Women's World Banking

Algorithmic Bias, Financial Inclusion, and Gender

A primer on opening up new credit to women in emerging economies

Sonja Kelly and Mehrdad Mirpourian
Women's World Banking

February 2021

Contents

Using the PDF navigation

To get to a segment quickly, you can click on the navigation menu on the top of the page. This feature will appear on most pages.

Acknowledgements

This report would not have been possible without our encouraging and patient colleagues across the industry.

In particular, we thank our key informant interviewees. We cannot name them because we promised anonymity, but their insights drove our inquiries. Second, our colleagues at Women's World Banking were instrumental in ensuring our work was intelligible to a broad audience. Finally, we are grateful for landmark support from Visa Foundation, without which we would not have had the freedom to pursue this topic.

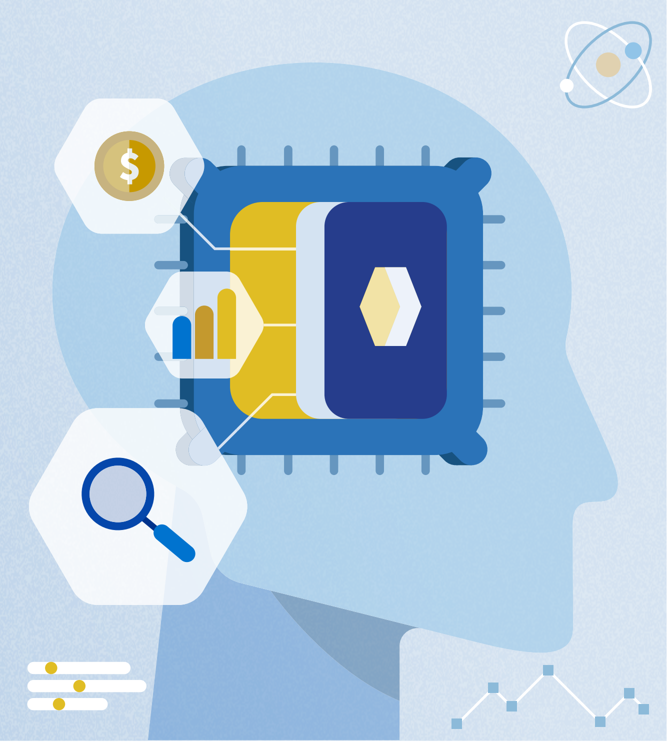
Introduction

Artificial intelligence and machine learning are changing financial services offerings to customers around the world.

In many markets, lenders are finding new ways of accelerating offerings to individuals and businesses, making decisions based on alternative data. The use of new data sources opens up new possibilities for thin-file customers. For women, who have historically been the victims of unconscious bias in lending decisions, algorithm-enabled credit decisions could create a level playing field. Do artificial intelligence (AI) and machine learning (ML) deliver on their promises to women customers? This report is part thought experiment and part primer, exploring the promises and pitfalls of using digital tools to open up new credit to women individuals and entrepreneurs.

Nearly every industry has considered leveraging computing power to improve its analytical toolkits. With the right data and under the right conditions, AI and ML can create a doorway to a more inclusive, fair, efficient future. Moving into this future (and not a dystopian science fiction storyline) requires active self-reflection, caution, and shared learning. It is no wonder that Alphabet (previously Google), Amazon, Apple, Facebook, IBM, Microsoft, and others are actively sharing their mistakes and successes as they innovate to improve consumers' experience with their products.

From these examples, we see glimpses of bias emerging across a diversity of industries. Amazon's recruitment AI learned bias against women applicants as it mimicked and amplified the decision-making of human HR representatives screening resumes. The tool was quickly shut down when it became clear that women applicants were systematically screened out of the list of viable candidates. Microsoft's Twitter chatbot, "Tay," was meant to demonstrate the ability of a machine to learn to talk seamlessly with humans. Instead, and within hours, it became a Hitler-saluting racist account, and was rightly asked to leave the platform soon after. We also know about instances of discrimination within the speech recognition systems that many of us have in our homes or in our handbags. Technology from Alphabet, Amazon, Apple, IBM, and Microsoft misidentified 35 percent of words from U.S.-born African-American users. Facial recognition software is ten times less accurate in identifying dark-skinned faces, with worrying implications for false positives when trying to identify a criminal suspect.



The financial sector is in the business of balancing risk and reward — as it should be. Many other industries pursue fairness as a rule of thumb. One economist we talked with as part of this research pointed out that the healthcare sector, for example, treats people based on need rather on whether they can pay for the service. In the financial sector, however, using data to distinguish "good" from "bad" credit risk among applicants allows financial institutions to approximate the risk they take on with each borrower, and sometimes to set a risk-based interest rate in response. Improved accuracy in these processes increases the efficiency of the lending process, making the institution more competitive and identifying the most appropriate premium for consumers. Some preferences are necessary and expected in order to make these processes work, since loan officers do not have a crystal ball for predicting the future.

We explore two questions here: First, where does gender-based bias originate? Second, how do we mitigate such biases in the emerging digital credit space? We hope that with these observations and other inputs, our industry can continue to learn from our mistakes and leverage new data sources for women's financial inclusion going forward.

