Contents

Executive Summary ........................................................................................................ 3
This Report .................................................................................................................... 4
Introduction .................................................................................................................. 5
1. Bias and Ethics in Machine Learning ....................................................................... 7
   1.1. Definitions .......................................................................................................... 7
   1.2. Types of Bias in ML ............................................................................................ 8
       1.2.1. Bias in data, collection and processing ......................................................... 10
       1.2.2. Other sources of Bias .................................................................................. 13
   1.3. Data Protection Laws and Bias ........................................................................ 13
2. Approaches for Ensuring Fairness ............................................................................ 14
   2.1. Governance Processes ....................................................................................... 16
   2.2. Technical approaches by Financial Institutions ................................................ 19
   2.3. Development of High-Level Principles ............................................................. 22
       2.3.1. Monetary Authority of Singapore (MAS) ..................................................... 23
       2.3.2. European Commission Ethics Guidelines for Trustworthy AI .................. 24
3. Case Studies ............................................................................................................. 26
   3.1. HSBC Case-Study: Techniques to Prevent Bias .................................................. 26
   3.2. Swiss Re Case-Study: AI and "Digital Governance Framework" (DGF) .......... 29
Conclusions ................................................................................................................... 31
Executive Summary

The adoption of Machine Learning (ML) in the financial sector has rapidly developed in recent years, and the technology’s capabilities offer substantial benefits for the industry, individuals and society. The ability for banks and insurers to improve the view of risk can result in better access to credit for customers.

However, these same technological implications raise concerns, including questions about how to ensure that important decisions made using ML are not discriminatory. This is because ML reflects historical data, influenced by human biases including unconscious ones. Given the automated nature of ML algorithms, the main concern is that misguided correlations could result in powerful implications, and inherent biases can produce data that amplifies biases already present in society.

In this context, the goal of ML is not only to maximize predictive performance but to do so subject to a fairness constraint. Financial institutions are in fact taking a cautious approach to the use of ML. They are implementing ML while considering data protection, security and integrity as part of their design process.

In fact, ML is already allowing firms to better understand and monitor the risks posed with each customer in a more holistic manner. Firms have been able to bridge data gaps, reduce information asymmetry that has led to better credit decisions, ultimately democratizing the access to financial services (see Box 2).

In this report, we address specific responses and solutions taken by financial institutions to ensure fairness. These include looking closely at the governance process for ML, as well as considering technical measures taken by many financial institutions.

Precisely because of its power and impact, ML requires a more collaborative effort between the industry and the supervisory community to ensure that it protects customers without stifling its adoption or stalling innovation in the financial sector.
This Report

Machine Learning (ML) has been rapidly evolving in recent years, its use spanning several sectors such as agriculture, health, marketing, and financial services amongst others. Given the opportunities to improve predictive performance, and the practical issues present, the IIF has been researching financial firms’ applications of ML as well as the challenges and opportunities that arise from the use of ML.

In March 2018, the IIF published its *Machine Learning in Credit Risk Report* (ML-CR Report), in which we surveyed a globally diverse sample of 60 firms (58 banks and 2 mortgage insurers) on their applications, motivations, experiences, and challenges as they apply ML techniques in credit risk. A similar study was published in October 2018, our *Machine Learning in Anti-Money Laundering Report* (AML Report) surveyed 59 firms (54 banks and 5 insurers), which included a substantial overlap with the same group of firms that were interviewed for ML-CR Report. Later this year, the IIF will publish an update on the key findings of our ML-CR Report to examine how firms have progressed in the last year.

The IIF Machine Learning Thematic Series addresses common challenges in the use of ML in the financial services industry, in particular in its use for credit risk and anti-money laundering. This is the second paper from this series focusing on the issue of Bias and Ethics in Machine Learning in Finance. The first paper in this series, “Explainability in Predictive Modeling,” was published in November 2018. A third paper to be published in 2019 will elaborate on recommendations to supervisors and regulators.

In this report, we discuss: (1) what is bias and ethics in ML; (2) the experience of financial institutions with issues of bias and ethics; (3) sources of bias and how firms identify bias; and (4) approaches to addressing these challenges.

The first section of the report focuses on the theoretical discussion around bias and ethics in ML, discussing definitions, common challenges identified, and the overlap with current data protection laws. In the second and third sections, we explore examples of firms’ approaches to tackling with bias and ethics as they implement ML systems.

---

3. We note that whereas “explainability” was frequently cited as a hurdle in implementation for use in credit risk, the same was not borne out in our AML study.
Introduction

As the use of Machine Learning continues to filter into the financial services industry, our recent reports highlight the innovative uses of Machine Learning and the leverage of big data in the financial system. Firms’ motivations around Machine Learning in finance have been about making customer experiences simpler and faster. For example, consumers expect faster lending decisions. Several firms have indicated that ML can enhance their ability to bank new customers, which helps democratize access to financial services.

There are currently numerous definitions of ML. For the purposes of all our ML reports, we identified four main characteristics that most ML approaches conform to:

1. A primary goal of optimizing out-of-sample predictive performance facilitated by well-tuned regularization.
2. A significant degree of automation in the model development process.
3. The use of cross-validation to model relationships in the data, i.e. divide data into random separate sets for purpose of training, testing and validation.
4. Applicable to very large volumes of data (although some techniques also work well on small data sets), in some cases including unstructured data sources.

In both our ML-CR and AML Reports, we find that banks use both supervised and unsupervised learning in these areas. With supervised learning the input data is well-known and labeled, and the machine is trained to predict an output given certain inputs. With unsupervised learning the input data is neither classified nor labeled, and the machine is trained to group unsorted information according to similarities, patterns and differences without any prior training.

In credit risk, ML techniques are most commonly used to improve prediction in the areas of credit scoring and decisioning. Its usage in credit scoring and decisioning has been to improve insights from existing data sources in core modeling processes; in automatic credit decisioning processes to handle applications, such as a credit scorecard linked to a decision system; and to incorporate new types of data. ML has also been used to develop ‘challenger models’ to serve as a benchmark to the champion model in production in regulatory capital and provisioning models and for stress testing models.

In anti-money laundering, ML has been used primarily to enhance existing safeguards and create new techniques, especially in transaction and client activity monitoring. Customer segmentation and risk scoring are also key areas of usage, in particular where firms have been able to identify new risk segments and patterns of unusual behavior, and use this new knowledge to build new typologies in existing risk models for both risk assessment and monitoring purposes. Additionally, by analyzing an entire customer base for similarities using unsupervised ML, firms have been able to identify outliers and investigate if further action is warranted. Through the use of supervised and unsupervised learning, firms are developing algorithms to more effectively detect illicit flows of funds.

In both surveys, we found the challenges of implementing ML were similar, in particular those related to data quality and data sharing. The main common challenge is to create data that is sufficiently organized and accessible for the use of ML. Legacy IT systems and restrictions on access to data pose barriers to developing such data. Firms’ face challenges to create an effective IT environment and data architecture for data consolidation and mining. The challenge for data governance is to effectively look at best practices in accessing a sufficient amount of quality data,
the ability to bring data together into a single source or data lake, and in having the skills to analyze and interpret it.

An additional common challenge raised is the finding and retaining the right talent.

In terms of regulators’ expectations around explainability of models, the results of the ML-CR Report differed to those of the AML Report. “Explainability” was frequently cited as a hurdle in implementation of ML for use in credit risk, but not in our AML Report. Two main factors may be at play for this difference (i) the regulators’ greater sensitivity to undue bias in credit risk, and (ii) that in terms of its use for AML, the regulatory and supervisory community share a common goal with financial institutions to increase the efficiency of suspicious activity prevention mechanisms to have stronger defenses against financial crime.

