary publication of a series of information regarding the purpose, safety, and other dimensions [60] . Finally, there are some initiatives to consolidate many of these documents, such as the SMACTR, an “end-to-end framework for internal algorithmic auditing” that incorporates some of the discussed documents [61] .

Choosing a process

Which process my company should pick? Given the high contextuality of every model and peculiarities of every company, there is no easy answer for this question. Companies must consider several dimensions, such as its corporate culture, how used to processes employees are, and how much resources can the company spend in developing processes. Ultimately, companies should choose a process that works. This means that is a process that accomplishes the fairness strategy of a firm.

For example, a B2B company may be more interested in having robust documentation practices to share them with their clients, who are other companies and may have the technical know how to interpret these documents. Conversely, B2C companies may decide to invest less in documentation and more in checklists during the development of a model, trying to be comprehensive in thinking about all its diverse customers.

Nevertheless, companies need to understand that any process for mitigating gender bias in Machine Learning cannot become a mere formality. The idea of having processes is to help a company address its strategic challenges, in our case in congruence with understanding fairness and aware of stakeholder needs.

Action Points

- Your company may seek inspiration in many existing processes to address fairness in Machine Learning
- Any process choice (or creation) should match your strategic goals and your company's resources
- Avoid turning processes into mere formalities. Any process should trigger critical thinking in teams, with the ultimate goal to create more fair systems.

Technology

Technological solutions

Over the past years, researchers developed several techniques and approaches to mitigate bias in Machine Learning. These approaches can be applied in different stages of the Machine Learning development pipeline, according to the source of bias. Bias in a model can come from virtually anywhere during its development, such as due to a underrepresented dataset, historical reasons, or when a system is deployed in unintended ways [62].

To tackle bias in Machine Learning, developers can rely on different bias mitigation algorithms. They can tackle the issue during pre-processing (correcting the training data used by a Machine Learning model), during in-processing (adjusting a classifier that is a dealing with a biased dataset), and at the post-processing stage (when the bias is corrected after a classifier is already trained) [13]. These algorithmics can be implemented through technical toolkits that address fairness in Machine Learning.

Build or Buy?

Do companies need to build their own technology to tackle gender bias in Machine Learning? Not necessarily. Although some companies have the resources and the strategic motivation to develop in-house skills to tackle this issue, there are alternatives. Companies may consider partnering with organizations such as the Algorithmic Justice League or the O'Neil Risk Consulting & Algorithmic Auditing (ORCAA), which provide algorithmic audits and consulting. Additionally, other companies may provide products to help companies monitor and audit algorithms, such as Arthur.ai.

Nevertheless, companies should understand how these alternatives integrate back to their businesses. Will the audits continue after the partner is gone? Who will be responsible for them? Wil the company adjust their models based on provided recommendations? All of these questions should be considered while deciding on whether to outsource technology or not.

Available toolkits

If a company decides to implement their own technology, they can count with numerous available open-source available toolkits. Developers continuously deploy technical toolkits – programming libraries – while developing software or Machine Learning models. These toolkits are recent and there are around ten of them publicly available [13].

Deciding for a technical toolkit is often the responsibility of the developers and engineers working in a project. However, companies may also need to address non-technical needs while choosing for which toolkit to use. For example, if a company already has scarce resources, it might want to opt for programming libraries associated with large companies, such as IBM's AI Fairness 360 or Google's TensorFlow Constrained Optimization (TFCO) programming library. These toolkits have probably a better chance of being maintained, even if they are open source.

Limitations

Companies should be aware of the limitations of these technical toolkits yet acknowledging how welcomed and important they are for measuring and mitigating bias in Machine Learning.

First, any technical intervention will quantify and simplify nuanced concepts, such as measuring fairness through fairness metrics [51]. Second, even if these toolkits are available, developers need to become familiarized with them and companies need to account for a learning curve while implementing them. Third, companies may have legal reservations about using open source software, especially if it involves chances of filing patents[9]. Finally, as research evolves and technology becomes more sophisticated, these toolkits may become obsolete and companies need to constantly keep up to date.