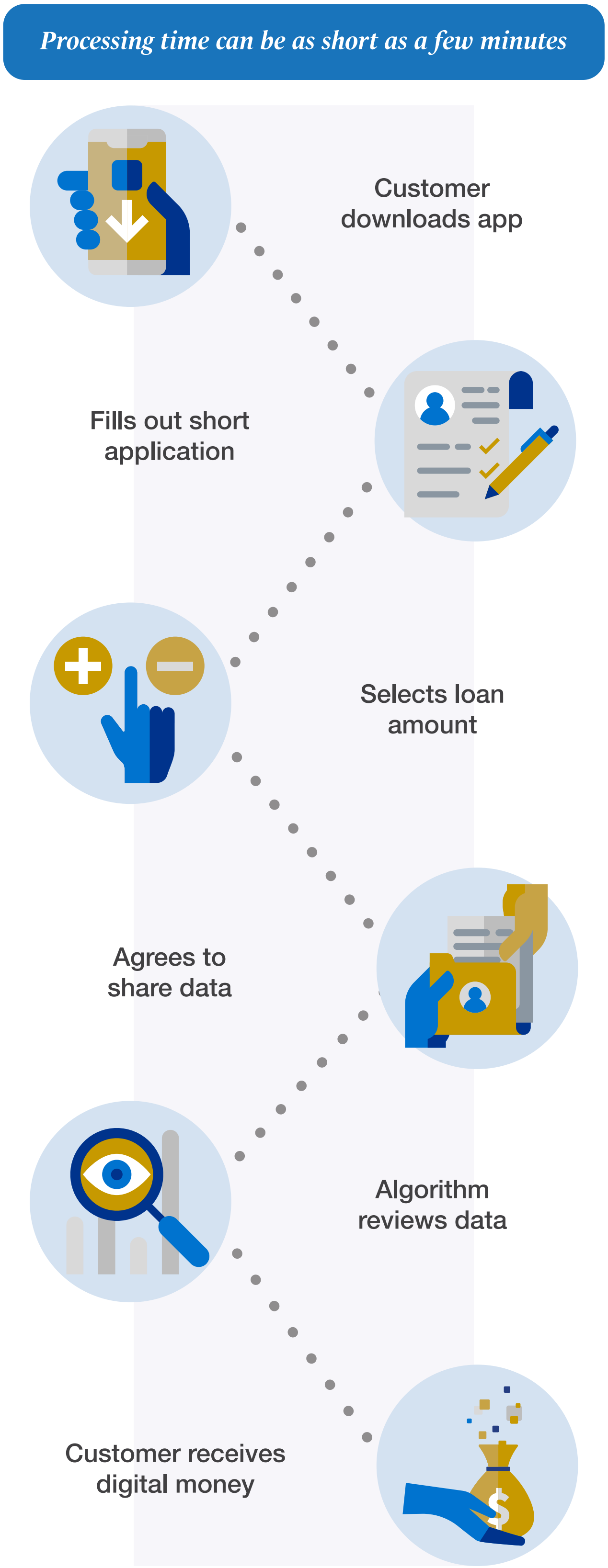
Emerging Credit Platforms

One reason this topic is important is because of the fintech companies that have emerged across the world to offer digital credit to consumers. Just over five years ago, entrepreneurs began to pilot this technology, which collects the data footprint of a user with a mobile phone. With this data[1], along with information on repayment, app-based companies built algorithms that approximated creditworthiness.

These algorithms used behavioral patterns such as whether a user capitalized the first letter of her contacts; whether the user engaged in gambling; what kind of phone the user had; what the available data was on the phone; and how many hours per day the borrower spent at his business[2]. The institutions used these data sources to make decisions about whether to offer a first loan to "thin-file" customers — people who would not be likely to have a traditional credit score. After the repayment of the first loan, companies had a de facto credit history on which to make subsequent decisions. We show the customer's perspective on an app-based digital credit journey in Figure 1.

Today, variations on this process offer loans for household consumption-smoothing and capital inputs for small and medium-sized businesses, and they have expanded beyond smartphone apps. Mainstream financial institutions and fintech startups alike leverage AI and ML to create and improve their credit-rating systems. Data is harvested from any internet-connected platform, including phones, tablets, computers, and other devices. You can see examples of this data collected based on our interviewee's responses and app-store disclosures in Figure 2 and Box 1. Data truly is the "new oil" in the financial system, and we are seeing thousands of "drilling rigs" emerging to capture this resource. Many consulting firms and research institutions have described this exciting emerging landscape, so we do not go into depth on the shape and scale of the industry in this publication. Nevertheless, we predict that it will only continue to grow as individuals, businesses, and financial institutions adapt to a far more digital world in response to Covid-19-related restrictions globally.

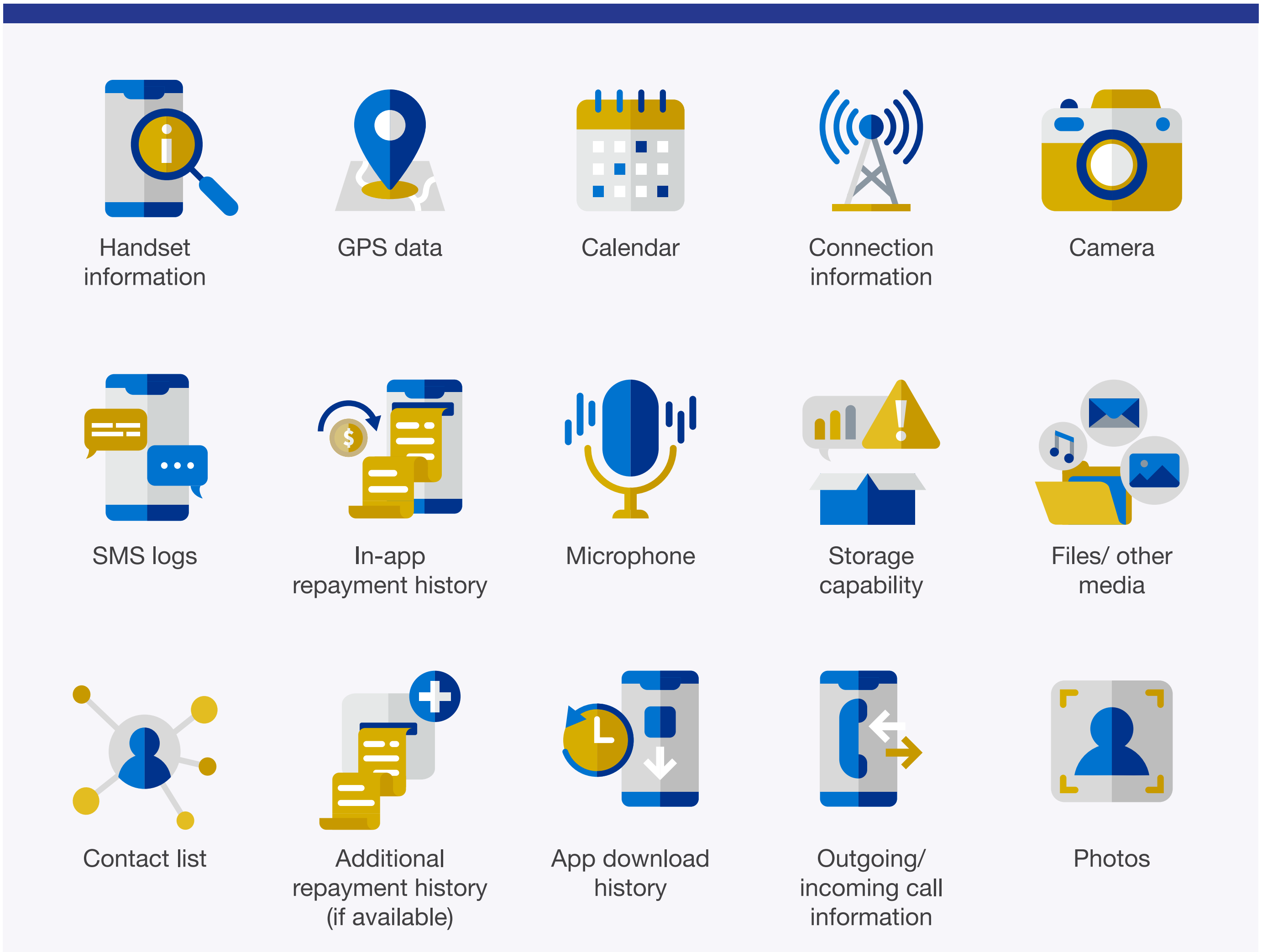
Figure 1: Following an app-based digital credit journey


Processing time can be as short as a few minutes
Customer downloads app
Fills out short application
Selects loan amount
Agrees to share data
Algorithm reviews data
Customer receives digital money
$

[1] *With apologies to the grammar police, we choose to make the word "data" singular throughout, as we are often talking about data as an idea rather than a particular collection of "datum."*

[2] *These are all variables that companies reported to be significant predictors of creditworthiness at some point in the life of their algorithm.*

Figure 2: Sample data collected by online or app-based digital credit companies



Our key informant interviewees were quick to remind us that the diversity of channels through which companies collect data on customers creates a diversity of data on which to base credit decisions. Any data can leverage AI and ML analysis techniques as long as the data set is large enough. We wanted our work to be applicable to a range of data types. We also wanted our applied research to be applicable to a range of institution types, and reflective of the growth in this industry. Accordingly, while at times we draw examples from the app-based digital credit industry, this report and the accompanying interactive model described in Box 2 are applicable and relevant to a broad range of institutions.