This report addresses rising concerns from financial institutions and their supervisors on the ethical use of data in employing ML, and the potential for algorithms to introduce new biases or perpetuate existing ones against or for a particular group, such as a legally protected group. This report also considers the opportunity for ML to detect and minimize biases, and improve consistency, and uniformity in decision making.
1. Bias and Ethics in Machine Learning

1.1. Definitions

**What is bias?** Bias has contradictory meanings, as the history of the word itself indicates, and can have various interpretations in different contexts.

The statistical definition of bias relates to systematic differences between a population sub-sample and the whole population at large, where selection bias occurs when the set of data inputs to a model is not representative of a population, and results in conclusions that could favor certain groups over others. In ML systems models are trained to predict the future based on the past, as such what the ML model learns depends on the data used to train it.

To evaluate the performance of a method on a given data set, we need to measure how well its predictions match the observed data. This means quantifying the extent to which the predicted response value for a given observation is close to the true response value for that observation, where the true response value is a previously unseen observation not used for training. The accuracy of a resulting prediction depends on two quantities: the reducible error and the irreducible error. ML algorithms aim at minimizing prediction errors, which results in a trade-off to balance reducible errors – i.e. bias and variance.

Variance refers to the amount by which the model prediction would change if it were to be estimated with different training data. Bias refers to the error that is introduced by approximating a real-life problem by a simpler model. High bias occurs when predictions are far off from the actual values, and high variance occurs when small changes in the training data can result in large changes to the model prediction for a given data point. The challenge is finding a method with both low variance and low squared bias. Underfitting occurs when a model is too simple to capture the complex patterns in the data, whereas overfitting occurs when the algorithm is picking up some patterns in the training data that are caused by random chance rather than true properties of the learning function. This trade-off between bias and variance is crucial in understanding the behavior of the model; not finding that balance can lead to the model being less accurate or less predictive for underrepresented groups in the data.

In contrast, the popular, or social, definition of bias is human judgement made based on preconceived notions or prejudices, as opposed as the impartial evaluation of facts. Society tries to address this with legislation to prevent discrimination against particular groups of people. Protected groups vary by jurisdiction.

Both the statistical and the social meanings of bias are relevant and may lead to discrimination. Laws typically evaluate the discrimination using two distinct notions: disparate treatment, and disparate impact.

Disparate treatment includes overt discrimination, as well as more subtle unjustified differences in outcome on a prohibited basis. For example, in a credit scoring system, disparate treatment can occur during the data development and input stage when a lender treats a customer applying for a mortgage loan differently on a prohibited basis by providing assistance to improve one applicant’s qualifications and not others. Disparate treatment can also occur if the system

---

6 Oxford Dictionary, Definition of Bias
includes factors that are highly correlated with a prohibited basis that serves as a proxy, for example a lender gives a higher cut-off score for applicants in a certain zip-code highly correlated with race. Whereas disparate impact occurs when a neutral policy or practice results in a disproportional exclusion or burden on certain group of people, whether or not the policy was created with the intent to discriminate. In practice, the “effects test” is used to determine disparate impact. The test measures whether a credit practice has a disproportionately negative effect of discriminating against a protective class.

Therefore, it does not suffice to consider bias as a problem of mathematical correctness, rather it should also be considered in terms of how to make sure ML algorithms are fair and support ethical standards. The most important step in designing and building ML systems with inclusion in mind is to identify where and how bias is present.

**What is ethics?** Ethics is a system of moral principles governing a person’s behavior or the conduct of an activity. In the case of the financial institutions, ethics bridges the gap between the regulated and non-regulated spaces — that is, firms know what they can do in accordance with relevant laws and regulations, but ethics guides firms on what they should do. What is deemed “ethical” varies between individuals, societies, and jurisdictions, and can change over time. In this context, institutions are composed of multiple individuals with different views on what is and isn’t ethical, therefore fostering a culture of ethics also means having those discussions, rather than simply imposing a set of rules. Firms define ethics as a framework for reflection, beyond legal and regulatory requirements, to make decisions about what is right or wrong. Financial institutions have long established ethical standards that are enshrined in firms’ values and codes of conduct, incremental to those that are adopted in response to regulatory requirements such as those relating to fair lending or best interest standards.

**What is fairness?** No consensus on the “best” definition of fairness exists in the literature. We define “fairness” as a principle that understands and minimizes discrimination based on legally protected characteristics and aligns with ethical principles. In this context, bias and fairness are included as sub-topics under “ethics”.

In this context, it’s important to re-define the main goal of ML as maximizing predictive performance subject to a fairness constraint.

**1.2. Types of Bias in ML**

Bias, unfair outcomes and discriminatory impacts are not exclusive to the use of new analytical techniques such as ML. To understand these concerns in the context of ML, we need to ground the discussion with an acknowledgement that human decision-making is inherently biased. Many decisions today, including credit decisions, involve human judgement. For example, in 2013 the U.S. Department of Justice and the Consumer Financial Protection Bureau settled allegations of discriminatory practices by National City Bank. The bank engaged in a pattern of discrimination that increased loan prices for African-American and Hispanic borrowers by charging these borrowers higher loan prices not based on borrower risk, but because of their race or national origin. Specifically, the bank gave discretion to its loan officers and mortgage brokers to vary a loan’s interest rate and fees from the price it set based on the borrower’s objective credit-related factors. This subjective and unguided pricing discretion resulted in
African-American and Hispanic borrowers paying more than similarly qualified non-Hispanic White borrowers\(^8\).

History has shown that humans make biased decisions, whether intentionally and caused by prejudice, or unintentional and caused by limitations in knowledge or experiences. Bias is unavoidable, in fact any data collected by human processes will always have some bias in it. More automated use of credit scoring systems was found to “reduce the possibility of unlawful discrimination by helping ensure consistency and uniformity and minimizing individual judgment and discretion.” Another more recent study analyzed the effects of the use of ML in mortgage lending in the U.S. The authors concluded that when deciding whether borrowers are granted mortgages, machine learning benefits disadvantaged groups.\(^9\) ML can help discover traditional prejudices that are present in historical data. Rather than thinking about the use of ML solely as a risk of increased bias in decision-making, policymakers should understand its potential to reduce this bias. This also depends on whether there is a performance metric in place to measure fairness, and in what the correcting methodology entails following the discovery of bias.

Needless to say, the need to avoid unjustified bias goes well beyond financial services. Non-discrimination laws and data protection laws demand ethics and fairness, and typically prevent people from being discriminated against on the basis of certain protected characteristics. In the U.S., firms are subject to fair lending statutes that prohibit discrimination in lending,\(^1\) amongst other standards,\(^2\) and to their own internal codes of conduct, all of which predates ML. Similarly, in the EU data protection laws have long included "special categories" of data.

What has changed is that ML has created the potential for machines to learn from data that reflects human biases, including unconscious ones, and then exhibit and perhaps even amplify those biases. The main concern, and the motivation for the increased scrutiny of bias in the world of ML, is that misconceived correlations could have powerful implications given the automated nature of ML algorithms, and inherent biases can produce data that amplifies the biases already present in society. Additional wider aspects related to social “perception” of ML algorithms might also come into play.

---


\(^11\) The Fair Housing Act (FHA) and Equal Credit Opportunity Act (ECOA) protect consumers by prohibiting unfair and discriminatory practices. The ECOA prohibits discrimination on nine bases – race, color, religion, sex, national origin, age, marital status, receipt of public assistance, or exercised any right under the Consumer Credit Protection Act. Regulation B implements ECOA. FHA covers residential real-estate credit transactions, prohibits discrimination on seven bases – color, race, religion, sex, national origin, handicap and familial status.