Action Points

- If your company does not have resources or a strategic reason to maintain an in-house ethical AI team, consider exploring alternatives and partnering with different organizations
- Ask your technical team to understand more about fairness in Machine Learning and share some of the existing toolkits with them
- Consider more dimensions before implementing a technical solution, such as legal considerations and a company's strategic priorities

[9] For a quick guide on the legal aspects of open source software, consult https://opensource.guide/legal/, prepared by Github

Gender Theory

Any Machine Learning (ML) model should be built with input from a subject matter expert. If a bank building an economical ML model consults with economists, companies working in mitigating gender bias in ML should pay attention to experts in gender theory, especially sociologists. Sociology, among other social sciences, brings important perspectives into the bias mitigation discussion, such as gender binarism and intersectionality.

Gender Binarism

There is a vast literature on theories of gender and sex[10]. A key concept from this field of inquiry is the idea that gender is defined by society, and not by biology. As a result, people who do not easily fall into the categories of female or male (such as transgender and non-binary people) may experience tensions in a society designed for gender binarism.

In the context of Computer Science and Machine Learning, this tension brought by gender binarism is visible in discussions of how to codify gender in systems. Often registration forms and computer systems request data about gender. Regardless of the actual need for gender data[11], companies may be faced with the decision of how to store that data.

The quick answer would be to simply institute two categories, male and female. Doing that may have negative consequences for people who do not identify with this simple categorization. Anthropologists refer to this experience as *torque*, “the individual experience of being twisted or pulled by classification systems” [63]. An

[10] Judith Butler is one of the most influential names in the field.

[11] For an interesting discussion on the need of gender, check Beyond Trans: Does Gender Matter? by Heath Fogg Davis.

example that addresses torque and expands away from gender binarism in systems is Facebook. The social network currently offers more than 50 options to its members to describe their gender [64].

Although most literature in mitigating gender bias in Machine Learning still treats gender as a binary construct, new initiatives are pushing this idea. During the Third ACM Conference on Fairness, Accountability, and Transparency (FAccT), for example, researchers developed a workshop to discuss human classification and the impacts of misclassification [63]. While more research is yet to be developed in this field, companies should be aware of this subject, engaging non-binary people and considering their needs.

Intersectionality

Companies should design ML systems considering women's multiple identities. These dimensions may include race, age, class, sexual orientation, gender identity, disabilities, and country of origin. For example, even if companies can build a model when comparing *women* and *men,* they may fail to achieve this fairness to racialized women or women in any other intersection of these dimensions.

Intersectionality as an analytical lens was cultivated by black feminism [65], mainly through the works of Kimberlé Crenshaw [66]. She explores "the race and gender dimensions of violence against women of color" and sheds light on how these identity dimensions should be considered simultaneously while understanding experiences and oppression faced by black women. In brief, we can think of intersectionality as how multiple levels of identity can amplify oppression when experienced simultaneously.

The Artificial Intelligence community is addressing intersectionality from multiple angles. Gender Shades exposed how the error rates in facial recognition software are higher for women, specifically women of colour [2]. Researchers also documented "overwhelmingly negative attitudes" towards Automatic Gender Recognition by transgender people [67]. Finally, researchers have hypothesised the various ways in which AI can negatively impact people with disabilities, such as voice recognition systems who may not work correctly for people with atypical speech or facial recognition software for people with different facial characteristics, such as people with down syndrome [68] . Besides raising awareness around intersectionality, researchers have also approached the problem by suggesting new metrics that address intersectional fairness. [69] [70].

Gender theory matters

Companies should make an effort to understand fairness issues through the lens of gender theory. Doing so provides critical subject matter expertise to tackle a complex problem of mitigating bias in Machine Learning. Concepts such as non-binarism and intersectionality may have practical implications to current work and should be considered by companies.