Box 1: Data collected by app-based digital credit companies

88%	Read the contents of your USB storage	PHOTOS/ MEDIA FILES
81	Modify or delete the contents of your USB storage	
38	Retrive running apps	DEVICE & APP HISTORY
6	Read your Web bookmarks & history	
75	Read phone status and identity	DEVICE ID & CALL INFO
88	Precise location (GPS & network based)	LOCATION
75	Approximate location (network-based)	
13	Access extra location provider commands	
75	Read phone status & identity	PHONE
31	Directly call phone numbers	
13	Read call log	
63	Take pictures and videos	CAMERA
75	Read the contents of your USB storage	STORAGE
69	Modify or delete the contents of your USB storage	
44	Read your text message (SMS or MMS)	SMS
38	Receive text messages (SMS)	
25	Read calendar events plus confidential information	CALENDAR
19	Add/modify calendar	
38	Find accounts on the device	IDENTITY
6	Read your own contact card	
6	Add or remove accounts	
63	Read your contacts	CONTACTS
38	Find accounts on the device	
13	Modify your contacts	

WI-FI CONNECTION INFO	View Wi-Fi connections	56%
MICRO PHONE	Record audio	13
OTHER	View network connections	100
	Full network access	100
	Receive data from internet	94
	Prevent device from sleeping	94
	Control vibration	44
	Run at startup	44
	Draw over other apps	25
	Connect and disconnect from Wi-Fi	25
	Change network connectivity	19
	Create accounts & set passwords	19
	Read Google service configuration	19
	Pair with Bluetooth devices	13
	Use accounts on the device	6
	Toggle sync on and off	6
	Control flashlight	6
	Read sync settings	6
	Access download manager	6
	Interact across users	6
	Change system display settings	6
	Modify system settings	6
	Read sync statistics	6

Source: Google Play Store, with authors' analysis conducted using a convenience sample of 16 of the largest digital credit applications including Branch, Easy Paisa, Eazzy Bank, GetBucks, KCB M-Pesa, L-Pesa, MoKash, M-Pawa, M-Pepea, Okash, PesaFlash, PesaZone, Saida, Stawika, Tala, and Timiza Loan. Current as of December 2020.

Research Methodology

We took two different approaches to this research. First, as this was Women's World Banking's first journey into this topic, we conducted key informant interviews with thought leaders and practitioners across the digital credit space. Interviewees included data scientists, digital finance experts, academics, entrepreneurs, app developers, and coders. We talked primarily to people familiar with the financial inclusion, digital finance, or mainstream finance worlds, but we did also include some interviewees with proficiency in other industries as well. We include findings from these interviews throughout, along with anonymous quotes from interviewees.

Second, we created an interactive tool using synthetic data to explore various bias scenarios. We studied how these biases affect credit decisions about customers, and what the business impacts of these biases might be. This exploration gave our project team a series of practical demonstrations of how biases affect underrepresented segments of customers, how a range of off-the-shelf algorithms treat different customer segments, and how the short-term costs and long-term benefits of fair algorithms impact business objectives. To share our findings with Women's World Banking's audience, we decided to make the tool publicly available. You can find this tool on the Women's World Banking website and use its Python code on the Women's World Banking GitHub page (see Box 2 on page 22 for more details).

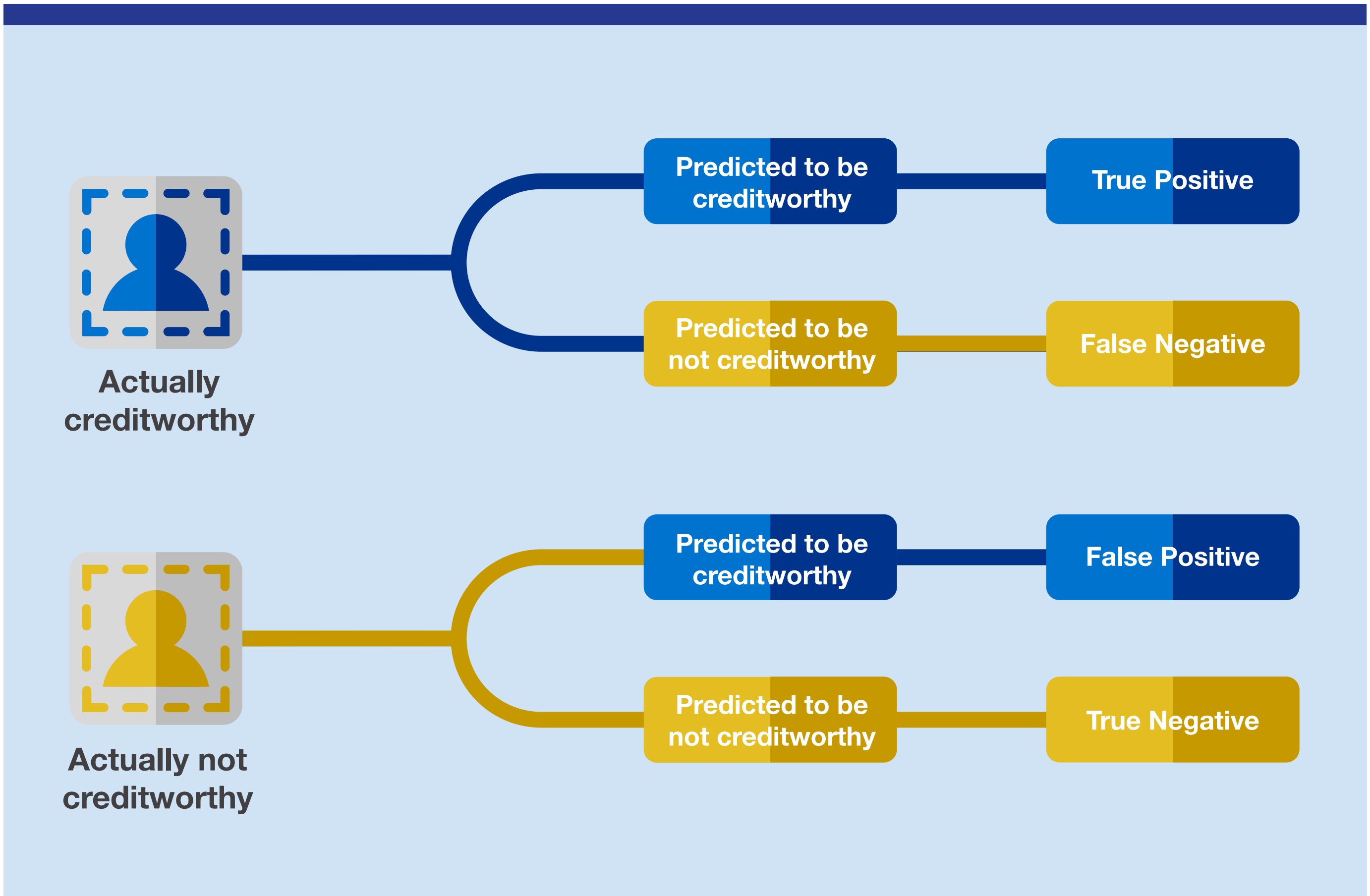
These two methodologies serve as Women's World Banking's introduction to the intersection of algorithmic bias, financial inclusion, and gender. We hope this paper is only a start as we move forward in exploring additional applications of this work to Women's World Banking's network, portfolio companies, and advocacy strategy.

