Managing Discrimination Risk of Machine Learning and AI Models

David M. Skanderson*

Introduction

Businesses frequently rely upon predictive models to support or make decisions about investments, hiring and retention of employees, financial reporting, customer service (including granting credit, pricing, and marketing), customer relationship management, capital adequacy, and various other purposes. As a general matter, the use of any model entails "model risk," which can be defined as "the potential for adverse consequences from decisions based on incorrect or misused model outputs and reports."[1] The potential sources of such risk include, but are not limited to, design flaws; errors in assumptions, mathematics, or programming; data errors; errors in model implementation; or the misapplication of models to purposes for which they were not designed. However, it also includes the risk that a model's use will result in the violation of laws or regulations (such as prohibitions on discrimination), or will cause costly reputational harm. Increased model complexity, uncertainty about data inputs and assumptions, a greater extent

*Ph.D. Vice President, Charles River Associates. This article was submitted to the *ABA Journal of Labor & Employment Law* in connection with the 72nd Annual Conference of the NYU Center for Labor & Employment Law. The conclusions set forth herein are based on independent research and publicly available material. The views expressed herein are the views and opinions of the author and do not reflect or represent the views of Charles River Associates or any of the organizations with which the author is affiliated. Any opinion expressed herein shall not amount to any form of guarantee that the author or Charles River Associates has determined or predicted future events or circumstances, and no such reliance may be inferred or implied. The author and Charles River Associates accept no duty of care or liability of any kind whatsoever to any party, and no responsibility for damages, if any, suffered by any party as a result of decisions made, or not made, or actions taken, or not taken, based on this paper. Detailed information about Charles River Associates, a registered trade name of CRA International, Inc., is available at www.crai.com.

1. Bd. of Governors of the Fed. Rsrv. Sys. Supervisory Ltr. SR 11-7, Guidance on Model Risk Management 2 (Apr. 4, 2011), https://www.federalreserve.gov/supervisionreg/srletters/sr1107.pdf [https://perma.cc/MX4S-MAL3]. This guidance was jointly issued with two other federal bank supervisory agencies: Off. of the Comptroller of the Currency, OCC Bull. 2011-12, Supervisory Guidance on Model Risk Management (Apr. 4, 2011), https://www.occ.treas.gov/news-issuances/bulletins/2011/bulletin-2011-12a.pdf [https://perma.cc/N7J6-V2AS]; Fed. Deposit Ins. Corp., Fin. Inst. Ltr. 22-2017, Adoption of Supervisory Guidance on Model Risk Management (June 7, 2017), https://www.fdic.gov/news/financial-institution-letters/2017/fil17022.pdf [https://perma.cc/CX8Q-ECUE].

of model use, and a greater potential impact of the model all increase model risk.

The risk of unknowing and unintentional discrimination is an increasing concern with the increased application of complex machine learning and so-called artificial or automated intelligence (AI) models to ever-expanding sets of available data to make decisions about consumers and actual or potential employees. Because allegations of discrimination can be damaging to a business enterprise—in terms of both direct financial costs and reputational damage—businesses relying on models with potential discrimination risks need to take care to identify, investigate, and manage those risks. This article discusses key considerations in managing the discrimination risk posed by predictive models, based on the author's experience as a quantitative economist in financial services, and explains how concepts of model risk management that have been developed in the financial sector may be applied to managing discrimination risk (and other business risks) in other sectors. We start with a non-technical overview of machine learning models, followed by a discussion of considerations in evaluating models for fairness, before turning to the subject of model risk management.

I. Machine Learning and AI Compared to Traditional Predictive Modeling

In various areas of the business world, complex new machine learning and AI models are augmenting or replacing traditional predictive modeling methods or are being used to automate tasks traditionally carried out by people. This trend has been particularly prevalent in the banking and consumer finance field, the securities industry, and increasingly in human resources.[2] Machine learning uses computers to create analytical models on a largely automated basis, and to make decisions without being programmed with a specific set of decision criteria (although discrete decision criteria may be used in addition to the model).[3] Based on the available data and the general model structure

2. *See, e.g.*, Dom Nicastro, *7 Ways Artificial Intelligence Is Reinventing Human Resources*, CMS WIRE (May 18, 2020), https://www.cmswire.com/digital-workplace/7-ways-artificial-intelligence-is-reinventing-human-resources [https://perma.cc/ZA7V-E5JM]; *see also Artificial Intelligence (AI) in the Securities Industry*, FIN. INDUS. REGUL. AUTH. (June 10, 2020), https://www.finra.org/rules-guidance/key-topics/fintech/report/artificial-intelligence-in-the-securities-industry; Martin Leo, Suneel Sharma & K. Maddulety, *Machine Learning in Banking Risk Management: A Literature Review*, RISKS 1–2 (Mar. 5, 2019), https://doi.org/10.3390/risks7010029.

3. In this paper, I focus chiefly on models used for "supervised learning" problems—that is, cases in which there is a known outcome variable (such as a job performance measure)—and a model is developed to predict that outcome. This contrasts with "unsupervised learning" problems, in which the features (characteristics) of a population are observed but there is no measurement or prediction of a specific outcome. Instead, unsupervised learning is directed at describing how a set of data can be organized or clustered. *See* TREVOR HASTIE, ROBERT TIBSHIRANI & JEROME FRIEDMAN, THE ELEMENTS OF STATISTICAL LEARNING: DATA MINING, INFERENCE, AND PREDICTION 1–2 (2d ed. 2009).

and business objectives defined by the model developer, a computer algorithm determines which data elements to use in predicting behavior or another outcome of interest based on a statistical optimization process. Some examples of machine learning model applications are the targeting of advertisements on social media platforms, product recommendations on Internet retail sites and marketplaces, detection of fraudulent transactions by credit card issuers, prediction of the likelihood a consumer will default on a loan, and prediction of the likelihood of recidivism by convicted criminals.[4]

Machine learning and AI modeling methods differ from traditional predictive modeling, and the differences present unique risks that must be understood and managed to avoid unfairness in decision-making, including illegal discrimination.[5] Traditional predictive modeling (which has been used for many years in consumer credit scoring) typically uses economic theory and analysis together with statistical analysis to derive a fixed formula based on a defined set of data attributes, each of which is assigned a fixed number of points, resulting in a numeric score that represents the likelihood of an outcome of interest (such as default on a loan).[6] Such models are usually adjusted or replaced on a relatively infrequent basis, typically no more often than annually. Perhaps the most ubiquitous and commonly known example of such a predictive model is the FICO® credit scoring model.[7] However, even some commercially available credit bureau scores have increasingly adopted machine learning techniques to improve the predictive power of their scores, even though they fundamentally still rely on traditional model formulations.