Action Points

- Treat gender theory as subject matter expertise while mitigating gender bias in your systems
- Consider concepts such as non-binarism and intersectionality while making decisions
- Incorporate concepts from sociology, anthropology, and other social sciences during analyses.

Integrating Concepts

Companies should act simultaneously in all three dimensions of our framework. Understand fairness, engage stakeholders, and building fairness skills all influence each other and play a role in mitigating gender bias in Machine Learning. For example, companies may change their fairness metrics based on the advice of their legal teams or in-house sociologists, while strong processes in place will make it easier for developers to implement technology that addresses these recommendations.

.

The following summary table may be useful to put all concepts together:

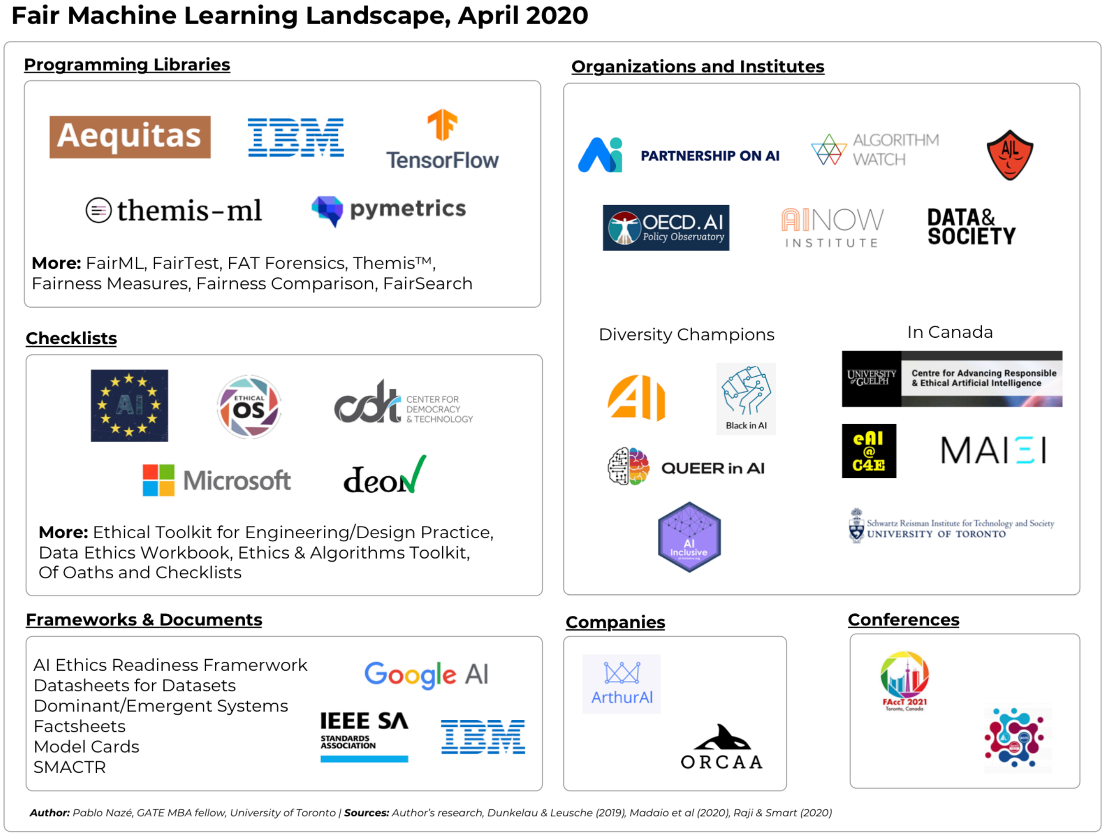
Table 1 – Framework Summary

		Key Actions
Understand Fairness		• Commit to understand fairness and its nuances; Embrace the many possibilities of fairness while refining your company's comprehension of the subject
Engage Stakeholders	Investors	• Use adjacent industries to make a business argument for mitigating gender bias in Machine Learning
	Regulators	• Make sure your legal team is aware of the latest Artificial Intelligence and industry-specific regulations
	End-users	• Be mindful of trade-offs between addressing customers' demands and increased development costs
	Civil Society	• Fully commit to fairness, avoiding ethics washing and technosolutionism
Build Fairness Skills	People	• Ensure senior leadership is committed to fairness and to providing the adequate organizational resources to tackle this issue
	Processes	• Avoid turning processes into mere formalities. Any process should trigger critical thinking in teams, with the ultimate goal to create more fair systems.
	Technology	• If your company does not have resources or a strategic reason to maintain an in-house ethical AI team, consider exploring alternatives and partnering with different organizations
	Gender Theory	• Treat gender theory as subject matter expertise while mitigating gender bias in your systems; consider concepts such as non-binarism and intersectionality while making decisions

Additional Resources

One of the outputs of this report was to map the current landscape of Fair Machine Learning, which can be viewed below.

Figure 3 – Fair Machine Learning Landscape, April 2020



Fair Machine Learning Landscape, April 2020

Programming Libraries

More: FairML, FairTest, FAT Forensics, Themis™, Fairness Measures, Fairness Comparison, FairSearch

Checklists

More: Ethical Toolkit for Engineering/Design Practice, Data Ethics Workbook, Ethics & Algorithms Toolkit, Of Oaths and Checklists