Thoughts on Fairness

Fairness is an intricate and multidimensional concept, and its definition depends on context and culture. It is impossible to give one specific definition of fairness that applies to all organizations' use cases, so here we will explore a set of definitions[3]. For the sake of considering gender-based bias in lending, the important thing is that financial institutions have examined and adopted a definition of fairness as they balance fairness and efficiency in their credit operations.

Most fairness definitions rely on the four segments of a confusion matrix (Figure 3). The stylized confusion matrix we have depicted here compares predicted and actual events. In credit assessment, if someone is predicted to repay the loan and actually repays it, the result is a "true positive." If they are predicted to repay the loan but do not repay it, the result is a "false positive," and so on.

Figure 3: Confusion matrix applied to credit scoring


Actually creditworthy
Predicted to be creditworthy
True Positive
Predicted to be not creditworthy
False Negative
Actually not creditworthy
Predicted to be creditworthy
False Positive
Predicted to be not creditworthy
True Negative

Statistical definitions of fairness rely on the likelihood of a prediction and its outcome falling into one of these four groups based on past performance. Fairness definitions typically fall under three main categories: statistical measures, similarity-based measures, or casual reasoning. We focus on statistical measures and similarity-based measures, which are most relevant to the approach we took in studying algorithmic bias in credit scoring. These fairness definitions are shown in Table 2.

[3] *In fact, "Arrow's impossibility theorem" says that it is impossible to satisfy all of the constraints that definitions of fairness ask at the same time.*

Table 2: A selection of definitions of fairness

Type of measure	Type of fairness	Example
Statistical Measures	**Statistical parity:** Subjects in protected and unprotected groups have an equal probability to be in the positive predicted class.	Male and female loan applicants have an equal chance of having a good predicted credit score.
	Conditional statistical parity: The same definition as statistical parity while controlling for a set of factors.	Male and female applicants have an equal chance of having a good predicted credit score, controlling for their credit history, income, or other factors.
	Predictive parity: Both protected and unprotected groups have an equal positive predicted value.	The probability that an applicant with a good predicted credit score actually has a good credit score should be the same among both male and female applicants.
	False positive error rate balance: Both protected and unprotected groups have an equal false positive rate.	The probability of incorrectly assigning a good predicted credit score to an applicant with an actual bad credit score is the same for both male and female applicants.
	False negative error rate balance: Both protected and unprotected groups have an equal false negative rate.	The probability that an applicant with an actual good credit score is assigned a bad predicted credit score is the same for male and female applicants.
	Equalized odds: Protected and unprotected groups have equal true positive and equal false positive rates.	The probability of correctly assigning an applicant with an actual good credit score, and the probability of incorrectly assigning an applicant with actual bad credit score, is the same for both male and female applicants.
	Conditional use accuracy equality: Equal positive predicted value as well as negative predicted value.	Male and female loan applicants from both positive and negative predicted classes have equal accuracy.
	Treatment equality: The false negative to false positive ratio is the same among protected and unprotected groups.	The false negative and false positive ratio is the same among protected and unprotected groups.
	Test-fairness: For any predicted probability score, individuals in protected and unprotected groups have equal probability to be in positive class.	For any given predicted probability score S, both male and female applicants have equal probability of having actually a good credit score.
	Well-calibration: A probability, “S” percent, of applicants should have a good credit score, if a classifier assumes that a group of applicants have a certain probability S of having a good credit score.	If a set of applicants have a certain probability of having a good predicted credit score, let’s say 10, means that 10 percent of applicants indeed have an actual good credit score.
	Balance for positive class: Individuals from protected and unprotected groups with an actual positive class have an equal average predicted probability score of S.	One group of applicants — women or men — with a good credit score would consistently receive a higher score than applicants with a good credit score from the other group.
	Balance for negative class: Individuals from protected and unprotected groups with an actual negative class have an equal average predicted probability score of S.	The expected value of probability assigned by the model to male and female loan applicants with a bad actual credit score is the same.

(Continued from page 11) Table 2: A selection of definitions of fairness

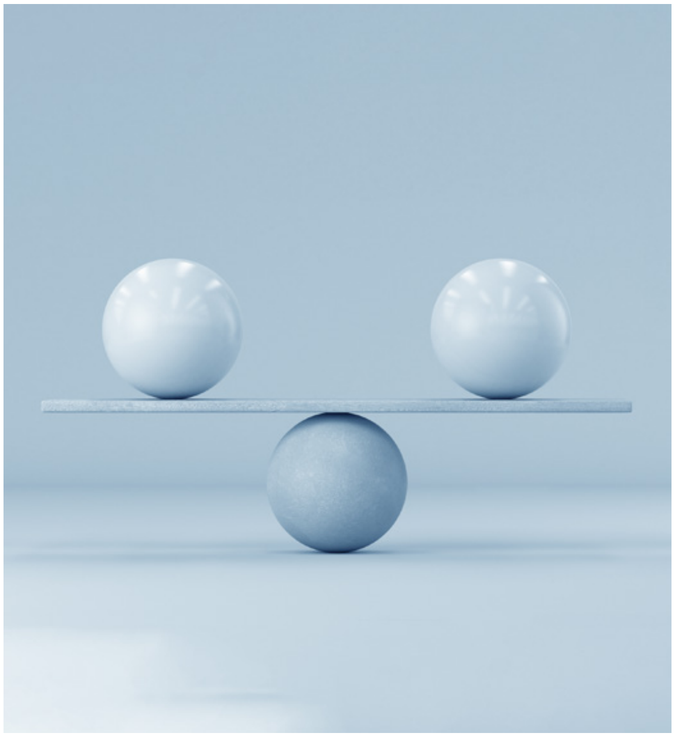
Type of measure	Type of fairness	Example
Similarity -Based Measures	**Causal discrimination:** The classification of any two subjects with the exact same attributes is the same.	A male and female applicant who otherwise have the same attributes are either both assigned a good credit score or both assigned a bad credit score.
	Fairness through unawareness: Sensitive attributes are not used in the decision-making process.	No gender-related features are included when training the algorithm.
	Fairness through awareness: Similar individuals should have a similar classification and distance metric.	The distance metric between two applicants would be 0, if all attributes except gender are identical, and 1 otherwise.

Source: Authors' elaboration based on classification by Verma and Rubin (2018).