By contrast with traditional models, the main focus of machine learning models is to make the best prediction of the outcome variable of interest without necessarily trying to model or explain the structure

4. Cathy O'Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy 74 (2016); Leo, Sharma & Maddulety *supra* note 2 at 6, 8; Dave Davies, *How Machine Learning in Search Works: Everything You Need to Know*, Search Engine J. (May 26, 2020), https://www.searchenginejournal.com/search-engines/machine-learning.

5. For a wide-ranging discussion of fairness issues related to machine learning and AI models, see O'Neil, *supra* note 4, at 3. For a survey of various concepts and measures of fairness see Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman & Aram Galstyan, A Survey on Bias and Fairness in Machine Learning (Sept. 17, 2019) (unpublished manuscript), https://arxiv.org/pdf/1908.09635.pdf. Examples of federal statutes prohibiting discrimination of various forms include Title VII of the Civil Rights Act of 1964, 42 U.S.C. §§ 2000e–2000e-17; the Age Discrimination in Employment Act of 1967, 29 U.S.C. §§ 621–634; the Americans with Disabilities Act of 1990, 42 U.S.C. §§ 12101–12213; the Equal Credit Opportunity Act of 1974, 15 U.S.C. §§ 1691–1691f; and the Fair Housing Act, 42 U.S.C. §§ 3601–3619.

6. Lyn C. Thomas, David B. Edelman & Jonathan N. Crook, Credit Scoring and Its Applications 41 (2002).

7. For an overview of the FICO score, see *What Is a FICO® Score?*, myFICO, https://www.myfico.com/credit-education/what-is-a-fico-score [https://perma.cc/YZ9P-LHFD].

of the underlying economic or behavioral relationship. Because such modeling methods largely automate the model-building process, they can be adept at crunching through large volumes of data to identify characteristics and their interrelationships that help to predict outcomes. Alternatively, machine learning methods may be used as tools for developing a model based on a more traditional modeling framework. In that case, the machine learning methods are used by the model to discover relationships in the data that might be considered for inclusion in a model. In machine learning, the automated model-building process determines which input variables (or features) are most useful and how to combine them to best predict a behavior or outcome based on the latest data available. In some cases, machine learning may combine the results of multiple models to increase predictive power.[8] Finally, machine learning methods tend to select attributes and combinations of attributes based purely on the strength of their correlations to the outcome being predicted. Less emphasis (or sometimes no emphasis) is placed on understanding whether logical economic or behavioral reasons underlie those correlations.

A key motivation for machine learning is the desire to identify and exploit subtle and difficult-to-observe relationships among disparate data elements from different sources that can be combined to better predict behavior. The underlying assumption is that some relationships are hidden within the data that either are too complex for a person to understand and ferret out, or that at least would require an inordinate amount of tedious effort to discover—effort which could be more efficiently performed by a computer. As a simple example, there may be a distinct difference in credit default risk between a consumer with multiple recent delinquencies on a single credit card account plus several recent applications for new credit with different banks, and who has also recently applied for a payday loan, compared to a consumer who is otherwise the same but has delinquencies across multiple accounts instead of a single account. In this example, what matters to the prediction is both the combination of values from three distinct data attributes, and a somewhat subtle distinction based one of those attributes.

It would be prohibitively time-consuming or impossible for a human analyst to evaluate all possible interrelationships among all available data elements to identify the characteristics that best predict default. Machine learning techniques, however, can allow an analyst to consider an arbitrary number of complex interrelationships among hundreds or thousands of candidate variables by letting the computer

8. Examples of this type of model are the so-called "gradient boosted" decision tree and "random forest" classes of models. *See* HASTIE, TIBSHIRANI & FRIEDMAN, *supra* note 3, at 359, 587.

grind through that tedious work. The appeal of such techniques has naturally grown as the amount of data available for analysis has grown, exponetially increasing the number of potential variable combinations and interactions that could be considered by a modeler.

In more specific terms, some of the key features that distinguish machine learning and AI models from traditional predictive models are as follows.

- In traditional predictive modeling, a person explicitly programs the computer with a set of specific instructions regarding how to predict the outcome of interest, based on the modeler's analysis. That analysis may include some of the same types of correlation analysis that would be performed in a machine learning context, but on a less automated basis and typically incorporating reliance on experience and judgment. By contrast, with machine learning, the modeler sets only general parameters regarding the modeling exercise and lets the computer build a mathematical model by discovering the underlying relationships between the outcome and a set of predictive variables in a less explicitly structured way based on a sample of data.[9] In basic terms, the modeler provides the computer a data sample containing lots of examples of "good" cases and lots of examples of "bad" cases, and provides a set of "guardrails" within which the model development process will operate, and then lets the computer discover which set of characteristics or combinations of characteristics are most common among the goods, most common among the bads, and thus allows one best to distinguish between the goods and the bads.
- Traditional predictive modeling selects and imposes a specific mathematical formula or equation (a "functional form") to represent the structure of the relationship between the outcome being predicted and the set of variables being used to predict it. That structure is typically informed by economic or behavioral theory, or at least intuition about the nature of the underlying relationship. By contrast, machine learning methods are atheoretical, and do not attempt to posit or impose any specific mathematical, economic, or behavioral relationship. They are agnostic about the structure or nature of the relationship and are generally divorced from any intuition about how and why a given set of predictive variables may be related to the outcome of interest. The question of *why* a specific attribute or combination

9. The parameters are typically referred to as "hyperparameters," which are parameters that guide and constrain the machine learning process. They are set by the modeler, rather than being estimated based on data as part of the modeling process. For an overview of machine learning (also known as "statistical learning"), see *id.* at 219, 232.

of attributes predicts the outcome is less important than the fact that it is correlated with the outcome, and that the correlation appears to be robust across different data samples.

- Traditional predictive modeling typically selects a single model—a single mathematical equation—that will be used to calculate a score (such as a credit score that ranks consumers in terms of their likelihood to default on a credit account) for individuals in a given population, although sometimes separate models are used to account for distinct population segments that exhibit different behavior or different relationships to predictive variables (e.g., consumers with some negative credit experience versus consumers with no negative credit experience, or consumers with extensive credit experience versus consumers with limited credit experience).[10] By contrast, many of the most popular types of machine learning methods used in the marketplace currently consist of multiple models—possibly hundreds or thousands of models—each of which is relatively simple, and the results of which are combined to generate a prediction.[11]
- In traditional predictive modeling, considerable human effort goes into attempting to discover and test which of the available data attributes best predict the outcome of interest and what form that relationship takes. Considerable automation may be involved in the process of testing alternative sets of predictive variables, but ultimately it is the human modeler who determines what should be included in the model through a combination of statistical analysis, theory, and practical judgment. By contrast, machine learning methods largely give over the responsibility for determining the set of variables to include in the model and how they should be included based on the strengths of their correlations to the outcome.[12] Based on a series of instructions provided by the modeler (parameters governing the overall structure of the model), the computer iterates through essentially all possible combinations of the available predictive variables (including all possible segmentations of each variable) and ranks them based on the strength of their predictive ability.