\(^12\) Section 5 of the Federal Trade Commission Act (FTC Act), 15 USC 45(a)(1) (UDAP), and Community Reinvestment Act (CRA). CRA requires the Federal Reserve and other federal banking regulators to encourage financial institutions to help meet the credit needs of the communities in which they do business, including low- and moderate-income neighborhoods.
Perhaps the most recent major example of bias in ML is from Amazon’s recruitment tool that was scrapped after it was found to be biased against hiring women. The recruiting tool was built to analyze job applications with the intention of spotting candidates who would be worth recruiting. In this particular case, the tool was trained using data submitted to the company over a decade and given the gender disparity in the tech-industry, the data mostly came from male applicants.

Amazon’s recruiting tool highlights the issue of inherent biases in training data, as well as the issue of correlation in the data. The former is an issue due to the disproportional representation of a particular class in the training data, while the latter is an issue of a particular data or certain values in that data being highly correlated with membership in a protective/sensitive class.

The tool taught itself that male candidates were preferable. It penalized resumes that included the word “women’s” and downgraded graduates of two all-women’s universities, while favoring candidates using verbs such as “executed” and “captured” which were more commonly used by male applicants.

Amazon no longer uses the system beyond rudimentary chores such as removing duplicate candidate profiles from databases.

Financial institutions are taking a cautious approach to the use of ML. Data protection, security and integrity are already a key part of the design process for banks. Statistical models used for credit decisioning are already subject to model governance, model risk frameworks, and fair lending assessments. Financial institutions can adapt current governance and risk management frameworks to develop approaches to ensuring the ethical use of new technologies such as ML.

For instance, during a fair lending assessment in the U.S. examiners obtain a list of the variables considered and may conduct a comparative analysis between approved and denied applications to examine whether there are indications of disparate treatment. Many firms use existing regulations and supervisory guidance as a starting point, and carefully determine how to adapt these processes to the use of new technologies.

One of the biggest concerns firms share are with biased data leading to biased algorithms, in particular related to protected/sensitive features that could create moral, ethical and legal problems.

The most critical step in designing and building with inclusion in mind is to identify where and how bias is present. Below we identify the main types of biases in building and designing ML systems identified by financial institutions.

### 1.2.1. Bias in data, collection and processing

Training data, i.e. historical data, is reflective of human history and previous decision-making. It carries the biases and prejudices contained therein.

#### a. Dataset bias

Dataset bias occurs when the data used to train the ML model doesn’t represent the diversity of the population. In other words, the dataset is used to make generalizations on a group of sub-populations that differ meaningfully but are not each represented well in the sample. It’s a

---

source of bias because it makes inferences based on little information, or because the dataset is too homogeneous.

Bias by omission (or exclusion) can occur when the data used is biased of who is represented and omitted, resulting in underrepresentation of a population in the dataset. For example, certain facial recognition software works well only for a small subset of users, typically lighter skin users, because the training data underrepresents other skin-tones. This is a source of concern given the fast-moving adoption of facial recognition software by governmental agencies and, police, and it will be a concern around self-driving cars if similar data is used to train them.

Another example is that of the Amazon HR tool (see Box 1). Because the historical data used to train applicants underrepresented women, the ML system drew erroneous conclusions.

**b. Association bias:**
Association bias occurs when some data in the training set correlates with certain protected/sensitive characteristics, such as race, that cannot be used explicitly. The training set data, while correlated with class membership, may also contain legitimate and relevant information.

Bank can, as required by some regulations such as GDPR and U.S. fair lending laws, withhold sensitive demographic information from the algorithm. But this does not solve the problem of “redundant encodings,” where membership in a protected/sensitive class is encoded in other data. Therefore, if the training dataset is rich in features that are granular and diverse, the algorithm given particular pieces of data and values that are highly correlated can then deduce a certain protected/sensitive characteristic, even if implicitly.

In the recent Amazon example, gender was encoded by other data. The ML system learned from this encoded data to discriminate on female candidates (see Box 1).

Similarly, a ML system trained to predict likelihood to repay a loan, in which protected characteristics are excluded, could learn from encoded data such as geographic area in which the applicant lives, to discriminate based on race. This could result in a financial institution denying loans to applicants because of the racial demographics of the geographical area.

For this reason, it’s crucial to look at data characteristics carefully to ensure that there isn’t any direct or encoded bias in the features. Encoded features may include protected/sensitive attributes.

On the other hand, forbidding a bank from even collecting sensitive demographic information makes it difficult to check whether a model is inferring that sensitive information and creating unintended discrimination.

---

17 GDPR art. 9 lists special categories of sensitive personal data and forbids its use except in some exceptions and under certain conditions.
One manner in which firms have been tackling association bias is to require more representative datasets and require model governance processes to track the lifecycle of the training dataset. We discuss this further in Section 2. This can be done by tracking, accounting for legal and ethical considerations, how the dataset was created and processed. By doing this, banks seek to understand what demographic skews are present in a dataset and then work to understand potential blind spots and biases. This involves modifying the ML process to prevent bias.

### c. Cleaning and Transformation bias

Bias can become embedded in data through the process of data cleaning and transformation. For instance, feature engineering, which creates tailored attributes based on input variables that are used for strategies and models, can impact ML systems. The new “features” that are created may augment and aggregate certain attributes, while minimizing others.

The concern, which is not unique to ML, lies with the unintended and long-term consequences of classifying and categorizing individuals into groups that have long lasting impacts. Therefore, the process of feature engineering requires careful exploration and analysis. Banks are involving subject matter experts in the feature engineering process to ensure that the features selected are domain-specific. In some cases, firms are focusing their efforts on techniques that require less features, such as random forest, and principal component analysis (PCA).

Furthermore, firms are also examining in more detail how data is acquired, understanding clearly which features are “raw” and which engineered. This step allows validation teams to check in more detail the derived features prior to training the model. In our *ML-CR Report*, many firms indicated using ML techniques for the function of model development, in particular for variable selection and feature engineering. Firms have carefully implemented data enrichment and model validation processes to ensure that new features are well documented and described in an understandable manner.

The FICO case study in our “*Explainability in Predictive Modeling*” paper[^9] is a great example of what firms are currently doing to ensure that data is appropriate and valuable for analytical purposes. FICO have been looking at the process of data enrichment with scrutiny, methodologically and efficiently tackling the process of data wrangling and enrichment. The thinking is that by methodologically enriching the data through the process of cleaning and transformation, ML can help firms transform the raw data to the right structure and volume for modeling, without inserting too much judgement into the process, which could bias the results.

Firms indicated that the use of ML can make complex features more transparent. This can be done graphically, through residual analysis and localized partial dependency plots. Visualization techniques help firms understand signals and patterns and iterate as needed. In some instances, this is done by taking a templatized approach to generating characteristics from the data, in others via text analytics, and network analysis (i.e. constructing networks of entities and automatically generating a large number of aggregated characteristics which describe the entities within a network).

1.2.2. Other sources of Bias

a. Interaction bias
This type of bias comes from humans deliberately tampering with continuous learning ML systems and creating biased results.

For example, humans deliberately inputting abusive language to a chatbot designed to learn dynamically resulting in it learning to say offensive things.

b. Automation bias
This is the tendency to favor the results generated by automated systems over those generated by non-automated systems, regardless of the error rates of each.

For example, “death by GPS” is a phenomenon where tourists relied so heavily on erroneous GPS instructions that they drove into the ocean. In the context of financial services there may exist the risk that investors, for example, blindly place faith in algorithms that recommend trading strategies, without questioning its suitability.

1.3. Data Protection Laws and Bias

The rapid growth in the ability of companies and governments to collect and analyze data creates new challenges to provide appropriate data safeguards and consumer control, while maximizing benefits. At the same time, consumers and privacy advocates are demanding greater transparency and control around the use of customer data. Policymakers in many jurisdictions are responding by updating privacy laws, many of which were enacted prior to the development of current technologies and are mismatched with the capability of emerging technologies, to increase requirements about how data is protected and used and what constitutes personal information.