Frameworks & Documents

AI Ethics Readiness Framerwork
Datasheets for Datasets
Dominant/Emergent Systems
Factsheets
Model Cards
SMACTR

Organizations and Institutes

Diversity Champions

In Canada

UNIVERSITY of GUELPH Centre for Advancing Responsible & Ethical Artificial Intelligence
eAI @ C4E
MAIEI
Schwartz Reisman Institute for Technology and Society UNIVERSITY OF TORONTO

Companies

Conferences

Author: *Pablo Nazé, GATE MBA fellow, University of Toronto |* ***Sources:*** *Author's research, Dunkelau & Leusche (2019), Madaio et al (2020), Raji & Smart (2020)*

Besides reviewing this report's reference list, companies can consult additional reports and use cases:

- [IEEE: A Call to Action for Businesses Using AI]
- [McKinsey: Notes from the AI frontier: Tackling bias in AI (and in humans)]
- [Brookings: Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms]
- [Use Case: Building Representative Talent Search at LinkedIn]

Conclusion

The goal of this report was to provide business with actionable ways to think about gender bias in Machine Learning. We drew insights from a vast body of studies, working papers, and conferences, as well as 8 semi-structured interviews with practitioners. We shed light on the complexities of fairness, introducing concepts from a myriad of disciplines, ranging from computer science, to organizational theory, law, and sociology. Finally, we presented a framework – understand fairness, engage stakeholders, and build fairness skills – to help companies navigate those concepts. Given that fairness is very nuanced and highly contextual on application and industry, we did not offer silver bullets or definite techniques, checklists, or technical toolkits.

There have never been more materials published about this subject. Companies have a timely opportunity to take advantage of the latest research to tackle fairness issues and build a more equitable future. We hope this report is a useful map to help companies in their journey, always moving forward towards mitigating gender bias in Machine Learning.

References

[1] United Nations Development Programme, "Tackling Social Norms: A game changer for gender inequalities," UNDP, New York, 2020.

[2] J. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, New York, 2018.

[3] M. R. Costa-jussà, "An analysis of gender bias studies in natural language processing," *Nature Machine Intelligence,* pp. 495-496, 14 October 2019.

[4] T. Telford, "Apple Card algorithm sparks gender bias allegations against Goldman Sachs," The Washington Post, 11 November 2019. [Online]. Available: https://www.washingtonpost.com/business/2019/11/11/apple-card-algorithm-sparks-gender-bias-allegations-against-goldman-sachs/. [Accessed 30 March 2020].