10. See, for example, Robert B. Avery, Kenneth P. Brevoort & Glenn Canner, *Does Credit Scoring Produce Disparate Impact?*, 40 REAL EST. ECON. S65, S77–S78 (2012).

11. For example, the often used "Gradient Boosting Tree" method is based on fitting relatively simple decision tree models on a series of many randomly selected subsamples of the training set, and the model's prediction is derived from the combined results of those models. *See generally* Jerome H. Friedman, *Greedy Function Approximation: A Gradient Boosting Machine*, 29 ANNALS STAT. 1189, 1216–17 (1999).

12. For a discussion if the differences in approaches between the goals and methods of traditional economic modeling versus machine learning, see Susan Athey & Guido W. Imbens, *Machine Learning Methods Economists Should Know About*, 11 ANN. REV. ECON. 685, 693 (2019).

II. Benefits and Risks

New modeling approaches and the automation of previously manual tasks have the potential to both reduce costs and provide new insights into human behavior, which could improve business decisions and profitability. In the credit market, for example, new models paired with nontraditional data sources are broadening credit access for consumers and small businesses, allowing lenders to better serve consumer segments that historically have been underserved, such as consumers who are unbanked, have low or moderate incomes, do not use traditional credit products, are self-employed, or have little established credit history.[13] Machine learning and AI models are also being used to protect consumers and financial institutions from increasingly complex forms of financial fraud, such as fraudulent credit card use, identity theft, and money laundering, based on a wide range of data regarding consumer spending patterns, merchant-level transaction information, electronic device information, and previously detected fraud, among other data.

Replacing judgmental human decisions with model-based decisions also has the potential to reduce overt and unintentional human bias in decision-making. In the labor market, for example, the use of predictive models in screening job applicants both frees recruiters from the tedious manual task of reviewing resumes and could reduce or eliminate the chance that conscious or unconscious human bias will affect applicant selection. However, automation of such tasks is not a guarantee of unbiased outcomes, because a model may engender bias through either the predictive variables that it uses or the data sample used to develop the model (the "training data").

In the author's professional experience, there is not a full appreciation among modelers of how the risk of discrimination may arise in automated, model-driven processes. The author has often heard modelers say something along the lines of, "We don't discriminate. Our models don't consider prohibited factors." Indeed, responsible model developers take steps to ensure that their training data are anonymized and that model developers have no access to data on prohibited factors. While that's a big step in the right direction, the risk of disparate impact may

13. See, for example, Julapa Jagtiani & Catharine Lemieux, *The Roles of Alternative Data and Machine Learning in Fintech Lending: Evidence from the Lending Club Consumer Platform* 1 (Consumer Fin. Inst., Fed. Rsrv. Bank of Phila. Working Paper No. 18-15, rev. 2019), https://www.philadelphiafed.org/consumer-finance/the-roles-of-alternative-data-and-machine-learning-in-fintech-lending [perma.cc/3EFC-NP9F]; *see also* Sumit Agarwal, Shashwat Alok, Pulak Ghosh & Sudip Gupta, *Financial Inclusion and Alternate Credit Scoring: Role of Big Data and Machine Learning in Fintech* 6 (Ind. School of Bus. Working Paper, Dec. 21, 2019), http://dx.doi.org/10.2139/ssrn.3507827 [perma.cc/R4FC-AQV3]; FinRegLab, The Use of Cash-Flow Data in Underwriting Credit 3 (2019), https://finreglab.org/wp-content/uploads/2019/07/FRL_Research-Report_Final.pdf.

not receive sufficient attention. Ostensibly neutral variables that predict behavior may nevertheless present a disparate impact risk if they are so highly correlated with a legally protected characteristic that they effectively act as a substitute, or "proxy," for that characteristic.

To understand how proxy relationships may create a disparate impact risk, it is helpful to consider an example. Some employers have used commercially available credit bureau scores, or other credit history information, as a factor in employment decisions. For example, a 2016 study found that forty-seven percent of employers used credit information in screening job applicants as of 2012.[14] Similarly, rental property managers often use credit scores and other credit reporting information in evaluating tenant applicants, and insurance companies often use credit scores in insurance risk rating.[15] Why is that? Presumably, credit information has been found to be, or is believed to be, predictive of preformance relevant to those contexts (i.e., the likelihood of paying rent on time or of filing an insurance claim).[16] Although it is not necessarily the case that people perform worse on the job or have auto accidents *because* they have bad credit history (i.e., the relationship is not necessarily causal), it may be the case that poor credit history is taken as a *signal* of some other underlying characteristic, quality, or behavior that is not directly observable and relates to the propensity to have poor job performance, make insurance claims, or be a good tenant.[17] Thus, in this example, credit history is acting as an indirect proxy for characteristics relevant to the performance of interest and that are unobserved, and which may be unobservable.

Why might the use of credit information in this way create a risk of disparate impact on a prohibited basis? Consider what would

14. Robert Clifford & Daniel Shoag, *"No More Credit Score" Employer Credit Check Bans and Signal Substitution* 2 (Fed. Rsrv. Bank of Bos. Working Paper No. 16-10, 2016), https://www.bostonfed.org/publications/research-department-working-paper/2016/no-more-credit-score-employer-credit-check-bans-and-signal-substitution.aspx [perma.cc/RP7Y-DJM7]. A more recent survey commissioned by the National Association of Professional Background Screeners found that twenty-five percent of employers performed credit checks for some job candidates based on the position and another six percent performed credit checks for all candidates. HR.COM & NAT'L ASS'N OF PRO. BACKGROUND SCREENERS, NATIONAL SURVEY, EMPLOYERS UNIVERSALLY USING BACKGROUND CHECKS TO PROTECT EMPLOYEES, CUSTOMERS AND THE PUBLIC 9 (2017), https://pubs.thepbsa.org/pub.cfm?id=6E232E17-B749-6287-0E86-95568FA599D1 [perma.cc/SCA5-TAM8].