Certain provisions of these strengthened data protection requirements relate to the fair and ethical use of personal information. However, given the data-intensive nature of ML techniques, these requirements may create challenges for the use of ML or even exacerbate existing biases.

The reason behind this principle is that ML systems learn patterns and associations that are present in the training dataset, and their out-of-sample predictive accuracy can be improved by using more data.

The EU General Data Protection Regulation (GDPR), which went into effect in May 2018, makes significant changes designed to better address the realities of an evolving digital world while increasing the level of compliance accountability for organizations processing personal data. Its principles include greater transparency, how and why personal data is processed, and it strengthens individuals’ rights in and control over their personal data.

Provisions in the GDPR that set out individual rights in the context of automated decision making that has a legal or significant impact on an individual are particularly relevant to ML. In an attempt to reduce the impact of potential biased outcomes, the GDPR grants individuals the right to object to automated processes, demand disclosure of how these decisions are made or require human intervention in the process.

---

20 https://www.nytimes.com/2016/02/14/opinion/sunday/ignore-the-gps-that-ocean-is-not-a-road.html
While these articles apply to all sorts of automated decision-making processes, they are particularly relevant to ML techniques if they make use of personal information. Made explicit in the GDPR is the principle of data minimization, which indicates that “personal data must be adequate, relevant and limited to what is necessary in relation to the purposes for which those data is processed.”

Considering dataset bias as an issue, allowing access to more data would allow far more representative curated datasets and would go a long way towards addressing some of these issues. On the contrary, denying access to data will not only reduce the volume of data, but would also limit the data available to train the algorithm on a specific subset of a population, usually those more concerned in privacy protection.

Furthermore, prediction accuracy in ML is generally correlated with the amount of data available for training the algorithm, having little data on a group will likely lead to incorrect predictions in that group. A predictor might end up flagging an individual incorrectly, and when that identification coincides with sensitive information (i.e. race, gender, disability, etc.), it can result in an unfair outcome.

Perhaps counterintuitively, by raising the requirements around how personal information is collected and used, GDPR may end up exacerbating existing biases in ML training data.

Ideally, firms would be allowed to use sensitive / protected data in order to search for bias and run counterfactual scenarios. Following the discovery of bias, correcting methodologies should be taken and documented. Restrictions in data protection laws on the use of sensitive personal information makes it more difficult for firms to determine if a ML algorithm discriminates on the basis of a protected characteristic. Similar restrictions exist in certain financial regulations and data protection requirements can prevent firms from doing so. For instance, in the U.S. outside of mortgage lending, firms are not permitted to collect such data, and must rely on estimates to assess fairness of models. Several academic and scientific studies show the pitfalls of such estimation.

In summary, policy makers should be mindful of the consequence of certain rules on innovative technologies, specifically in the rules which determine how personal information can be processed.

2. Approaches for Ensuring Fairness

In this section we discuss some of the approaches taken by firms to address fairness, and then discuss concrete case studies provided by participating firms.

The key question here is how can firms ensure that their ML models haven’t learned a biased view of the world based on the shortcoming of training data, the model or their objectives. As firms mature in their application of ML, there is greater adoption of governing principles related to fairness, accountability, ethics and transparency.

---

22 GDPR, article 5 (1)(c)
The Amazon example suggests that companies should design, assess and build these systems with inclusion in mind from the beginning.

Firms acknowledge that to be successful there first needs to be an awareness of the potential correlations between features in the training data. As we discussed a methodological approach to data mining is key, and should encompass data cleaning, data wrangling, as well as validation and monitoring.

It’s important to caveat the discussion about challenges related to addressing fairness in ML with examples of ML helping overcome existing biases. One North American firm indicated that new ML models were effective at correcting biases against minorities, indicating that by placing a high emphasis on the detection of outliers they were able to consider groups in which the firm has little data on. This resulted in the ability to place into proper and fuller context the income gaps (such as maternity leave) of working moms (see Box 2).

Scotiabank’s Chief Risk Officer Daniel Moore indicated that ML was used to identify which credit card customers are likely to pay late, and they were able to engage these customers proactively to help. This resulted in a reduction of arrears by 10%.24

Many firms indicated that the use of Machine Learning has allowed them to understand and monitor the risk posed by each customer in a more holistic manner. If we take the example of women on maternity leave, the use of new techniques allowed for a reduction in information asymmetry and have led to better credit decisions, democratizing the access to financial services.

---

**Box 2: Identifying Income Gaps**

Many traditional models overlook the explanations for income gaps, this is typically due to the difficulty of traditional models to account for data gaps and missing values. For instance, in many traditional decisioning models an incomplete data field in a customer’s credit application may result in the application being suspended. In fact, data omissions can cause skews in historical datasets, and therefore be a source of model bias.

Two North American firms indicated that through the use of ML they were able to correct biases against minorities and women. This was accomplished by placing a high emphasis on the detection of outliers and on missing value imputation.

Rather than suspending an application, or declining credit to a customer, ML solutions used by these firms allowed them to have a greater capacity to generate a modeled output without that missing data point.

Whereas, a traditional model may have advised against extending credit to customers with income gaps and thus cause rejection of proposed loans to clients who took maternity leave, the use of ML allowed firms to overcome such data gaps. This is especially important, both in terms of overcoming biases, but also in financial institutions’ motivations to implement ML, which are heavily oriented towards better use of data and better modeling accuracy.

---

2.1. Governance Processes

There are different stages in the data mining governance process where firms can help prevent or reduce ML bias, below we examine six stages to evaluate existing biases. These can be adjusted to each firm’s own internal process, and where appropriate combined with the technical approaches listed on Sec. 2.2.

1. Conceptual soundness

As we have mentioned before, it is unrealistic to aim for a decision-making process that would be guaranteed to never include any form of bias whatsoever. A more sensible approach is to design and build with inclusion in mind. Therefore, the first and most important review point starts with the critical step of identifying where and how bias is present, reviewing key assumptions and limitations, and assessing the applicability of the model to models in scope.

This would also ensure that while the application of the technology is new, it is embedded into known processes, potentially making it more accessible to internal units and supervisors who are less experienced in their use.

The direction in which many institutions are heading is to design and implement high-level standards, that take into account ethics, accountability, transparency, and fairness that need to be achieved within the different business units of the firm.

Additionally, many of the firms we surveyed are concerned with cognitive diversity in their firms, in particular in the teams of people that build ML systems. The thinking is that in order to avoid homogeneous non-inclusive AI, companies need teams with diverse backgrounds and experiences to recognize bias.

2. Data use

Since machine learning applications rely on large amounts of data, and often multiple datasets, there should be an understanding of what data is being used, if it can and should be used, and an assessment of the potential risks that could arise from the use of that data. When appropriately used, data can facilitate new and improved products and services, increase revenue and mitigate risk. When not properly governed, certain data practices can damage a company’s reputation and cause a loss of customer and client trust. This goes beyond embedded bias in the data and to the broader question of how data can be used responsibly and ethically.

For example, one firm has implemented an enterprise-wide data use governance framework to ensure that it handles and uses data properly. Their program assesses the risks of new uses of data; while governing bodies are authorized to approve use of the data, they may impose additional controls to mitigate any identified risks. Data use decisions are guided by a set of firmwide data use principles to ensure that all sensitive information of the firm, its customers and its clients is protected and that the appropriate use of data creates a positive impact for all stakeholders. Their principles state, amongst other things, that data be safeguarded, that there is transparency in how data is used, that data owners have appropriate control over their data, and that reputational, operational and other risks are minimized.