Origins of Bias in Algorithms

Bias can refer to any form of preference. In this section, our focus is on unfair bias or discrimination. Discrimination happens when some prioritized groups receive a systematic advantage (being offered credit, for example) and other groups are placed at a systematic disadvantage (being denied credit, for example). Discrimination can be based on race, religion, language, gender, nationality, age, sexual orientation, and other categories. In this section, we focus on gender bias.

Gender-based algorithmic bias happens when an algorithm creates results that are systemically prejudiced against people, with gender explaining the bias. Algorithms, created and run by machines, can be biased just as humans can be. Algorithmic bias usually stems from conscious or unconscious prejudices introduced by the individuals — data scientists, coders, developers, or others — who create the algorithms. There are two ways algorithms might mirror an individual's unintended cognitive biases or real-life prejudices. As one of our interviewees puts it, "Algorithms were written by people, and people come to the desk with a preexisting bias that will get coded in." First, the algorithm itself can be biased because of how it is constructed. Second, an individual could introduce biases because he/she uses incomplete, faulty, or prejudicial data sets as the input that "trains" the algorithm. Biases might include:



"Algorithms were written by people, and people come to the desk with a preexisting bias that will get coded in."

Sampling Bias

With sampling bias, one population is overrepresented or underrepresented in a training data set. An example of this would be a digital credit app in a market in which men are more likely than women to have smartphones. If all of the customer data is used to train the algorithm, the algorithm will rely more on men's data than on women's.

Labeling Bias

Labeling is how data scientists annotate and classify certain properties and characteristics of a data point in order to make it searchable by an algorithm. An example of this would be labeling loan applicant occupations as "doctor" versus "nurse" rather than as "healthcare worker." Doctor and nurse would quickly become proxies for gender among loan applicants, whereas "healthcare worker" would continue to mask gender.

Outcome Proxy Bias

Outcome proxy bias occurs when the machine learning assignment is not well-defined. For example, if an algorithm uses address of residence as a proxy to predict the likelihood of credit default, the decision-making suffers from outcome proxy bias. The data is biased because default might be higher in neighborhoods with lower incomes, but this correlation does not guarantee the individual will default.

Once an algorithm is biased, it can deploy these biases at scale or evolve to amplify bias over time.

Locating & Mitigating Bias Through Algorithms

Locating and mitigating bias is not as easy as deleting the column labeled "gender" in a data set. Implicit biases still remain because of sampling bias, labeling bias, outcome proxy bias, or other biases, creating a situation in which women are unfairly more likely than men to fall into the "false negative" category in the confusion matrix. In other words, creditworthy women may be more likely to be denied credit than creditworthy men. These biases are not always visible at first glance, but they are major challenges to pursuing fairness through machine learning and artificial intelligence. One financial inclusion thought leader we talked with simply noted, "There are very few sources of data that would not be vulnerable to bias." Technologies do not ignore biased data unless explicitly told to do so. We therefore turn now to a discussion about how to locate and mitigate biases by leveraging algorithms.

"There are very few sources of data that would not be vulnerable to bias."

Just as algorithms can systematically keep women from accessing credit, they can also ensure fairness for women borrowers. Like all technologies, algorithms are not good or bad — they merely amplify the intent of their user. Mitigating bias depends first on the kind of bias the data contains. For example, if the data contains labeling biases, the mitigation strategy would be different than if the data created a sampling bias. Here, we have assembled some mitigation strategies to get institutions started on locating and addressing bias[4].

[4] *We draw many of these suggested bias mitigation strategies from the AI Fairness 360 Toolkit, a widely cited and open-source library, written by IBM research, that contains techniques to detect and mitigate bias in machine-learning models. The tool is available in Python and R here: https://aif360.mybluemix.net/*

In general, there are three stages during which data and model fairness can be measured. These three stages are the pre-processing, in-processing, and post-processing steps, each of which has its own benefits and challenges. Addressing bias in one stage does not guarantee fairness, but it certainly increases the likelihood.

1 Pre-processing

Before data is even used to train an algorithm, there are a number of non-technical decisions that lead to particular mitigation techniques. For example, if the data is not representative, data scientists can use machine learning or artificial intelligence to re-weight the data. Using this technique, if there is sampling bias, the underrepresented population will not have a smaller effect on the algorithm's decisions. Similarly, in pre-processing, algorithms can detect and point out labeling biases.

2 In-processing

In-processing algorithms incorporate fairness into the machine learning training task itself. These methods put a "penalty" on undesired biases or add fairness constraints to the model. For example, an in-processing mitigation strategy might establish that women should be accepted at the same rate as men. Some of the most widely used in-processing algorithms remove prejudice or align the algorithm's decisions with a defined outcome.

3 Post-processing

Sometimes it is challenging to explain how a machine-learning model really works. This might be the case in "deep learning models" in which the machine makes decisions it does not, nor cannot, explain. For such cases, it is necessary to use post-processing algorithms. Post-processing algorithms usually focus on reducing bias by working on the model output predictions. Although applying these techniques is relatively simple, the challenge is in maintaining the model accuracy while reducing bias.

The challenge for institutions will be in fitting mitigation strategies to the data, and importantly, in using such mitigation to balance the fairness and efficiency of the model. Some of these mitigation strategies may only impact the credit-scoring model moderately, and some may have a massive effect. Institutions have to test mitigation strategies on their own data to understand the relative tradeoffs with respect to model efficiency.

Mitigating bias through algorithms is not the only way to address gender bias in machine learning and artificial intelligence. In the next section, we consider real-life examples of how institutions use their own operational processes, management, and norms to decrease bias in algorithm-based lending methodologies.

Locating & Mitigating Bias Through Operational Processes and Norms

Staff at all levels, and with varied degrees of sophistication in technical skills, can play a role in mitigating biases. The most successful institutions we talked with — ones that intentionally acknowledged the potential for bias and actively addressed it — were leveraging a range of strategies to combat gender bias in their credit decision-making. One lender put it this way: "We don't use gender in our models, but that's not enough to make sure we have fairness. We look for equal access, where every group has the same chance of approval, and fair pricing." Here, we share three overlapping mitigation strategies that rely more on operational processes and institutional norms:

First, the entire organization — not just the data science team — should be involved in mitigating bias. One institution we talked with that was shifting its credit underwriting from people to technology described its inclusive human resources strategy: "Every member of the bank has gone through an entire digital transformation journey. Every member of the bank has sat in meetings, classrooms, to learn how to transform into a digital thinking organization. Organizational culture involves a team of leaders that actually believe in what data has to offer." Furthermore, organization leadership should be able to objectively review and understand how credit decisions are made. One CEO we talked with described a situation in which a data scientist made a poor decision.

The CEO had to step in, but recognized that this situation was only extraordinary in that the bias was caught: "In my case, I didn't leave it up to the tech team or the developers. One of our developers is 23 [years old]. He has very little professional experience, and he made decisions on his own that were flat-out incorrect."

Most developers, coders, and data scientists are hired for their ability to create algorithmic efficiency, driving profit for the institution. Candidates for this highly technical role in an institution are unlikely to be hired from within the markets in which algorithms are being deployed. One of our interviewees pointed out that in most of the largest digital credit institutions serving low- and middle-income markets, nearly all of the data science team is based outside of the markets in which they work.In addition, these staff tend to be men. One institution we talked with lamented the limited number of women candidates for developer jobs: "In the technology team, we only have two women and we have 20 men. We really want to bring more women in, but it's really hard to find women who study computer science or I.T."