15. FED. TRADE COMM'N, CREDIT-BASED INSURANCE SCORES: IMPACTS ON CONSUMERS OF AUTOMOBILE INSURANCE: A REPORT TO CONGRESS BY THE FEDERAL TRADE COMMISSION 2 (2007), https://www.ftc.gov/sites/default/files/documents/reports/credit-based-insurance-scores-impacts-consumers-automobile-insurance-report-congress-federal-trade/p044804facta_report_credit-based_insurance_scores.pdf.

16. See *id*. at 81, fig. 21, for information about the predictiveness of credit information in the insurance context.

17. For example, Clifford and Shoag, posit a theoretical framework of an employer's screening problem in which the true quality of a job candidate is unknown but credit check information provides a signal about quality, which is measured with random noise (or error). Clifford & Shoag, *supra* note 14, at 9–10.

occur if an economic recession in a particular employment market happened to disproportionately affect Black workers. Specifically, suppose a recession disproportionately increased unemployment among Black workers, resulting in a disproportionate incidence of financial hardship and associated credit difficulties—increased delinquency on credit accounts, increased bankruptcy and mortgage foreclosure, etc.[18] In this case, an employer who used a credit score or other credit history information in employment decisions may tend to disproportionately exclude Black applicants from employment simply because they were harder hit by the recession than white applicants or applicants from other race groups and not because they are necessarily less reliable or less skilled than other job candidates. Thus, the use of a proxy in modeling or decision-making can have unintended disparate effects. Such disparate effects may or may not constitute unlawful discrimination depending upon other relevant evidence.

As another example, consider how bias in a data sample used to develop a predictive model could result in a disparate impact. Suppose we trained a model to score applicants for employment based on resume data regarding current and past employees together with their job performance information. The resume information includes employment history information, but also educational information such as schools attended and degrees earned. Let's suppose that none of the employees in our data sample happened to have attended a historically Black college or university (HBCU). In this case, if a model trained on that data sample found the school attended to be strongly predictive of job performance, it would tend to attach relatively lower scores to candidates who attended an HBCU (other things equal) than to candidates from universities observed in the data and associated with good job performance, because HBCUs are not observed in the training data sample and, thus, are not observed to be associated with good job performance. This scenario would present a disparate impact issue. This is similar to the experience of Amazon, which reportedly experimented with developing a machine learning tool for ranking software developer job candidates.[19] Amazon's machine learning specialists reportedly found that their model would tend to disproportionately exclude

18. For example, differential rates of mortgage delinquencies and foreclosures based on race in the 2007–2009 economic recession were documented by Patrick Bayer, Fernando Ferreira, & Stephen L. Ross, *The Vulnerability of Minority Homeowners in the Housing Boom and Bust*, 8 Am. Econ. J.: Econ. Pol'y 1, 22 (2016). Race-based differences in unemployment rate increases in the 2007–2009 recession are reported in Hilary Hoynes, Douglas L. Miller & Jessamyn Schaller, *Who Suffers During Recessions*, 26 J. Econ. Persp. 27, 28 (2012).

19. Jeffrey Dastin, *Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women*, Reuters (Oct. 10, 2018, 6:04 PM), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G [perma.cc/YZ3F-7NAA].

women from employment as a result of the training data being highly skewed toward men.

Disparate impact can also be a subtle phenomenon, one that is both difficult to avoid and difficult to detect. In fact, in some cases, excluding from the model development process explicit identifiers of protected demographic characteristics (such as sex, race, national origin, and age) could cause variables that are correlated with those characteristics to be included in the model *because of* their correlations with the protected characteristics. This could occur in a situation where a protected characteristic itself has significant power to predict the outcome of interest.[20]

For example, suppose that senior citizens are considerably less likely than younger people to respond to a direct-mail credit card offer, but that the underlying reasons for that lower propensity to respond are not observable in our training dataset. In this case, senior citizen status would have significant power to predict response to a credit card offer, even though the relationship between senior status and response propensity is not causal, because senior status is a proxy for an unobserved aspect of credit card demand. In this example, the model's predictive power would be enhanced by explicitly including age group in the model, but one may want avoid that option because of the legally protected status of age under the Equal Credit Opportunity Act. However, excluding senior status from the model would necessarily create a gap in the information available to the model about response propensity. Machine learning models, which are driven by correlations among variables, would naturally to try to fill that gap through recourse to other information available in the training data set.[21] Specifically, because age is excluded from the training data and certain unobservable factors common among senior citizens are helpful in predicting response to a credit offer, a machine learning technique will tend to fill the gap by selecting alternative variables from the available data set that are correlated with senior status, even though they may have no direct or causal relationship to the propensity to respond to a credit offer. In other words, the machine learning model development process

20. For a more in-depth discussion of this "proxy discrimination" issue, and particularly the distinction between proxy discrimination and other forms of disparate impact, see Anya E.R. Prince & Daniel B. Schwarcz, *Proxy Discrimination in the Age of Artificial Intelligence and Big Data*, 105 IOWA L. REV. 1257, 1274–75 (2020).

21. Prince and Schwarcz similarly argue that AI models

> armed with big data are inherently structured to engage in proxy discrimination whenever they are deprived of information about membership in a legally-suspect class that is genuinely predictive of legitimate objectives. Simply denying AIs access to the most intuitive proxies for predictive but suspect characteristics . . . simply causes AIs to locate less intuitive proxies.

Id. at 1257.

would tend to select variables that are proxies for senior status because senior status happens to be a proxy for unobserved factors that cause the behavior that is actually of interest and not because they are directly related to the behavior the model is intended to predict. In this example, explicitly including age as a predictive variable in the model would cause such added variables to cease to be predictive. In this example, it is a second-order proxy effect that creates a potential disparate impact issue.

Such issues can be difficult to erradicate because machine learning methods will inherently tend to search for replacements for any valuable omitted information. Like nature, machine learning abhors a vacuum. This sort of unintentional disparate impact is quite different from the sort of discrimination issue that was the focus of early disparate impact claims—that is, intentional discrimination through indirect or covert means, or through aguably arbitrary rules that tended to perpetuate historical patterns of exclusion or segregation.[22]

Because of the issues discussed above, guarding against unfairness in the use of predictive models is not simply a matter of ensuring that they do not utilize overtly discriminatory variables. A model developer must also consider the possibility that ostensibly neutral predictive variables are directly or indirectly correlated with prohibited factors or that the data sample used in developing the model is not sufficiently representative of the demographics of the population to which the model will be applied. If the data sources used are not representative of the population of potentially qualified consumers or job applicants of interest and/or systematically exclude certain segments of the population, they may tend to create feedback loops that perpetuate or reinforce historical biases.