3. Data governance

ML systems may be trained using legacy data that has been collected across multiple systems and which may not necessarily have been collected with the explicit purpose of training algorithms. Not all data is suitable for training ML systems in terms of its quality and accuracy. Datasets generally require some amount of preparation before they can yield useful insights.
There should be a robust data governance strategy that focuses on the data landscape, reference data and data quality. The proper identification of authoritative data sources versus datasets that may be conflicting or unnecessary allows efforts to improve data quality to be focused in the right place. It allows firms to address the quality of the data to ensure there are no issues such as missing or incomplete values, improper formatting or other issues that could make it difficult to process, and ultimately reduces the probability of creating insights that may be false or misleading.

4. Modelling

The modelling process of ML systems typically includes the process of choosing and training a ML algorithm, tweaking it and validating it on holdout data. Therefore, the choice of the design and deployment typically can introduce subjective human bias in terms of the chosen ML system. Many aspects can be taken into consideration when assessing for biases, such as the purpose of the model, reviewing the documentation of the rationale for the selected ML system, and using supportive analytics for key decisions (i.e. feature importance assessments).

Another aspect that should be considered is the importance of materiality in ML decision-making, i.e. evaluating the impact of decisions driven by the use of ML. The Monetary Authority of Singapore, a frontrunner on the use of AI, provided a series of indicative examples of consideration to be used when determining materiality (see sec. 2.3.1).

5. Outcome analysis and controls

Control frameworks play an important role in this context. The data selection and the results of the algorithm could be subjected to clear 1st, 2nd and 3rd level controls much in the same way that other business activities are. Sample testing of previous results could be used to identify if there is an apparent bias in the decision-making process, or if mistakes have been made in the data selection process. The precondition to an effective control framework would of course be to determine a “benchmark”, i.e. to agree which factors and results decide if the algorithm produces biased results.

Interpretability measures can also assist in extracting insights of the model that can be used to minimize bias. At the very least a feature importance assessment should be done to identify which features in the model are the most important, as well as attempt to identify for any single feature its related explanatory power individually.

For example, one firm noted that this is particularly useful when using more complex algorithms (i.e. those that necessitate all features to be used) where a feature importance assessment allows to select the top 20 features, dismiss those that do not drive the prediction, and those that can potentially bring bias into the system. Our Thematic Series paper “Explainability in Predictive Modeling” identifies some of the many different techniques available to provide satisfactory explanations of ML decisions.

6. Tuning and monitoring

Results and improvements deriving from these controls need to be fed back to address concerns. This means that it should be possible to amend which information is selected, or to build a system that allows to interfere with the calculations of an algorithm by a human (in practice this is currently the case). At the same time, wherever firms have had indications of biased decisions in the past, this information should be leveraged from the beginning to prevent it from happening again. For instance, one firm has a human investigator (checker) that can overrule
the ML outcome, which results on that specific logic tree being blacklisted for three months. The algorithm continues to learn from that experience.

A proactive documentation of the steps that have been taken to select datasets as well as their source is key. In the same context, audit trails are necessary regarding the algorithms that are applied, and which changes have been made to them (e.g. when applying supervised ML techniques). These data mining governance processes and their considerations need to be specified further. However, these specifications should not occur solely by regulators, but should be the result of a cooperative approach by the financial industry and the public sector. This would allow to reduce the uncertainty due to the lack of experience with these technologies on both sides and allow to build a sensible framework that protects the interests of all parties involved, while still promoting innovation.

The fact remains that even after a model is deployed there are various ways in which unfairness can be introduced. The solution is to be laborious in terms of model validation (see Box 3), and evaluation and monitoring of the system.

**Box 3: The importance of model validation**

For some firms, the introduction of ML techniques has changed the concept of model validation as part of the model development process. The argument is that given that data is the main driver in the modeling process, the actual framework for model governance also needs to adapt. Such firms are in the agreement that amending the current framework to improve model governance processes is necessary.

Nevertheless, other firms believe that in most jurisdictions introducing a new model governance rulebook is unnecessary and rather small adjustments or case by case considerations to existing model governance frameworks can be done to cover ML models. For instance, in the U.S. Federal Reserve’s “Guidance on Model Risk Management” requires firms to know a model’s use, who uses it, as well as demonstrate data governance, explain mitigation solutions, and regularly monitor its performance. For this reason, ML cases are mostly being handled with traditional validation, but making adjustments as needed.

However, all firms agree on the importance of ensuring that proper safeguards or technical considerations are taken when choosing data, understanding the assumptions, limitations and weaknesses in the model for instance. Additionally, many firms remark that validation would need to be revisited more frequently. In some cases, the use of a ML may necessitate a more dynamic validation, one in which monitoring is done on an ongoing basis.

In many instances, firms have identified potential for ML algorithms to improve current processes, in particular given the layer of bias in human decision-making (e.g. two individuals may reach a different conclusion based on the same information about a customer). There is currently some thinking behind automating processes to increase transparency of decision making, in particular cases in which tracking results may be more transparent.

With ML models there is greater risk that biases change more drastically based on incoming data (i.e. incoming bias / outgoing bias). Additionally, concepts such as interpretability and explainability touched on in previous papers are also motivations for tailoring governance frameworks.
2.2. Technical approaches by Financial Institutions

The approaches listed below are not an exhaustive list but represent some of the main responses firms have taken when addressing bias and ethics in ML. These should be considered as an overlay to our previous discussion on the stages in the data mining process (Sec. 2.1).

There are no uses of ML by member firms today that are completely autonomous, i.e., the model is used primarily as a tool to supplement human decision making. While it is desirable for humans to maintain involvement in automated processes, we note that there will be cases where a human broad override does not make sense: for example, in the detection and prevention of bad conduct, or the prevention of dangerous trading decisions. In these situations, there should be human oversight and accountability for the AI system, and individuals might need the authority to request the review of an automated decision when they have compelling cause. However, the individuals interacting with the AI system should not be able to directly override the AI system without cause, as this would undermine its protective function.

a. Tracking the lifecycle of the training dataset:
In terms of biased data, it’s important to track the lifecycle of the training dataset to understand how it was built and what the demographic skews might be in dataset. This allows data scientists to understand potential “blind spots” that may be embedded throughout the ML development pipeline, including product teams’ own blind spots. Product teams may struggle to identify which subpopulations they need to consider for specific kinds of ML applications. This approach should be considered when assessing “conceptual soundness,” as well as for the “data use” stage in the data mining process.

This technique entails tracking how the dataset was created, composed, the pre-processing done, and how legal and ethical considerations are taken into account.

One area of focus by institutions has been on bias in word embeddings, where in addition to tracking the lifecycle of the training dataset, firms have been focusing on improving the representativeness of datasets. It is vital for data scientists that develop algorithms to shape data samples in a way that biases are minimized.

Once potential biases are identified, firms have taken different approaches, such as:

- Eliminating problematic components of the input dataset. For example, a bank may strip the data to keep only data that is directly relevant to whether a customer will repay a loan.
- Expand training dataset. This would mean including more information to counterweight problematic data. For example, Los Angeles-based ZestFinance, founded by former Google CIO Douglas Merrill, found that the inclusion of more data and ML techniques reduced default rates significantly without added portfolio risk.25

b. Suitability assessment of customer’s profile
A seemingly easy practice but incredibly important is to perform a suitability assessment of the customer’s profile. This approach is crucial when considering the “modelling” and “outcome analysis and controls” stages in Section 2.1.

25 https://www.bizjournals.com/chicago/news/2019/03/05/discover-to-use-ai-to-score-your-credit.html
This entails not automating decision-making and having a clear understanding of the use-case for each model, the context, and model purpose.

To illustrate, a model that was trained and calibrated with data from customers ages 30-40 years old, is applied on a customer that is 85 years old, where clearly the attributes and risk profile of the customer are different.

c. **Oversampling**
This technique involves assigning heavier statistical weights to underrepresented data. This approach can be considered during the “data governance” and “modeling” stages discussed in Section 2.1.