[5] A. Jobin, M. Ienca and E. Vayena, "The global landscape of AI ethics guidelines," *Nat Mach Intell,* vol. 1, no. 9, p. 389–399, 2019.

[6] R. Chan, "Salesforce is hiring its first Chief Ethical and Humane Use officer to make sure its artificial intelligence isn't used for evil," Business Insider, 16 Dec 2018. [Online]. Available: https://www.businessinsider.com/salesforce-hires-paula-goldman-as-chief-ethical-and-humane-use-officer-2018-12. [Accessed 9 Mar 2020].

[7] Accenture Netherlands, "Responsible AI: with opportunity comes responsibility," Accenture, 17 Oct 2018. [Online]. Available: https://www.accenture-insights.nl/en-us/articles/responsible-ai-with-opportunity-comes-responsibility. [Accessed 9 Mar 2020].

[8] K. Holstein, J. W. Vaughan, H. Daumé, M. Dudik and H. Wallach , "Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?," in *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow, Scotland, UK, 2019.

[9] M. Hardt, E. Price and N. Srebro, "Equality of Opportunity in Supervised Learning," in *Advances in Neural Information Processing Systems 29 (NIPS),* 2016.

[10] D. Patil, H. Mason and M. Loukides, "Of oaths and checklists," 17 July 2018. [Online]. Available: https://www.oreilly.com/radar/of-oaths-and-checklists/. [Accessed 27 March 2020].

[11] B. Rakova, J. Yang and C. Chowdhury, *Tutorial: Assessing the intersection of Organizational Structure and ACM FAT* efforts within industry,* Barcelona: ACM Conference on Fairness, Accountability, and Transparency, 2020.

[12] J. P. Kotter, "Leading Change: Why Transformation Efforts Fail," *Harvard Business Review,* pp. 96-103, January 2007.

[13] J. Dunkelau and M. Leusche, *Fairness-Aware Machine Learning: An Extensive Overview,* Heinrich Heine Universität Düsseldorf. Working Paper Series: Fairness in Artificial Intelligence Reasoning, 2019.

[14] IBM, "AI Fairness 360 Open Source Toolkit," Sep 2018. [Online]. Available: http://aif360.mybluemix.net. [Accessed 9 Mar 2020].

[15] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton and D. Roth, "A comparative study of fairness-enhancing interventions in Machine Learning," in *ACM Conference on Fairness, Accountability, and Transparency*, Atlanta, 2019.

[16] A. Chouldechova, "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments," *Big Data,* vol. 5, pp. 153-163, 2017.

[17] Center for Data Science and Public Policy, The University of Chicago, "Aequitas," 2018. [Online]. Available: http://www.datasciencepublicpolicy.org/projects/aequitas/. [Accessed 9 Mar 2020].

[18] G. Harrison, J. Hanson, C. Jacinto, J. Ramirez and B. Ur, "An empirical study on the perceived fairness of realistic, imperfect Machine Learning models," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Barcelona, 2020.

[19] Starbucks, "Ethical Sourcing: Coffee," [Online]. Available: https://www.starbucks.com/responsibility/sourcing/coffee. [Accessed 27 March 2020].

[20] Nike, Inc., "FY19 NIKE, Inc. Impact Report," 2020.

[21] T. Whelan and R. Kronthal-Sacco, "Research: Actually, Consumers Do Buy Sustainable Products," 19 June 2019. [Online]. Available: https://hbr.org/2019/06/research-actually-consumers-do-buy-sustainable-products. [Accessed 27 March 2020].

[22] Ponemon Institute, "Cost of a Data Breach Report: 2019," 2019. [Online]. Available: https://databreachcalculator.mybluemix.net. [Accessed 27 March 2020].

[23] Georgian Partners, "The 11 Principles of Trust: How to Create Business Value Through Trust," [Online]. Available: https://georgianpartners.com/principles-of-trust/. [Accessed 27 March 2020].