Avoiding bias may require adding human analysis and judgment to the process of model development, as well as oversight to ensure that models with the potential for discriminatory effects are developed in a sound and rigorous manner, and are defensible in terms of the business justification for the model inputs and sample selection. It requires recognizing which variables in a data set may be correlated with prohibited demographic characteristics, which may be beyond the knowledge or experience of the typical model developer. In short, it requires considerable focused analysis beyond that typically considered in the model development process.

22. *E.g.*, Phillips v. Martin Marietta Corp., 400 U.S. 542, 543–44 (1971); Weeks v. S. Bell Tel. & Tel. Co., 408 F.2d 228, 229–31 (5th Cir. 1969); Rosenfeld v. S. Pac. Co., 444 F.2d 1219, 1220 (9th Cir. 1971); Griggs v. Duke Power Co., 401 U.S. 424, 426–27 (1971).

III. Some Key Questions to Consider in Evaluating a Model for Fairness

Assessing, quantifying, and weighing the discrimination risk of predictive models and alternative data sources is a complex technical endeavor. It can be difficult to identify and understand the specific sources of a disparate impact because machine learning and AI models are complex, are inherently difficult to interpret, and tend to be far from transparent, and because the discrimination that may occur is typically an unintentional result of the model's development process and use, rather than a conscious choice by the model's developers. Even if the source of a disparate impact can be identified, it can be difficult to determine how to remedy it. Nevertheless, legal and compliance personnel can get a sense of the potential risk by asking the right questions and evaluating whether necessary controls are in place to manage risk. More broadly, as discussed in the next section, understanding and managing the risks posed by statistical models require establishment of an appropriate governance structure for model development and use.

"Interrogating" a model along the following lines can provide insights into where risks may arise and what additional analysis may be required to diagnose the extent of any risk. Once the risk is understood, necessary corrective actions may be taken, including potentially modifying or entirely scrapping the model, or implementing measures to limit and monitor the model's impacts as it is applied in practice.

1. Is it clear that no prohibited bases were used—explicitly or implicitly—in developing the model? This includes confirming not only that none of the variables used in the model explicitly represent a legally prohibited characteristic,[23] but also that none of the variables appear to act effectively as a stand-in (proxy) for a prohibited characteristic by virtue of a strong correlation with a prohibited basis. It also requires confirming that no prohibited bases were used explicitly or implicitly in selecting or generating the training data sample on which the model was built. The data sources and predictive variables being used as inputs to models and decision rules should be scrutinized accordingly. In the event that some of the variables used are perceived as likely to have strong correlations with legally prohibited factors or are potentially controversial,

23. In the case of consumer credit, one exception to the prohibition on considering the enumerated prohibited bases is the use of the age of an applicant, which is permitted by the Equal Credit Opportunity Act and its implementing Regulation B under certain conditions. Specifically, age may be taken into account to determine minimum legal requirements for a credit obligation, and a so-called "judgmental credit scoring system" may take age into account only to treat elderly applicants more favorably than younger applicants. In addition, a credit scoring system may specifically score differences in credit risk that may be related to a consumer's age or may use different sets of predictive variables for different age groups if it is "empirically derived, demonstrably and statistically sound," provided that the age of an elderly applicant is not assigned a negative factor or value. 12 C.F.R. § 202.6(b)(2) (2020).

or if the use of the model is likely to be subject to regulatory scrutiny, it is particularly important to ensure that the model was demonstrably developed in a statistically sound manner and that the business justification for the model and its predictive variables is well documented. Where feasible given the available data, one can test whether any perceived or suspected associations with prohibited bases actually exist and perform rigorous analysis to determine the extent of any disparate impact. Quantitative measures of disparate impact can then be weighed against the strength of a variable's or model's business justification. It is also important to evaluate the representativeness of the data used to develop a model, and any tendency to assume that data are inherently neutral and unbiased should be resisted. The provenance of the data should be understood and the likelihood that it underrepresents or excludes legally protected groups should be considered.

2. Do the variables used in the model each have a clear, intuitive, and explainable relationship to the outcome that the model is designed to predict? It is important to evaluate the relevance of the variables in a model to the behavior or outcome that the model is designed to predict. Data elements that appear to have predictive power but have no intuitive relationship to the behavior or outcome being predicted should receive extra scrutiny. If a clear intuitive relationship is lacking, then it is likely that the variable is merely acting as a proxy for some unobserved relationship. If the variable in question is correlated with a prohibited factor, then there is a heightened risk that the variable may be acting as a proxy for the relationship between the outcome of interest and the prohibited factor, thus creating a potential disparate impact issue. The key here is to always ask "why?" When potentially risky or questionable variables are encountered, it is important to evaluate how much they actually contribute to the predictive power and business objectives of the model and to weigh those benefits against potential legal or reputation risk in deciding whether the variables should be used. A variety of statistical tools can be useful in evaluating the tradeoffs.

3. If any of the model's predictive variables are correlated with a prohibited basis, do the variables' relationships to the outcome being predicted exist independently of protected class membership? If a predictive variable is truly neutral with respect to prohbited factors, it should have essentially the same predictive relationship to the outcome or behavior of interest when evaluated on data samples that are restricted to members of each protected class (e.g., the same relationship for Blacks as for Hispanics as for whites, or the same relationship for females as for males). If protected class information is known or can be estimated, then it is possible to quantify whether and to what extent the predictive power of the model differs across protected class groups.[24] If the relationship

24. In consumer financial services regulation and public health research, statistical processes have been used to estimate likely race and ethnicity using surnames and address locations, based on statistical associations between race/ethnicity and those attributes. *See* Marc N. Elliott, Peter A. Morrison, Allen Fremont, Daniel F. McCaffrey, Philip Pantoja & Nicole Lurie, *Using the Census Bureau's Surname List to Improve Estimates*

differs among protected class groups, then the variable may have either unfavorable or favorable differential impacts on certain groups.