This results in the algorithm being trained to pay more attention to the underrepresented data than what the data library might suggest. This technique can also be applied to algorithms that study age, race, gender, nationality, etc.

This can be done so that the algorithm doesn’t discriminate against or ignore specific types of people. Insufficient data could impact credit decisions for classes of borrowers to which a bank has not previously lent, but wants to do so in the future. It can also be used to add weight to make up for shortage of data in the dataset.

d. **Only include top performing variables that can be understood and explained**
ML systems require an iterative process, therefore performing a feature importance assessment every time the model is generated is crucial. This approach should be used during the “modeling” stage.

This entails performing a feature importance assessment on every single set of performing variables that are used, and then exploring the explanatory power of each performing variable individually.

This is done by firms to identify whether the model is taking the right features, and in cases in which the features are problematic these can be removed.

e. **Improving accuracy**
This is one of the main ways in which firms tackle unfairness. The below techniques are typically used in the “modeling”, “outcome analysis and controls” and “tuning and monitoring” stages indicated in Section 2.1.

Some techniques are concerned with:

- “accuracy equity” aimed at looking at overall accuracy of a predictive model for each group
- “conditional accuracy equity” which looks at accuracy of a predictive model for each group conditional on their predicted class
- “disparate mistreatment,” considers differences in false positive rates between groups

f. **Demographics or equal representation**
This technique can mean including an “equality of opportunity” or “affirmative action” measure. It can be used in the “modeling” process described in Section 2.1.
This technique helps to scrutinize predictors to uncover possible concerns. The main idea behind this technique is that those who qualify for a desirable outcome (receive a loan) should have an equal chance of being classified correctly.

**g. Averaging results**
A few firms are using this technique currently. This can be used in the “modeling” stage detailed in Section 2.1.

This entails a posteriori bias correction, where the sensitive attribute is kept in the dataset and the model decision is run twice. For example, let’s consider the following simplified example of a credit scoring model that has several inputs variables (X) plus the input variable of “gender” (G in male “M” and female “F”), and where the credit score (S) is the output.

\[(X, G) \rightarrow S=S(X,G)\]

To prevent discrimination with respect to gender, or to correct for bias, then you may give the following as an output:

\[S_{\text{non_discr}}(X, G) = \frac{1}{2} (S(X, G=\text{M}) + S(X, G=\text{F}))\]

By averaging the model decision of both the female and male attributes, the resulting score will not differ for a male or female having the same attributes X. For this posteriori bias correction to be applied, the sensitive attribute (G in this case) needs to be known.

**h. Challenger models**
Making use of a “challenger model,” in which the ML model is used to produce insights about particular correlations in the data, which can then potentially be applied to enhance production models. This approach has been used in the “outcome analysis and controls” and also for “tuning and monitoring” stages.

This entails having a ML model that serves as a benchmark to the “champion” model in production.

It has been used by firms in the context of bias to investigate mismatches between the ML system and the simpler regression-based model. For example, a firm used a simple regression-based model to test the ability to predict when a customer will default. It then created a more complex “challenger” model that used more advance ML techniques and had better prediction accuracy. The model was used to confirm that the base regression model was operating in line with expectations. It also helped verify that the ML system was not bringing unintended biases.

Several firms also indicated using challenger models as part of the validation process for their capital and provisioning models. Some of the strategies discussed in our Thematic Series paper “Explainability in Predictive Modeling,” are also applicable to addressing bias and ethics in ML.

**i. Blacklisting:**
This entails adding blacklist of terms that the algorithm should ignore (for example, adding profanities or racial epithets). It has been done by firms as part of the “tuning and monitoring” stage.

For example, one firm indicated that after a human investigator overrules a ML outcome, that specific logic tree ends up being blacklisted for three months, and the algorithm learns from that experience.
j. **Post processing calibration**
This entails calibrating classifiers parameters to give them the same acceptance ratio for all subgroups of sensitive features (e.g. race, etc.).

k. **Approaches using sensitive/protected attributes**
The following approaches are not commonly used by firms given the limitations with the use of sensitive/protected factors in the data. However, firms raised them as important techniques that should be considered in minimizing bias.

- **Causal reasoning**
  This is only possible if sensitive data is used.
  This entails modeling sensitive/protected factors such as sex and race, to ensure that their impact is captured and does not directly affect the output variable.

- **Counterfactual scenarios**
  Another technique that is likely to be used only if sensitive/protected attributes are used.
  In this technique members of the “protected group” are considered members of the “unprotected” group. For example, reclassify a woman as a man in the system.

2.3. **Development of High-Level Principles**
Banking supervision is already adapting to technology. Supervisors are staying abreast of how technology is changing risks in financial institutions, and many are examining ways in which AI advances can be used to help them supervise banks more effectively. For instance, in a recent speech James Proudman of the UK Prudential Regulatory Authority discussed the application of advanced analytics in prudential supervision.²⁶

Many actors have started working on high-level governing principles that take into account fairness, bias, transparency and accountability, some recent examples are from Microsoft Research’s FATE (Fairness, Accountability, Transparency and Ethics)²⁷ and the Monetary Authority of Singapore’s FEAT (Fairness, Ethics, Accountability and Transparency) Principles.²⁸

Financial institutions have also started to develop similar governing principles, and some institutions are already operationalizing these. As a first step this requires executive leadership to improve communication between technical, business and control functions, and to clarify the organization’s social and ethical values. Many firms are working together with the heads of the many different business units to identify key case studies in order to start operationalizing these principles. For many this is as an opportunity to lead by improving outcomes for customers, monitoring for unacceptable outcomes, being accountable for mistakes and misuses of results.

Supervisors are also working in many cases together with the industry to develop high-level principles that address issues of fairness, transparency, ethics and accountability. Two of the most relevant developments are included in our discussion below.

2.3.1. Monetary Authority of Singapore (MAS)

The Monetary Authority of Singapore (MAS) is currently the only financial supervisor that has developed high-level governing principles to promote fairness, ethics, accountability and transparency (FEAT) specifically catered for the financial services industry. Many participating firms indicated that the pragmatic approach taken by the Singaporean supervisor is useful in terms of implementation. MAS worked closely with senior industry partners through a special committee to develop the principles, taking into account their financial institutions feedback, as well as Fintech firms, technology providers and academia.²⁹

One key characteristic of FEAT is that governing principles are not intended to replace existing internal governance frameworks, rather the intent is for them to serve as a foundation when making ML driven decisions. Firms are to consider the materiality of ML driven decisions as firms calibrate actions and requirements under their internal governance framework.

Previously, we introduced the concept of materiality, and in this section, we use MAS’s illustrative examples of materiality as a starting point in our discussion. For instance, MAS considers the extent to which ML is used in decision-making, the complexity of the ML model, the severity and probability of impact on different stakeholders, the monetary and financial impact, and its regulatory impact.

In terms of fairness, MAS’s principle of Fairness focuses on two main aspects – firstly, justifiability, and secondly, accuracy and bias of ML driven decisions; both in line with the scope and discussion in our paper.

Justifiability is defined in this case as (i) individuals are not “systematically disadvantaged” through ML driven decisions, unless these decisions can be justified, and (ii) the use of personal attributes as input factors is justified.

One interesting example provided in the MAS Report is that in order to provide a more relevant product offering for different customer groups using information such as customer’s age may help decide whether to offer retirement-related products. The use of the personal attribute may therefore be justified given the relationship between age and retirement.

In terms of bias, MAS’s principle of “Accuracy and Bias” focuses on model validation and minimizing unintentional bias. This principle highlights our comments regarding the importance of model validation in ML driven decisions, the need to regularly review and validate for accuracy and relevance, ensuring that models behave as designed and intended.