The conceptual distance between developers and low-income women in emerging markets — as well as the gap between the skills of algorithm developers and their ability to make decisions about fairness — necessitates leadership involvement in the conversation. It may also involve lower-skilled team members in the markets in which the models are being deployed. At all levels of the organization, an awareness of what bias is and a willingness to spot it might be critical to prioritizing bias mitigation on a technical level.

Second, the algorithm must be reviewed regularly as part of an operational process. One organization we talked with mentioned that in organizations that are heavily relying on algorithms for their credit underwriting, the lead data scientist is often on the management committee, influencing decisions about how much oversight the algorithms receive. A support organization we talked with indicated that the success of bias mitigation may be based on the commitment to fairness of individuals at the organization: "This stuff has to be institutionalized in some way. What if another data scientist joins? You may find nuanced thinkers in the institutions, but no written policies assuring us that as they scale, bias will continue to be a way to examine the data." A thought leader we talked with shared why they became interested in this topic: "I was talking to more and more lenders who didn't actually know how their scores worked."

With algorithm review, people can make decisions about alancing fairness and efficiency rather than leaving it up to the model. Among the strategies we heard was the value of performing quarterly checks that compare algorithm results by population sub-groups, to ensure that the distribution of scores across subgroups correlates to repayment rates (in other words, ensuring the distribution in the confusion matrix is equal in both men and women sub-segments). We also heard about the importance of conducting a strategic review between old and new models every time there is an update. One organization we talked with, aware of its own bias, looks at the top 20 variables that drive the score, checks to see if these variables are highly correlated with gender, and then decides whether it is comfortable with the bias. For example, if amount of time spent at the business location is highly correlated with gender, the institution might decide this is a valid variable, even though it emphasizes efficiency rather than fairness across gender lines.

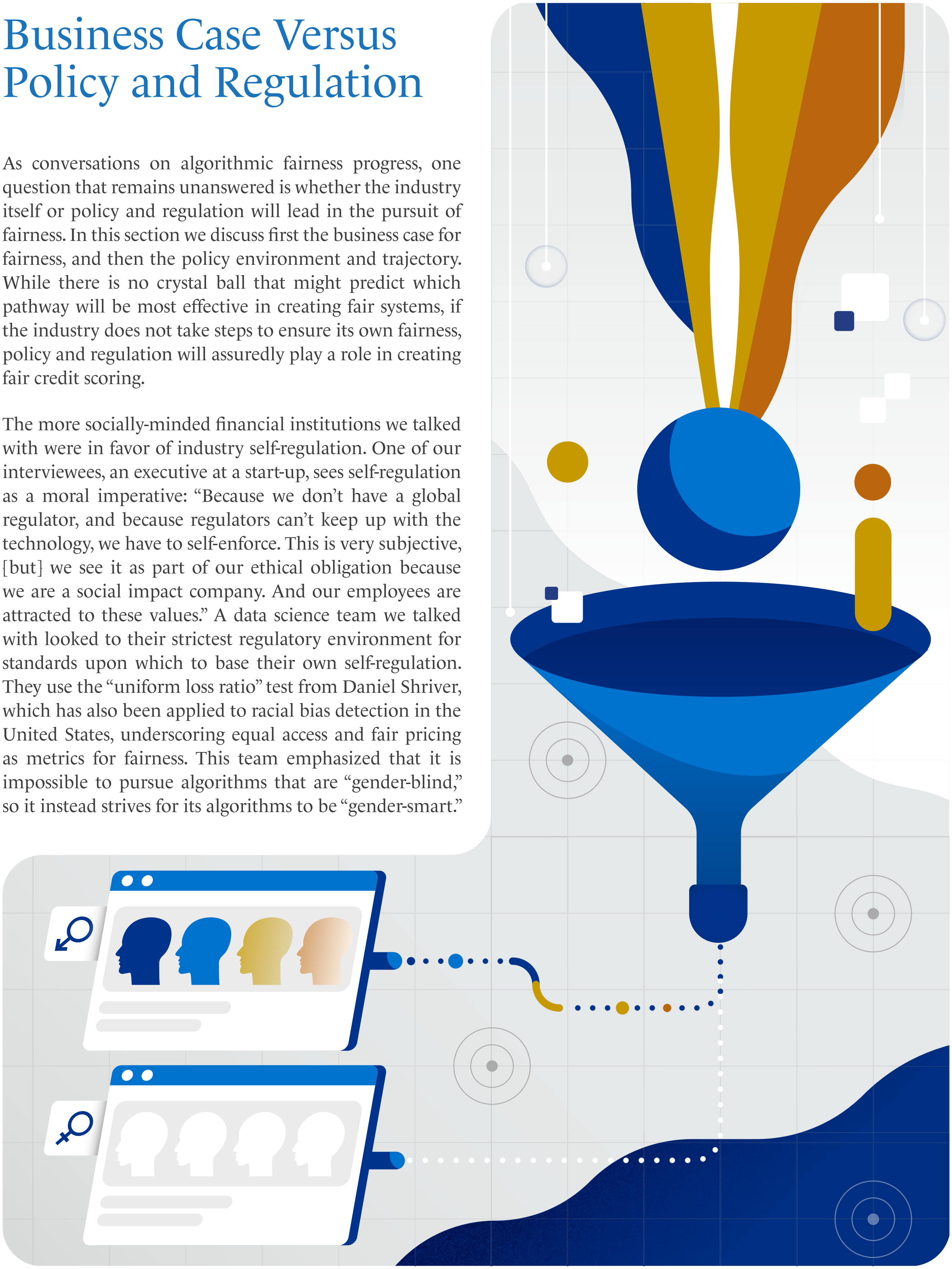
Third, the algorithm should be discontinued and reviewed in case of an observable systemic shock. Covid-19 is a perfect example of the necessity of this mitigation strategy. One lender we talked with was forced to suspend lending to all new customers, and to only focus on existing customers. The pre-Covid algorithms that the lender was using were completely ineffective at predicting creditworthiness post-Covid. Another lender we talked with said, simply, in reference to Covid-19, "Algorithms are no longer predictive of creditworthiness. We are basing decisions on behaviors that no longer exist." A third said, "Covid-19 is a protracted crisis for which these products were not designed." Under these circumstances, finding a balance between fairness and accuracy is impossible without retraining the algorithm. Once the algorithm is retrained, organizations can re-seek fairness.