[24] Strategic Cyber Ventures, "2018 Cybersecurity Venture Capital Investment," 16 January 2019. [Online]. Available: https://medium.com/@scv_group/2018-cybersecurity-venture-capital-investment-c50e1f25fe23. [Accessed 27 March 2020].

[25] McKinsey and Company, "Delivering through diversity," January 2018. [Online]. Available: https://www.mckinsey.com/business-functions/organization/our-insights/delivering-through-diversity. [Accessed 27 March 2020].

[26] K. Das, "The diversity industry is worth billions. But what do we have to show for it?," Fast Company, 22 October 2019. [Online]. Available: https://www.fastcompany.com/90419581/the-diversity-industry-is-worth-billions-but-what-do-we-have-to-show-for-it. [Accessed 27 March 2020].

[27] P. S. M. B. A. K. A. M. a. A. R. Muhammad Ali, "Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes," in *Proceedings of the ACM on Human-Computer Interaction*, New York, 2019.

[28] The Organisation for Economic Co-operation and Development (OECD) , "OECD AI Policy Observatory," 27 Feb 2020. [Online]. Available: https://oecd.ai. [Accessed 2020].

[29] High-Level Expert Group on Artificial Intelligence set up bt The European Commission, "Ethics Guidelines for Trustworthy AI," The European Commission, Brussels, 2018.

[30] The European Commission, "White Paper on Artificial Intelligence: a European approach to excellence and trust," Brussels, 2020.

[31] Government of Canada, "Directive on Automated Decision-Making," 1 Apr 2029. [Online]. Available: https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592. [Accessed 2020].

[32] Government of Canada, "Algorithmic Impact Assessment (AIA)," 31 May 2019. [Online]. Available: https://www.canada.ca/en/government/system/digital-government/modern-emerging-technologies/responsible-use-ai/algorithmic-impact-assessment.html. [Accessed 2020].

[33] The Office of the Privacy Commissioner of Canada , "Consultation on the OPC's Proposals for ensuring appropriate regulation of artificial intelligence," 28 Jan 2020. [Online]. Available: https://priv.gc.ca/en/about-the-opc/what-we-do/consultations/consultation-ai/pos_ai_202001/. [Accessed 2020].

[34] The U.S. Equal Employment Opportunity Commission, "Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures," *Federal Register,* vol. 44, no. 43, 1979.

[35] J. Sánchez-Monedero, L. Dencik and L. Edwards, "What does it mean to 'solve' the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Barcelona, 2020.

[36] U.S. Federal Trade Commission, "Your Equal Credit Opportunity Rights," Jan 2013. [Online]. Available: https://www.consumer.ftc.gov/articles/0347-your-equal-credit-opportunity-rights. [Accessed 24 Mar 2020].

[37] D. Pedreshi, S. Ruggieri and F. Turini, "Discrimination-aware data mining," in *KDD '08: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, Las Vegas, 2008.

[38] Pwc, "22nd Annual Global CEO Survey: CEOs' curbed confidence spells caution," https://www.pwc.com/gx/en/ceo-survey/2019/report/pwc-22nd-annual-global-ceo-survey.pdf, 2019.

[39] Forbes, "50 Stats That Prove The Value Of Customer Experience," 24 September 2019. [Online]. Available: https://www.forbes.com/sites/blakemorgan/2019/09/24/50-stats-that-prove-the-value-of-customer-experience/#344cf1974ef2. [Accessed 26 March 2020].

[40] Amazon, "Amazon Leadership Principles," [Online]. Available: https://www.amazon.jobs/principles. [Accessed 26 March 2020].

[41] CNBC, "Amazon beats Apple and Google to become the world's most valuable brand," 11 June 2019. [Online]. Available: https://www.cnbc.com/2019/06/11/amazon-beats-apple-and-google-to-become-the-worlds-most-valuable-brand.html. [Accessed 26 March 2020].