In terms of ethics, MAS also addresses this principle head on, requiring ML use to be aligned to the firm’s ethical standards, values, and codes of conduct. This very much mirrors our own internal discussions in the development of this report, and the definition of ethics that we provided. Our firms agree that ML-driven decisions should be held to at least the same ethical standards as human decisions.

2.3.2. European Commission Ethics Guidelines for Trustworthy AI
In April 2019, the European Commission’s High-Level Expert Group finalized its *Ethics Guidelines for Trustworthy AI*. There is broad support for this initiative, and the Commission’s decision to focus on this important topic. However, these are not specific to the financial sector, but rather generalized in scope to encompass all types of industries.

The guidelines set out four ethical principles (Respect for human autonomy, Prevention of harm, Fairness, Explicability) and practical ways in which they may be implemented in the form of seven voluntary requirements. These requirements relate to human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental well-being; and accountability. The approach seeks to balance maximizing AI benefits with prevention and minimization of its risks and acknowledges that many considerations are not black and white. There is also a holistic focus on not just the AI system itself but the broader environment, including people and processes.

The guidelines endeavor to provide firms with practical methods for implementing ethical principles, including the use of a voluntary assessment framework. The Commission has also launched a pilot process for the assessment framework to test the approach and gather practical feedback from various industries on how the assessment list can be improved, which will inform the Commission’s periodic review of the guidelines in early 2020.

Similar to the MAS initiative, the EU recognizes objectives for trustworthy AI, which includes principles of ethics, and “unintentional harm”.

For instance, the “Principle of Respect Human Autonomy” presented in Chapter I gives the human the right to decide whether to interact with the AI system, and opt out or withdraw from the interaction. This provision ties into the concept first established by GDPR. This valid principle should ensure that individuals interacting with the AI system are not able to override the system without cause. To illustrate this, should a financial institution decide to use only ML-based systems for credit decisioning, the customer would realistically opt out of the entire product range solely because of the FI’s decision making process. This would overlook any economic benefits on both sides of the potential relationship.

Chapter II of the EU Ethics Guidelines sets 7 requirements to realize trustworthy AI and among those requirements is “Privacy and Data Governance”. In this regard, we note the following:

Firstly, financial institutions highlight that data which has been used to train a model should be retained, to ensure control activities to identify bias and address the situation. The data governance principle rightly states that identified biases should be used to improve processes and instructions and lead to improved decision making and stronger institutions. We note that this has been identified by the IIF as a stage in the data mining governance process (see Sec. 2.1. on the importance of a sound control framework). It should be highlighted that an improved decision-making process does not necessarily mean favorable decisions for all customers. A stronger institution also means clearer insights into hidden risks.

Secondly, we concur that higher ethics standards should not be asked only for AI-driven decisions, but rather that all decisions and all technologies should be subjected to the same rules.
Lastly, materiality is explicitly recognized by MAS, and in the EU Ethics Guidelines this concept also materializes under the “Principle of Explicability”. Herein, the degree of explicability is “highly dependent on the context, and the severity of the consequences if the output is erroneous or otherwise inaccurate.”\textsuperscript{39} To illustrate this concept, the EU Ethics Guidelines provides an example that compare inaccurate shopping recommendations generated by an AI system with an AI system that serves as a tool to evaluate whether a person will be convicted of a criminal offense if they are released from parole.

\footnote{39 AI Ethics Guidelines, p. 13}
3. Case Studies

3.1. HSBC Case-Study: Techniques to Prevent Bias

The use of Machine Learning brings new governance challenges for regulators and the industry.

We spoke with Giles Spungin, PhD currently Global Head of Analytics, Regulatory Compliance and Operational Risk at HSBC about the role of quantitative analytics, big data and the increasing use of Machine Learning for regulatory compliance and operational risk.

With a wide-ranging stakeholder group, Dr. Spungin works closely with other lines of businesses and functions within HSBC. “Risk management doesn’t work in isolation, there is a significant amount of engagement and coordination needed with all parts of the bank,” said Dr. Spungin. In his case the activity examples range widely from capital modeling all the way to AI/ML capability in many different contexts, where the team utilize open source algorithms (which they develop themselves), as well as rely on vendor platforms, deploy cloud-based capabilities, and produce visual and behavioral analytics.

For model development they have put significant importance on model validation, model testing and model performance monitoring. Not only from the point of view of putting together the model, but also verifying that the model fits the intended purpose, and that it does not introduce any unintended biases. Additionally, there’s a strong focus on on-going model performance, “...we do not deploy the model and forget about it. In fact, this is where the work really starts to ensure that the model continues to perform going forward,” said Dr. Spungin, “for example, in the context of conduct risk analytics capability, we need to ensure that there isn’t any model performance drift, and for that a number of different techniques are used.”

Additionally, in terms of model risk governance process, Dr. Spungin is also responsible for ensuring that the models build follow SR 11-7,31 and HSBC’s internal guidelines based on the regulatory requirements and best industry practice. Guidelines need to be followed not only to ensure that model performs as expected, but also to ensure that they are in line with those requirements.32 Dr. Spungin remarked on the importance of a clear understanding of the dataset the model is trained on, to ensure that the model is fit for the intended purpose.

The introduction of Machine Learning changes the concept of model validation, as these models perform and behave differently from well-established modeling approaches. Below Dr. Spungin provides insights on a blue-print of best practices:

- **Keeping track of the lifecycle of the data:** As part of model development and model validation knowing what data is used for training, what are the characteristics of the data, if there are breakages of the data those need to be understood very well, and decisions need to be made whether to exclude that dataset or include it but focus on how breakages impact model accuracy. This doesn’t happen only at model design stage, but when the model is used.

- **Validation:** The use of Machine Learning may necessitate the need to employ benchmark models by using other machine learning models to validate, as well as expert-

---

32 SR 11-7 requires firms to know a model’s use, who uses it; demonstrate data governance and understanding of assumptions, limitations, and weaknesses of models; as well as mitigation solutions while regularly modeling performance.
judgment based backtesting. The latter is very important in cases of unsupervised learning, where the algorithm is designed to categorize the data without human intervention.

- **Model Drift:** Equally important is the need to understand whether a model depreciates, or the model accuracy drops in time. Here it is useful to look into changes in distributions of the actual data and features that drive prediction. Overtime if there aren’t any significant changes in model inputs, but dramatic change in model output is observed, model needs to be carefully investigated.

- **Recalibration:** In order to ensure on-going model performance, the assessment of the performance drift is important, as well as clear articulation of the use cases, and that the model is regularly recalibrated. Recalibration can be done by regularly feeding fresh data into the model and examining how predictions change compared to the past. Frequency of calibration is dependent on the number of transactions that are coming in for a specific model, for example if there a sizable shift of the population which happens every six months then the model needs to be recalibrated at least every six months. This is where the business use-case knowledge comes in.

Model bias needs to be tackled at the model development stage and also monitored for after model deployment.

* A priori, it’s crucial to look at data characteristics carefully, this ties in with keeping track and understanding the data, to ensure that there isn’t any direct or indirect bias in the features. Indirect features may be related to “sensitive attributes”, such as gender being related to the color of a car, or zip code being highly correlated to income or race.

Making sure that the datasets are representative is also crucial, as making predictions on a subset of population that is not represented in the model sample will introduce bias. For instance, in the hypothetical example that a model has been trained with data from customers ages 30-40 years old, and have calibrated model to that dataset, and then this model gets applied on a customer that is 85 years old, clearly the attributes and risk profile of customer may be different.

Another technique to prevent bias, is unbalanced dataset correction, this should be done when the dataset is highly skewed to a certain customer group, for example. Dr. Spungin’s view is that “you cannot entirely automate the decision-making in model building, you need to have a good understanding of the use-case, the context, what you are trying to accomplish, model purpose, and that needs to come together.”