4. Has the model's statistical validity been independently verified? It is important to ensure that models receive a rigorous statistical validation by a qualified, independent internal or external party to ensure that the models are statistically sound and were developed according to generally accepted statistical methods. At a general level, validation includes conceptual review to determine that the model design is sound and consistent with the model's intended purpose, data validation, replication and testing, review of mathematics and programming code, assessment of the model's limitations, evaluation of the ongoing monitoring plan for the model, and documentation and reporting of the validation findings. Model validation should be performed in the context of the specific intended use of the model, rather than in the abstract.[25] Statistical validation is aimed in part at confirming the evidence of a model's business justification (i.e., demonstrating the fitness of the model for its intended purpose and its accuracy, or predictive power, in accomplishing that purpose).[26] Therefore, statistical validity is an important line of defense against potential disparate impact claims, in addition to being a key means of understanding and controlling the business risk associated with using a model.[27] If a model or variable is found to have a disparate impact on a prohibited basis, it may still be legally permissible if its use is supported by a sufficient business justification.[28] In addition, if a model is, in fact, independently confirmed to use the most predictive combination of

---

of Race/Ethnicity and Associated Disparities, 9 HEALTH SERV. & OUTCOMES RSCH. METHODOLOGY 69, 70 (2009). This approach has been adopted by the U.S. Consumer Financial Protection Bureau. CONSUMER FIN. PROT. BUREAU, USING PUBLICLY AVAILABLE INFORMATION TO PROXY FOR UNIDENTIFIED RACE AND ETHNICITY: A METHODOLOGY AND ASSESSMENT 5–6 (Sept. 17, 2014), https://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf. Similarly, consumer first names can be used to estimate likely sex. However, such methods are subject to an unknown degree of error, which may be fairly large depending upon the population to which they are applied and protected class group of interest.

25. The use of a model for purposes other than those for which it was designed may undermine its statistical validity and may cause discrimination risk in some cases. For a discussion of model validation concepts and procedures, see OFF. OF THE COMPTROLLER OF THE CURRENCY, OCC BULL. 2000-16, RISK MODELING: MODEL VALIDATION 2 (May 30, 2000), https://ithandbook.ffiec.gov/media/resources/3676/occ-bl2000-16_risk_model_validation.pdf [https://perma.cc/2NQW-SKGN].

26. *Id*.

27. Regulations, regulatory guidance, and case law do not directly address the relationship between statistical model validation and defenses against a disparate impact claim, as far as the author is aware. However, because model validation directly addresses the question of whether a model advances a valid business interest, and advancing a valid interest is a key element of the burden of proof in disparate impact claims, it seems reasonable to infer that evidence of a model's statistical validity would be a key element of a defense against such claims. See, for example, *Texas Department of Housing & Community Affairs v. Inclusive Communities Project, Inc.*, 576 U.S. 519, 527–41 (2015) and HUD's Implementation of the Fair Housing Act's Disparate Impact Standard, Final Rule, 85 Fed. Reg. 60,288 (2020), for a discussion of the burden of proof framework for disparate impact claims under the Fair Housing Act.

28. *Tex. Dep't of Hous. & Cmty. Affs.*, 576 U.S. at 527–45; 24 C.F.R. § 100.500 (2020).

available predictive variables, then it generally should be unlikely that there is an equally effective but less discriminatory alternative available.[29]

5. Are the relationships captured by the model stable over time? A model's performance should be regularly monitored to ensure that it is truly capturing a stable predictive relationship and is not based on a fluke of the data sample used to develop the model. Periodic monitoring is often also referred to as "ongoing validation," because its purpose is to confirm (or refute) the continuing statistical validity of a model as it is applied in practice.[30] If the predictive power of a model quickly degrades over time and model developers need to frequently revise or replace the model as a result, that is a sign the model might not have been statistically valid to begin with, that the correlations on which the model was originally based may have been idiosyncratic to the particular data sample or time period used to develop the model, or that there has been a fundamental change in the underlying behavior or relationships being predicted. Regardless, if a model requires frequent revision, the reasons for that change should be investigated to determine whether fundamental issues exist with the model's validity. Ensuring that models retain their validity over time requires establishing appropriate performance metrics for each model, with reporting to the appropriate level of management. The monitoring process ideally would include setting predetermined tolerance thresholds for the performance metrics that would trigger escalated attention and analysis of the sources and causes of any apparent model performance degradation. In addition, because changes to a model may undermine its statistical validity, standards for model change control are needed to ensure that the risk impacts of changes to a model are well understood and controlled and that decisions regarding any resulting change in an organization's risk exposure are made at the appropriate level of management.[31]

6. Is the model and its development process sufficiently documented? Retention of relevant documentation and data is an important component of managing the discrimination risk posed by predictive models, because appropriate documentation is necessary for demonstrating the business justification of a model, decision rule,

29. Because of the correlation-driven nature of machine learning models, it is quite possible that alternative variables could be substituted for model variables without a significant loss in overall predictive power, though this is more likely to be the case for variables that have relatively low importance to the model prediction. Whether such substitutions would result in less discriminatory outcomes would depend upon whether or not the substitute variable has as strong a correlation with the prohibited factor of interest as the variable for which it is substituting.

30. Off. of the Comptroller of the Currency, *supra* note 1, at 12–13.

31. *Id*. Changes to a model might include such things as adding or removing one or more predictive variables, adjusting model coefficients or weights, recalibrating the model by re-estimating it on a more recent data sample, or applying the model to a different use. *Id.*

or predictive variable.[32] A requirement to document the model development process also helps to impose rigor on the development process by requiring explanations and justifications of such things as the use and objectives of the model; the data sources and sampling methods, and their appropriateness; the modeling methodology, assumptions, and reasons for key decisions in the process; performance testing and sensitivity analysis results; and implementation requirements. The statistics, logs, and other documentation created during model development may provide the requisite evidence justifying the use of each predictive variable in the model, as well as for the weight each variable receives in the decision process. The validation process typically ensures that such evidence has been independently verified. Documentation also provides a level of transparency to a process that otherwise would be opaque and not easily subject to independent oversight.

Documentation can be challenging on a practical level, because model developers often are not naturally disposed toward documenting their work or doing so in a way that is interpretable by non-technicians. However, failing to document can result in a much greater effort being spent later to develop business justifications *ex post* if a disparate impact concern arises. The passage of time and changes in personnel may thwart attempts at documentation and justification after the fact. Also, the data used to develop a model, the model development documentation, and the validation documentation should be retained because they will be needed in the event it is necessary to defend against a disparate impact claim. It is also helpful to retain the data used in each decision made by the model for the same reasons. If data are updated and overwritten over time (as sometimes occurs in business databases), it may be impossible to confirm in a retrospective review why the model rendered a particular decision for a particular individual and thus to demonstrate that the decision was justified on a nondiscriminatory basis. Ensuring that models are appropriately and consistently documented and that the requisite data are retained requires establishing and enforcing formal model documentation and retention standards at the enterprise level.