[42] R. Socher, "Why Ethical AI Is A Critical Differentiator," Forbes Insights With Intel AI, 27 March 2019. [Online]. Available: https://www.forbes.com/sites/insights-intelai/2019/03/27/why-ethical-ai-is-a-critical-differentiator/#5157dfeb63ab. [Accessed 26 March 2020].

[43] C. Lieber, "Topshop billionaire Philip Green is at the center of a #MeToo scandal," Vox, 25 October 2018. [Online]. Available: https://www.vox.com/the-goods/2018/10/25/18024504/philip-green-topshop-sexual-harassment-claims-metoo. [Accessed 26 March 2020].

[44] G. G. Fuster, *Tutorial: Gender: What the GDPR does not tell us (But maybe you can?),* Barcelona: ACM Conference on Fairness, Accountability, and Transparency, 2020.

[45] G. Dove, K. Halskov, J. Forlizzi and J. Zimmerman, "UX Design Innovation: Challenges for Working with Machine Learning as a Design Material," in *CHI '17: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, Denver, 2017.

[46] V. K. Rangan, L. Chase and S. Karim, "The Truth About CSR," Harvard Business Review, [Online]. Available: https://hbr.org/2015/01/the-truth-about-csr. [Accessed 26 March 2020].

[47] S. Kaplan, "Why the ‘business case’ for diversity isn’t working," Fast Company, 2 December 2020. [Online]. Available: https://www.fastcompany.com/90462867/why-the-business-case-for-diversity-isnt-working. [Accessed 26 March 2020].

[48] E. Bietti, "From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy," in *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Barcelona, 2020.

[49] N. Statt, "Google dissolves AI ethics board just one week after forming it," The Verge, 4 April 2019. [Online]. Available: https://www.theverge.com/2019/4/4/18296113/google-ai-ethics-board-ends-controversy-kay-coles-james-heritage-foundation. [Accessed 26 March 2020].

[50] S. Lindtner, S. Bardzell and J. Bardzell, "Reconstituting the Utopian Vision of Making: HCI After Technosolutionism," in *CHI '16: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, San Jose, 2016.

[51] M. A. Madaio, L. Stark, J. W. Vaughan and H. Wallach, "Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI," in *2020 ACM CHI Conference on Human Factors in Computing Systems (CHI 2020)*, Honolulu, Hawaii, 2020.

[52] The Algorithmic Justice League, "The Algorithmic Justice League," [Online]. Available: https://www.ajlunited.org/about. [Accessed 26 March 2020].

[53] The Algorithmic Watch, "The Algorithmic Watch," [Online]. Available: https://algorithmwatch.org/en/what-we-do/. [Accessed 26 March 2020].

[54] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (A/IS), "A Call to Action for Businesses Using AI," 2020. [Online]. Available: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead/ead-for-business.pdf. [Accessed 24 Feb 2020].

[55] The Glossary Committee, a Committee of The IEEE Global Initiative, "Ethically Aligned Design: First Edition Glossary (Draft form)," [Online]. Available: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e_glossary.pdf. [Accessed 27 March 2020].

[56] J. Clark and G. K. Hadfield, *Regulatory Markets for AI Safety,* arXiv preprint: arXiv:2001.00078 [cs.CY], 2019.

[57] A. Gawande, The Checklist Manifesto, Metropolitan Books, 2009.

[58] T. Gebru, J. Morgenstein, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III and K. Crawford, *Datasheets for Datasets,* arXiv preprint: arXiv:1803.09010 [cs.DB], 2018.

[59] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. R. Raji and T. Gebru, "Model Cards for Model Reporting," in *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency*, Atlanta, 2019.

[60] M. Arnold, R. K. E. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilovic, R. Nair, K. N. Ramamurthy, D. Reimer, A. Olteanu, D. Piorkowski, J. Tsay and K. R. Varshney, *FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity,* ArXiv preprint: arXiv:1808.07261 [cs.CY], 2018.

[61] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron and P. Barnes, "Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing," in *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Barcelona, 2020.