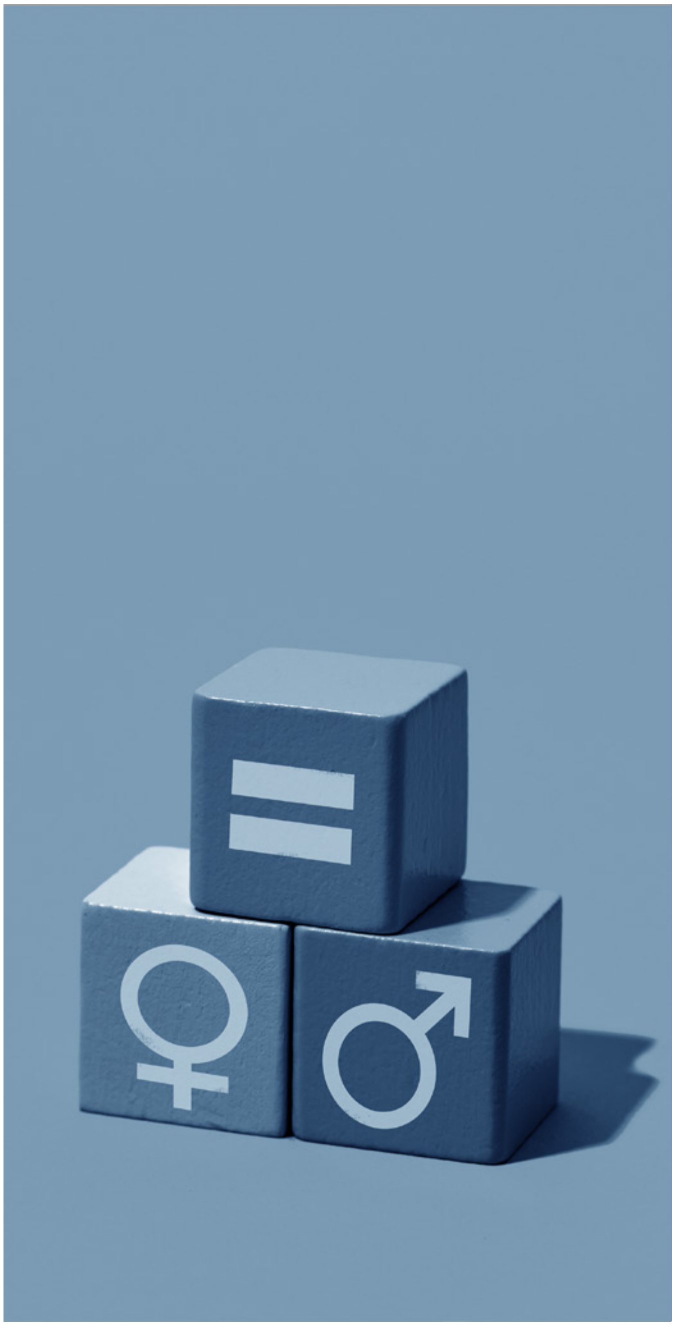
"I was talking to more and more lenders who didn't actually know how their scores worked."

Business Case Versus Policy and Regulation

As conversations on algorithmic fairness progress, one question that remains unanswered is whether the industry itself or policy and regulation will lead in the pursuit of fairness. In this section we discuss first the business case for fairness, and then the policy environment and trajectory. While there is no crystal ball that might predict which pathway will be most effective in creating fair systems, if the industry does not take steps to ensure its own fairness, policy and regulation will assuredly play a role in creating fair credit scoring.

The more socially-minded financial institutions we talked with were in favor of industry self-regulation. One of our interviewees, an executive at a start-up, sees self-regulation as a moral imperative: “Because we don’t have a global regulator, and because regulators can’t keep up with the technology, we have to self-enforce. This is very subjective, [but] we see it as part of our ethical obligation because we are a social impact company. And our employees are attracted to these values.” A data science team we talked with looked to their strictest regulatory environment for standards upon which to base their own self-regulation. They use the “uniform loss ratio” test from Daniel Shriver, which has also been applied to racial bias detection in the United States, underscoring equal access and fair pricing as metrics for fairness. This team emphasized that it is impossible to pursue algorithms that are “gender-blind,” so it instead strives for its algorithms to be “gender-smart.”





Social responsibility is not the only reason why providers might pursue more ethical algorithms. One interviewee we talked with said that one way to encourage companies to ensure their algorithms are not discriminating by gender would be to show that their bias is making them incorrectly assess the risk of female borrowers — and leading them to therefore miss out on an important consumer segment. Showing a lender that there is an untapped market in female borrowers could lead to a conclusion that additional female borrowers would improve the institution's bottom line. More gender-fair algorithms, therefore, would be a strategic rather than ethical imperative. Particularly in markets where the regulator might not have the capacity to regulate — or where regulating would hinder innovation in the industry — an ethical or business case argument may be the only option for achieving fairness.

Other interviewees we talked with see regulation as inevitable in markets where authorities have the capacity to create and enforce policy. One thought leader we talked with who works in Sub-Saharan Africa said, "If there's not a law, are you going to be anti-discrimination?" This interviewee was skeptical that the industry would be able to self-regulate. Examples of a legal approach include recent data privacy regulation or guidance in Kenya, Uganda, and Nigeria that draws on some of the principles in Europe's General Data Protection Regulation (GDPR). While there is regulation in this area, to date there is very little supervision in emerging markets. One additional jurisdiction concern is that many digital lenders do not sit under a financial regulator. One of our interviewees, a thought leader with a global view, asked, "Who would even ask to see an algorithm? There is no one at the moment. I wouldn't expect sophisticated regulations soon." Regulation in this area may be more likely to come from the authorities responsible for data protection, which is a growing but still very nascent regulatory environment in emerging markets.

Even with regulation, data fairness might still be elusive. In the United States, for example, gender is a protected classification, and lenders are not allowed to use gender when making credit decisions (just as they are not allowed to use race, sexual orientation, religion, or other factors). Keeping gender out of a model, however, will not automatically eliminate the three biases — sampling bias, labeling bias, or outcome proxy bias — we identified toward the beginning of this primer. One lender we talked with used this very defense when they insisted their algorithm is not biased: "To make sure our model isn't biased, we don't tell it whether the applicant is a man or a woman." This prevention technique is not enough in the pursuit of fairness.

Where to From Here?

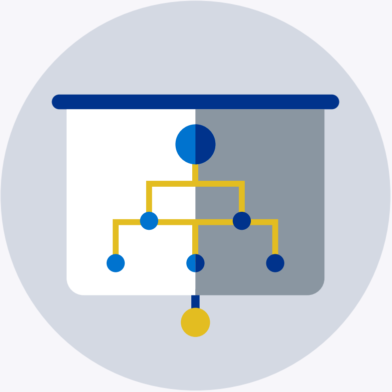
It might be easy to grow weary of pursuing fairness, given the complexity of the problem and the variety of solutions — which have varying levels of demonstrable success. There are two factors that should inspire hope. One factor is the volume of people around the world, across a diversity of industries, who are thinking about this same challenge. The second factor is the availability of technology solutions to the challenge of fairness. The same technology that might exhibit bias can be used to pursue and ensure fairness. In this primer we explained what fairness is, discussed where bias might emerge, and shared some mitigation strategies. Now we turn to practical suggestions for organizations looking to more actively balance algorithmic fairness and efficiency in their lending methodologies.