7. Are any of the model's predictive variables likely to attract heightened regulatory scrutiny or adverse publicity, even if they are defensible? Even though the components of a model may be defensible, controversial predictive variables are not without risk. For model users subject to supervisory examinations (such as in consumer financial services, for example), variables that are unconventional, that are not clearly related to the purpose of the model, and/or that are believed to be related to protected class characteristics could result in a costly and time-consuming regulatory inquiry or formal investigation. In addition, even if a model is intended to be confidential, it is worth considering the potential reputational damage and risk of litigation that a variable's use could create if it were to leak to the public, because leaks do occur despite the best

32. *See* OFF. OF THE COMPTROLLER OF THE CURRENCY, *supra* note 25, at 3–4. A key theme in model documentation standards is to require that (at a minimum) sufficient documentation be created to allow replication of the model by an independent party. *Id.*

> security precautions. Therefore, consideration should be given to the "headline risk" attached to potentially controversial or questionable model variables. Where such risks are identified, it is useful to investigate the importance of the variable(s) in question to the model's predictive power and to weigh the variable's contribution to business objectives against the perceived regulatory or reputation risks to decide whether the variable(s) should be used.

If potential discrimination risks are identified, rigorous statistical analysis may be needed to evaluate the severity of the risk. Given that many economic and personal characteristics of individuals have some degree of correlation with legally prohibited factors, such as race, sex, and age, it is inevitable that some models and predictive variables will have disparate effects in relation to prohibted factors. However, disparate effects do not necessarily translate into illegal discrimination if the variables in question have a sufficient business justification. Ultimately, the question of whether or not to use a variable or model often comes down to a practical risk judgment, taking into account both the magnitude of the disparate effect and the strength of the business justification supporting the model or variable.

Performing statistical analysis to evaluate discrimination risk appropriately can be challenging. This is because establishing that there is an inequality of outcomes on the basis of protected class membership is a necessary condition for establishing an unlawful disparate impact, but it is not a sufficient condition because of the business justification defense.[33] Such analysis may also include quantifying the extent to which individuals with inherently equal or reasonably similar qualifications are treated differently by a model (e.g., people with similar credit risk are assigned different scores by a credit scoring model, or people with similar skills and aptitudes for a given job are ranked differently in candidate selection).[34] Therefore, there is typically a need to translate the legal concepts of disparate impact and business justification or necessity into statistical terms and to devise a way to compare or weigh the two against each other. At the most basic level, there are questions of measurement: how to quantify the existence and magnitude of any disparate impact and the strength or magnitude of a model's business justification. Methods for testing for fairness in predictive models, especially machine learning models, are rapidly evolving. Various researchers have proposed different ways of approaching the problem, and, as yet, there appear to be no standard or generally accepted methods.[35]

---

33. *See Tex. Dep't of Hous. & Cmty. Affs.*, 576 U.S. at 542 (discussing burdens of proof under the Fair Housing Act).

34. See, for example, the reasoning articulated in *Wards Cove Packing Co. v. Atonio*, 490 U.S. 642, 658–60 (1989).

35. Examples of approaches that focus on evidence that a variable or a scoring model acts as a "proxy" for a prohibited basis include Fed. Trade Comm'n, *supra* note 15,

Evaluating the risk of unlawful disparate impact ideally requires that the extent of any disparate impact be defined in terms of measures that can be compared to quantitative evidence of business justification or necessity. In addition, as suggested by the discussion regarding proxy discrimination above, it is necessary to determine (if possible) the reason for a model's disparate impact—such as whether particular variables with a disparate effect have strong predictive power independently of correlations with prohibited factors, or because of their correlations with prohibited factors. Finally, once the statistical measures have been defined, there is still a question of how much business justification is "enough" to compensate for a given amount of disparate impact. The question of whether to accept a model for use when it displays some evidence of disparate impact appears to be an inherently subjective process of weighing risks and benefits, and deciding how much legal, regulatory, or reputation risk the organization is willing to accept.

IV. Managing Model Risk Through Model Governance—Lessons from Financial Services

The process of managing model risk is already well developed in the financial sector. Non-financial industries and business functions can look for guidance to the regulatory standards and industry practices in that sector. Financial sector regulatory guidance regarding model risk management has been developed jointly by the various U.S. financial regulators and by international organizations concerned with financial stability.[36] Those guidelines require that a financial institution establish a regimented model governance process, the elements of which can be distilled down to several key components.

First, an enterprise should establish policies and procedures defining roles and responsibilities for managing and overseeing model risk. At a high level, the policies include assignment of primary responsibility and accountability for managing model risk to each model owner,

---

at 50–73, and Avery, Brevoort & Canner, *supra* note 10, at 66–67. Examples of approaches that have been proposed for evaluating model fairness specifically for machine learning models include Yair Horesh, Noa Haas, Elhanan Mishraky, Yehezkel S. Resheff & Shir Meir Lador, *Paired-Consistency: An Example-Based Model-Agnostic Approach to Fairness Regularization in Machine Learning*, *in* MACHINE LEARNING AND KNOWLEDGE DISCOVERY IN DATABASES 590 (2020), and Cyrus DiCiccio, Sriram Vasudevan, Kinjal Basu, Krishnaram Kenthapadi & Deepak Agarwal, *Evaluating Fairness Using Permutation Tests*, PROC. 26TH ANN. ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY & DATA MINING 1467 (2020), https://arxiv.org/abs/2007.05124.pdf.

36. For an overview of U.S. bank regulatory expectations regarding model risk governance, see Off. of the Comptroller of the Currency, *supra* note 1, at 2–7. As an example of international standards, see Commission Regulation 575/2013, 2013 O.J. (L 176) (amending Commission Regulation 648/2012, 2012 O.J. (L 201)) (EU) on prudential requirements for credit institutions and investment firms.

subject to oversight by independent departments.[37] The policies include a critical role for senior management (including business, risk, legal, compliance, and audit leaders) in the oversight of model use and model risk, with governance and reporting processes to ensure both appropriate visability of risks within the organization and decision-making regarding risk at the appropriate level of management. Key roles and responsibilities defined in the governance process also include a role for internal audit, to confirm that the defined policies and procedures are being followed.

Second, a scheme should be devised for classifying models with respect to the level of risk that they pose to the enterprise (typically specified in terms of three or four risk tiers). The assignment of risk ratings to models is typically based on the nature of their use, extent of their impact, and the potential consequences of model error or failure (including financial, operational, regulatory, legal, and reputational impacts).[38] A model's assigned risk rating guides the extent of oversight to which the model is subjected, including model validation and reporting requirements. For example, low-risk models may have fewer or less stringent requirments than high-risk models regarding the frequency, intensity, formality, and/or degree of independence of model validation activities.