[62] H. Suresh and J. V. Guttag, *A Framework for Understanding Unintended Consequences of Machine Learning,* ArXiv preprint: arXiv:1901.10002 [cs.LG], 2019.

[63] A. Hanna, D. Baker and E. Denton, "Algorithmically encoded identities: reframing human classification," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020.

[64] The Mercury News, "Facebook offers more options for members to describe their gender," 13 February 2014. [Online]. Available: https://www.mercurynews.com/2014/02/13/facebook-offers-more-options-for-members-to-describe-their-gender-2/. [Accessed 27 March 2020].

[65] Black Feminisms, "Intersectionality 101: A Reading List," 2017. [Online]. Available: https://www.blackfeminisms.com/intersectionality-reading-list/. [Accessed 21 February 2020].

[66] K. Crenshaw, "Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color," *Stanford Law Review,* vol. 43, no. 6, pp. 1241-1299, 1991.

[67] F. Hamidi, K. M. Scheuerman and M. S. Branham, "Gender Recognition or Gender Reductionism?: The Social Implications of Embedded Gender Recognition Systems," in *CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Montréal, 2018.

[68] A. Guo, E. Kamar, J. W. Vaughan, H. Wallach and M. R. Morris, *Toward Fairness in AI for People with Disabilities: A Research Roadmap,* arXiv preprint: arXiv:1907.02227 [cs.CY], 2019.

[69] G. Morina, V. Oliinyk, J. Waton, I. Marusic and K. Georgatzis, *Auditing and Achieving Intersectional Fairness in Classification Problems,* arXiv preprint: arXiv:1911.01468 [cs.LG], 2019.

[70] J. R. Foulds, R. Islam, K. N. Keya and S. Pan, *An Intersectional Definition of Fairness,* arXiv preprint: arXiv:1807.08362 [cs.LG], 2018.

[71] B. Marr, "What Is The Difference Between Artificial Intelligence And Machine Learning?," Forbes, 6 December 2016. [Online]. Available: https://www.forbes.com/sites/bernardmarr/2016/12/06/what-is-the-difference-between-artificial-intelligence-and-machine-learning/#19b0b9142742. [Accessed 30 March 2020].

Appendix – Interview Details

Format: Semi-structured interviews, conducted in person or over the phone, lasting from 30 to 90 minutes.

Sample: 8 practitioners of data science or Machine Learning. Practical experience ranged from a few months (students) to years

Interview Protocol:
Hello, my name is Pablo and I'm conducting an project about gender bias in Artificial Intelligence and Machine Learning as part of my fellowship at the Institute for Gender and the Economy (GATE). Your identity and all of your answers will be kept confidential and anonymous .

Guiding Questions:

How long have you been working with AI/ML?
Could you tell me about the main technology stack that to you use?
Which subdomain(s) of AI/ML are you most comfortable working in?
Could you share with me your experience working in a AI/ML project that used data about people?
Could you tell me about a project where gender was one of the variables? Could you tell me more?
Why do you think gender was one of the variables? How did it impact your work? Why? Why not?
How did you measure the outcomes of that project? Which variables did you use to analyze the results? Why? Why Not?
Now I'd like you to imagine the following situation. You're working developing/managing a recruiting tool for a big company, and the media cracked a story that a group (race, gender, etc) of candidates were being systematically discriminated. How would you approach this situation? What would it take to solve this situation?
Now imagine that another company hired you to create/manage an internal recruiting tool. They asked explicitly for the tool not to discriminate against any group, but yet to keep a very high performance in terms of accuracy of prediction. Which steps would you take to start working on this?
Some researchers created a checklist to tackle bias in a project. (Presents checklist) What are your thoughts? What do you think it would take for it to be succesfully implemented by a company?
Are you familar with any debiasing technique or fairness metric you'd like to share?
Which methods do you think companies can use to decrease gender bias in AI/ML?
Do you have any questions for me?