From a very high-level standpoint, we can summarize this task by dividing it into the following phases:

1 Build a fairness implementation team

This multidisciplinary team should bring a group of legal, business, and machine learning experts together. Legal advisors define what the legal constraints are or could be, identifying what the minimum threshold of compliance might be — and how to design for future regulation. Business experts think about what definitions of fairness fit well with their strategy. If there is a social mission to serve low-income customers, and women in the market are more likely to be low-income, institutions have a mission-based imperative to design for this segment. Finally, machine learning experts are the ones who deal with pre/in/post processing algorithmic solutions to algorithmic bias problems. They design or modify algorithms to satisfy the desired fairness metrics they received from business and legal experts. Although legal, ML, and business teams have their own specific tasks, collaboration ensures that models are examined on multiple levels.

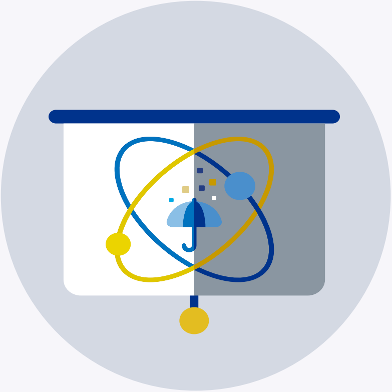
There is a business case for the development of this internal capacity. Fair algorithms provide win-win solutions in the mid- to long-term. Algorithms that focus on maximizing profit in the short-term might exclude segments of customers (filtering them into the false-negative category). Instead of allowing these excluded groups to serve as future sources of income for the organization, the unfair algorithm deprives both the organization and the customers from staying in a win-win equilibrium. On the consumer side, customers are becoming more socially aware and prefer to receive their products and services from like-minded organizations and institutions. As markets develop and customers have more choices, they may be willing and able to self-select more fair institutions.



2 Make strategic decisions

Leadership — whether executive leadership or a cross-functional task force — needs to put time and effort into defining where the organization falls on the spectrum between fairness and efficiency, and it needs to make these conscious decisions itself, rather than allowing data scientists or the algorithm itself to make them. Some of these decision points include:

- What is the universe of labels to include in customer data?
- Are there any protected attributes to keep from the model or to use to evaluate the fairness of the model?
- How much should these protected attributes matter?
- What strategies will the team use to evaluate whether the algorithm is staying within these defined parameters in its decision-making?
- How often will the data science team report back on the success of these mitigation strategies?
- What information does leadership need in order to make future decisions about the algorithm?



3 Implement mitigation strategies

Data scientists will be most involved at this step, but organization leadership must be able to articulate what these mitigation strategies are. Some questions leadership should be able to answer include whether mitigation strategies will focus on pre-, in-, or post-processing; who is responsible for bias mitigation; and what incentives are in place to ensure bias mitigation is prioritized.

Leadership should get updates in a format it can understand so it can assess how well this process is going, provide input at key decision points, and see how mitigation affects the business lines or the organization. Leadership should also know whom to celebrate when the institution has success in detecting and mitigating bias, and whom to blame when things go wrong.

4 Organizations must institutionalize these processes

It may be natural for a set of exemplary employees who care deeply about fairness to create a way of working on fairness. Unless these steps turn into clearly defined institutional and organizational processes, however, they could get deprioritized due to staff turnover and resource allocation.

5 Institutions that care about mitigating bias over the long term will increase representation at all levels

Hiring processes can seek out data scientists living and working in the same country as credit recipients. Institutions can hire for diversity of gender, age, race, or experience in order to encourage a range of perspectives among developers. This long -term capacity development will contribute to more fair and equitable algorithms, and it will also have the spillover effect of making progress on other organizational priorities.

Algorithmic bias is complicated, and requires multiple approaches to ensure the automated processes that improve efficiency do not translate into unfair treatment of women customers. The good news is that machine learning and artificial intelligence, while part of the problem, can also be part of the solution. Technology, along with effective management and organizational processes, provides new solutions for bias mitigation. We look forward to journeying into a solution together with the financial sector, opening up credit markets previously closed to many women customers.

Box 2: Want to play with your data? Apply these principles to your own institution.

Women's World Banking recently created a Python-based toolkit to show how financial services providers can detect and mitigate gender biases in credit score models.

The first step in the toolkit is a series of questions on portfolio size, sex ratios among clients, likelihood of women versus men applicants being extended credit, and a number of other factors. By asking these questions, the tool can model a particular institution's credit portfolio.

Next, based on user input, the tool creates a synthetic dataset for the user and provides insight on both bias detection and mitigation. Visit the tool at github.com/WomensWorldBanking.

Bias detection

It shows the user what kinds of gender biases are implicit in the dataset or credit score model.

Bias mitigation

Based on the biases within the model, the tool shows how the user can apply different methods to mitigate those biases. The tool demonstrates the fairness impacts of both pre-processing methods and in-processing methods for mitigation.

References

Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. "Fairness in criminal justice risk assessments: The state of the art." ***Sociological Methods & Research*** (2018): 0049124118782533.

Chouldechova, Alexandra. "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments," ***Big Data*** (2016).

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. "Algorithmic Decision Making and the Cost of Fairness," In Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining, pp. 797-806 (2017).

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. "Fairness Through Awareness," ***Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*** (2012).

Galhotra, Sainyam, Yuriy Brun, and Alexandra Meliou. "Fairness Testing: Testing Software for Discrimination," In Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering, pp. 498-510 (2017).

Hardt, Moritz, Eric Price, and Nati Srebro. "Equality of Opportunity in Supervised Learning," ***Advances in Neural Information Processing Systems*** (2016).

Kamiran, Faisal, and Toon Calders. "Data preprocessing techniques for classification without discrimination." ***Knowledge and Information Systems 33***, no. 1 (2012): 1-33.

Kilbertus, Niki, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. "Avoiding Discrimination Through Causal Reasoning," ***Advances in Neural Information Processing Systems*** (2017).

Kleinberg , Jon M., Sendhil Mullainathan, and Manish Raghavan. "Inherent Trade-Offs in the Fair Determination of Risk Scores," ***ITCS*** (2017).

Kusner, Matt J., Joshua R. Loftus, Chris Russell, and Ricardo Silva. "Counterfactual Fairness," ***Advances in Neural Information Processing Systems*** (2017).

Nabi, Razieh, and Ilya Shpitser. "Fair Inference on Outcomes" ***Association for the Advancement of Artificial Intelligence*** (2018).

Pearl, Judea. ***Causality: Models, Reasoning, and Inference***. New York: Cambridge University Press, 2000.

Simoiu, Camelia, Sam Corbett-Davies, and Sharad Goel. "The Problem of Infra-marginality in Outcome Tests for Discrimination," The Annals of Applied Statistics 11, no. 3, no. 3 (2017).

Verma, Sahil, and Julia Rubin. "Fairness Definitions Explained," ***IEEE/ACM International Workshop on Software Fairness (FairWare)*** (2018): 1-7.

Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. "Fairness Beyond Disparate Treatment and Disparate Impact: Learning Classification Without Disparate Mistreatment," ***Proc. of WWW'17*** (2017).

Zliobaite, Indre. "On the Relation between Accuracy and Fairness in Binary Classification," ***CoRR abs***/1505.05723 (2015).

Women's World Banking

@womensworldbnkg

www.womensworldbanking.org

Algorithmic Bias, Financial Inclusion, and Gender

A primer on opening up new credit to women in emerging economies