Third, an enterprise should establish and maintain a comprehensive inventory of models used by the enterprise, including both internally developed and vendor models, to facilitate oversight and management of model risk. Model inventories typically include such information as the purpose and intended use of the model; the model owner and developer; the model's assigned risk rating; the model's main data input sources and outputs; the dates that the model was developed and implemented; the dates of completed and planned validation activities; the expected retirement date of the model; and identification of model interdependencies (i.e., other models upon which a given model is dependent for inputs or other models that depend upon the model for inputs). Identifying, cataloguing, and classifying all of an enterprise's models can be a major project in a large enterprise, but is an essential one because it facilitates the visibility to sources of model risk that risk management decision makers need to identifying and overseeing that risk at the enterprise level.

Fourth, the model governance process should specify requirements for model validation, including the frequency of validation and

37. The "model owner" is typically defined as the organizational unit or individual that is the primary user of the model. The model owner typically determines the business requirements for the model and is responsible for correct implementation of the model.

38. For a survey of model risk rating practices in the financial sector, see Nick Kiritz, Miles Ravitz & Mark Levonian, *Model Risk Tiering: An Exploration of Industry Practices and Principles*, J. Risk Model Validation, June 2019, at 47.

revalidation, and requirements for the degree of independence of the party performing the validation. Typically, an initial validation by an independent party is required prior to model use (at least for higher-risk models), with ongoing performance monitoring and validation typically performed by either the model owner or an independent party. In general, validation should be performed by a party with sufficient expertise to provide "effective challenge" to the model developer, which typically requires both expertise in model development methods and an understanding of the business context or application of the model. Model validation standards typically specify requirements regarding how material weaknesses or limitations of a model uncovered by a validation exercise must be addressed (such as mitigation through model changes or limits on the model's use). Validation requirements apply to both internally developed and vendor models, although the specific requirements for vendor models may be different because they are often "black boxes" and the information available to the model user about the model's development and inner workings is usually proprietary to the model vendor.

Fifth, a model governance process should define the specific requirements for documentation of the model development process and the validation. Documentation standards are typically specified and enforced through the use of standardized documentation templates. As noted above, documentation standards act as a control both to provide sufficient visibility and oversight to the development and use of models and to impose formal rigor on the development and management of models.

Lastly, a sound governance process also includes a defined model change management process, which ensures that the risks associated with any changes to a model are evaluated, documented, independently reviewed, appropriately approved, and controlled. Like in information technology, change control helps to prevent disruptions to a business or unplanned risk-taking that may result from *ad hoc* model changes or a lack of appropriate quality control and impact assessment prior to model changes.

As outlined above, model risk management standards for financial institutions require a deliberate process of management and oversight, which includes various checks and approvals throughout the model life cycle, from the beginning of model development through model testing, use, and decommissioning.[39] Good model oversight and model risk governance also helps to guard against the risk of both discrimination and other adverse business outcomes.

Non-financial companies that use predictive models with potential for discrimination risk can use the financial sector's model risk

39. Off. of the Comptroller of the Currency, *supra* note 1, at 2–7.

management principles and processes outlined above to guide their own management of model risk, even though they are not subject to the same sorts of regulatory requirements as financial institutions. The processes need not be as elaborate as that required of financial institutions, but the same principles may be applied effectively in the narrower context of controlling discrimination and other legal risks. Companies developing and using machine learning and other predictive models can design an effective system for model risk management by ensuring that there is an appropriate level of model governance structure that provides mechanisms for identifying, evaluating, and controlling the discrimination and other legal risks associated with the models.

A process that is more narrowly focused on managing the discrimination risk associated with models would distill the key elements of the general model risk management framework outlined above to the aspects that are directly relevant to discrimination risk, although a more comprehensive approach to model risk management is also beneficial. Specifically, the key elements of risk management policies to focus on are formally defining the types or categories of models that are subject to legal and compliance review or approval requirements, and defining who in the organization is responsible and accountable for ensuring that models comply with legal standards. The operational processes involved in managing discrimination risk would include a process for engagement of legal and compliance personnel by model owners in the organization, which would include defining the stages of the model "lifecycle" at which legal/compliance consultation or approval is required for models posing discrimination risk.[40] At the beginning of the model lifecycle (prior to the start of model development), the discrimination risk management process might require vetting of the proposed data sources and inputs planned to be used or considered in model development. That stage of the process, however, could also include defining sets of pre-approved "safe" data that might be used by model developers without additional approval. This may include neutral data attributes that have been previousl vetted for discrimination risk. Later in the model lifecycle, legal or compliance personnel might have either veto power or a defined escalation path regarding particular types of models if they appear to pose unacceptable risk.

Just as in the case of a broader model risk management framework, development and maintenance of an inventory of models are central pieces of the puzzle, but the inventory could more narrowly focus on the set of models that pose a risk of discrimination and that,

40. The "lifecycle" of a predictive model refers to various stages in creation and use of a model, such as the initial concept proposal prior to approval for development, model development and testing, implementation, performance monitoring, updates/revisions or other changes, and decommissioning.

therefore, require legal oversight. Similarly, the model risk-rating framework could be narrowly focused on characterizing salient aspects of risk that relate to fairness and discrimination, including the nature of the legal exposures or potential consequences of a discrimination issue, and the nature and size of constituencies affected by the model (e.g., customers or employees).

Requirements for model documentation and validation should be no different when focusing on discrimination risk compared to other business risks, because documentation and validation both are key controls on the risk of unlawful discrimination and provide the basis for a business justification defense against a discrimination claim. However, validation requirements could be augmented in this case by defining requirements for disparate impact testing (which is not part of a standard risk model validation process). This process might include the specificaton of criteria that would determine whether and how often a model must undergo disparate impact testing based on the model's risk rating or intended use, either prior or subsequent to implementation. Similar to the case of a more general risk model validation standard, a narrowly tailored fairness validation standard would specify circumstances in which remedial action is required based on the validation findings, because of either a material model weakness or a finding of disparate impact risk.

Conclusion

New predictive modeling technologies and data sources offer the prospect of great benefits to various functions within business enterprises. However, those benefits can come with significant risks of illegal discrimination when applied to such uses as credit granting, insurance underwriting, hiring, and personnel management. Legal and compliance personnel responsible for overseeing those business functions need to be familiar with how predictive models and automated decision tools are being used in their organization, and such personnel should establish processes for diagnosing and managing the potential discrimination risks.

By integrating legal risk management with broader risk model governance, a business enterprise can become more efficient in managing the entire risk associated with predictive models, including discrimination risk. When structured appropriately, legal and compliance personnel can have the information and engagement with model owners in the organization necessary to identify and mitigate potential discrimination risk issues. With proactive risk management, an organization's model owners may have a greater chance of avoiding potential regulatory action or litigation and a reduced chance of incurring expenses associated with *ex post* detection of discrimination issues, which can result in the costly retooling of models already in production.