Once you have produced the model, there are many different ways to look at explainability. Simplest and most common method is partial dependency plots (PDP), with more involved methods including LIME and Shapley. This is an area of active research, and we are seeing increased focus on models with imbedded explainability, such as explainable neural nets (xNN). Dr. Spungin remarks that explainability assessment is an iterative process, that needs to be done every single time a model is generated and monitored for performance. Model features need to be continuously examined for explanatory power and suitability. This approach also helps to control for possible bias that could be introduced post model deployment by new data.

There are many different use cases where AI/ML can be issued in risk management, utilizing broad variety of data types, ranging from transactional data (e.g. customer and market),
surveillance data (e.g. market manipulation and conduct risk), loss data across the bank (e.g. operational risk and economic capital), and broader behavioral data.

In terms of behavioral data, for example, they look at how customer interact with their products, as some of the internal processes generate unstructured and non-numerical data and that necessitates usage of Natural Language Processing (NLP), sub-branch of ML. For example, pattern analysis in customer complaints is a good example where NLP adds value in product risk management activities.

**Mis-selling Case Study in Retail and Wealth Management**

There’s a very clear distinction between “execution-only transactions” and “advice-driven transactions”. “Execution-only” transactions are those in which the customer doesn’t receive advice from a relationship manager, whereas “advice-driven” transactions are those in which relationship manager advised a customer to buy or sell a particular product. HSBC has been using Machine Learning to mitigate mis-selling risk, by looking at specific data characteristics distinguishing these two transaction types.

Specifically, the team has created a model based on unsupervised learning, an anomaly detection model based on a technique called isolation forest, which is designed to seek out different characteristics/features of a transaction, and tries to look for the combination of which features most likely represent a product which was purchased on the basis of advice, and which features are representative of an “execution-only” transaction.

There is a marker in the system in which every time a product is sold, indicates whether the product was “execution-only” or “advice-driven”. With the advice route there is a clear expectation of a certain process to be followed. Interestingly, one of the features of advice/not-advice is customer’s demographics including age, because there may be a risk of a vulnerable customer mis-treatment due to one’s age profile, where the ethics angle is captured.

Another potential angle is product suitability, which applies not only to this mis-selling case study but more generally in the use of Machine Learning. As part of the process there needs to be an assessment of product risk level, and an assessment of the customer’s profile. For example, you may have a product-customer mis-match, where risk averse customer should not be sold highly risky products. The same intuition applies to complex product distribution use-cases.
3.2. Swiss Re Case-Study: AI and "Digital Governance Framework" (DGF)

Artificial Intelligence encompasses the convergence of machine processing, learning, perception and control supported by an exponential increase in volumes of data. Swiss Re perceives that it is important to monitor and understand proposed uses of data and technology in order to ascertain compliance with existing regulation, anticipated regulation and ethical considerations.

However, the scope of digital governance extends beyond these topics only, and includes topics such as cloud governance, information security or vendor governance. Such fragmentation slows down, or may even prevent, digital innovation.

Therefore, in order to ensure that digital services introduced at Swiss Re fulfill all internal and external governance requirements, and to make the life of owners of such services easier, Swiss Re recently developed and introduced DGF as a cross-functional, one-stop shop assurance process for assessing and managing the risks related to innovation in digital services, bringing the previously fragmented landscape of governance requirements under one roof.

The individual DGF assessments (currently ~35) are grouped within 15 "Gates" (the risk areas) and are triggered based on the answers provided to a self-service questionnaire. This dispatching step is followed by the actual gap assessment process.

*Figure 1: DGF provides a one-stop-shop assurance process covering all digital governance topics but only triggering the relevant ones*

One of the key risk areas of DGF is "Digital & Information Governance" (DIG), which assesses digital services against Swiss Re’s principles and policies put in place and followed. DIG ensures that the collection, storage and usage of data are consistent with brand values; Group ethics; and legal, regulatory and compliance related responsibilities.

In terms of the usage of personal data, firstly Swiss Re verifies that there is consent and transparency of the usage of personal data, that the purpose of data processing and the length of time the data will be used is disclosed, and that mitigation solutions on these aspects are upheld (such as anonymization of data).
Secondly, Swiss Re examines which rights may arise in respect of that data. This is done by taking into account: applicable confidentiality obligations, identifying the owner of the data, any rights and restrictions, as well as any copyright legislation application to that data, and identifying whether they have rights to use all of the data.

Finally, the use of data by AI products is the aspect that the DIG “risk area” investigates the most. Specifically, DIG currently works on integrating elements developed in the context of Swiss Re’s Digital Responsibility Framework and Digital and Smart Analytics Peer Review Framework. These aspects ensure, among others, that neither the AI algorithm itself nor the very nature of the case produces biased results. This means that bias should not be considered only as a technical problem, which it can be if unbalanced training datasets are used, rather it should be considered within a more pernicious source of bias one that may have happened at problem formulation. For example, by formulating a problem such as "do females generate higher costs than males", a bias is introduced long before any technology is applied. Identifying these sources of bias is therefore the prerequisite for proper bias prevention, given that the mitigation measures are of very different nature. Additionally, on the technical front a peer review framework for every analytics case that is run has been established.

*Figure 2: DGF provides a structured and calibrated framework to assess all requirements related to digital services*

DGF is a living framework that continuously evolves to adjust to the latest regulatory, governance and technology realities. By paying careful attention to governance, Swiss Re can extract value from data and technology while mitigating the main risks associated with digital services.
Conclusions

Financial institutions have progressed cautiously in their implementation of ML, applying intensive scrutiny across model development, testing and validation, and in terms of the broader model risk governance process. The use of ML in the financial industry has attracted concerns within firms’ management and supervisory community of the potential for algorithms to perpetuate existing biases against or for a particular group.

In this context, it is also important to consider that ML is already delivering benefits to individuals and society in terms of using its power to overcome biases. As discussed in this report, the use of ML techniques was found to reduce unlawful discrimination by ensuring consistency and minimizing human judgement, and as indicated in Box 2 has led to granting credit to groups of individuals (i.e. minorities and women) that would have been overlooked due to data gaps and missing values that are present in the data.

The aim of minimizing and preventing biases for ML models needs to be seen in context. Most existing linear or more traditional methods cannot be considered free of bias, in fact any data collected by human processes will always have some bias in it. Therefore, the use of ML needs to be understood in terms of its potential to reduce this bias.

It is also important to acknowledge that it is not sufficient to consider bias only as a problem of mathematical correctness, rather it should be considered in terms of how to ensure that ML algorithms are fair and support firms’ ethical standards. Although firms’ main goal for the use of ML is to maximize predictive performance, this should be subject to a fairness constraint.

Concurrently, policy makers should carefully consider that certain provisions to strengthen data protection requirements related to fair and ethical use of personal information, could create challenges for the use of ML.

We have identified practical approaches in six different stages of the data mining process (conceptual soundness, data use, data governance, modelling, outcome analysis and controls, and tuning and monitoring) for reducing or preventing ML bias.

The most critical step is that of conceptual soundness, which includes identifying where and how bias is present, reviewing key assumptions and limitations, and assessing model applicability. Many financial institutions have indicated that they are currently designing high-level standards that take into account ethics, accountability, transparency, and fairness. These institutions concurrently recognize and highlight the need to be laborious in terms of model validation, and evaluation and monitoring of the ML system.

Looking ahead, the IIF will be developing a set of recommendations for supervisors and regulators in the coming months. These recommendations can aid supervisors as they face the considerable task of aligning the different legal, technical and policy-related perspectives, while also ensuring that these do not stifle innovation.
Natalia Bailey
Associate Policy Advisor, Digital Finance
nbailey@iif.com

Other Contributors

Brad Carr
Senior Director, Digital Finance
bcarr@iif.com

Adrien Delle-Case
Policy Advisor, Digital Finance
bcarr@iif.com