Algorithms and bias: What lenders need to know

Financial institutions must mind their fintech solutions to ensure they do not generate unintended consequences

Algorithms and bias: What lenders need to know

The algorithms that power fintech may discriminate in ways that can be difficult to anticipate—and financial institutions can be held accountable even when alleged discrimination is clearly unintentional.

By Kevin Petrasic, Benjamin Saul, James Greig, Matthew Bornfreund and Katherine Lamberth

Much of the software now revolutionizing the financial services industry depends on algorithms that apply artificial intelligence (AI)—and increasingly, machine learning—to automate everything from simple, rote tasks to activities requiring sophisticated judgment. These algorithms and the analyses that undergird them have become progressively more sophisticated as the pool of potentially meaningful variables within the Big Data universe continues to proliferate.

When properly implemented, algorithmic and AI systems increase processing speed, reduce mistakes due to human error and minimize labor costs, all while improving customer satisfaction rates. Credit-scoring algorithms, for example, not only help financial institutions optimize default and prepayment rates, but also streamline the application process, allowing for leaner staffing and an enhanced customer experience. When effective, these algorithms enable lenders to tweak approval criteria quickly and continually, responding in real time to both market conditions and customer needs. Both lenders and borrowers stand to benefit.

For decades, financial services companies have used different types of algorithms to trade securities, predict financial markets, identify prospective employees and assess potential customers. Although AI-driven algorithms seek to avoid the failures of rigid instructions-based models of the past—such as those linked to the 1987 "Black Monday" stock market crash or 2010's "Flash Crash"—these models continue to present potential financial, reputational and legal risks for financial services companies.

Consumer financial services companies in particular must be vigilant in their use of algorithms that incorporate AI and machine learning. As algorithms become more ingrained in these companies' operations, previously unforeseen risks are beginning to appear—in particular, the risk that a perfectly well-intentioned algorithm may inadvertently generate biased conclusions that discriminate against protected classes of people.

As algorithms become more ingrained in these companies' operations, previously unforeseen risks are beginning to appear.

Input bias could occur when the source data itself is biased because it lacks certain types of information, is not representative or reflects historical biases.

EVOLUTION OF ALGORITHMS AND BIAS

The algorithms that powered trading models in the 1980s and 1990s were instructions-based programs. Designed to follow a detailed series of steps, early algorithms were able to act based only on clearly defined data and variables. These algorithms were inherently limited by the availability of digitized data and the computing power of the systems running them.

The development of Big Data, machine learning and AI, combined with hardware advances and distributed processing, has enabled engineers to design algorithms that are no longer strictly bound by the parameters in their operational code. Algorithms now run off data sets with thousands of variables and billions of records aggregated from individual internet usage patterns, entertainment consumption habits, marketing databases and retail transactions. The complexity of the interconnections and the sheer volume of data have spurred new data processing methods.

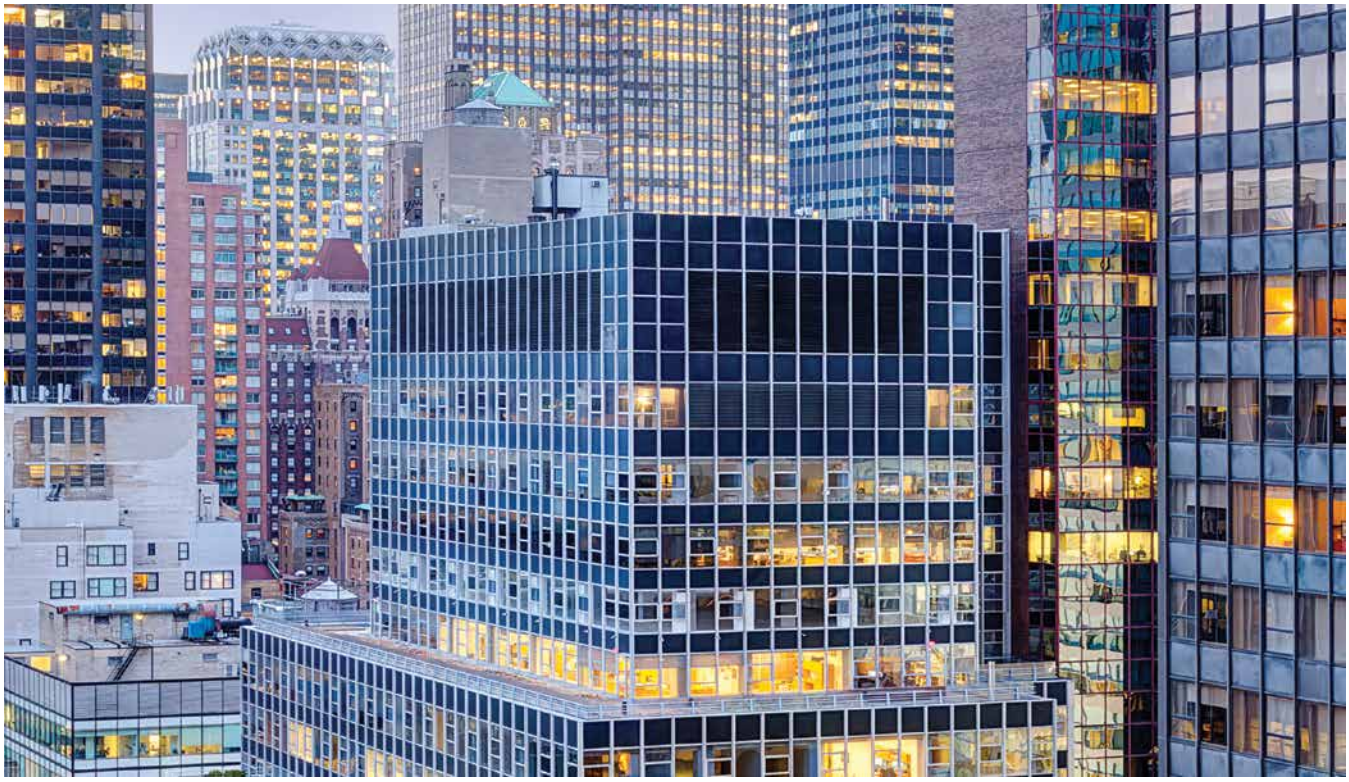
The rise of financial technology (fintech) since 2010 coincides with an intensifying focus on AI by computer scientists, prominent information technology companies and mainstream financial firms. The push into AI is driven in part by the need to derive and exploit useful knowledge from Big Data. Although still short of artificial general intelligence—the kind that appears to have sentient characteristics such as interpretation and improvisation—specialized AI systems have become remarkably adept at independent decision-making.

The key to developing these "smart algorithms" is using systems that are trained by recursively evaluating the output of each algorithm against a desired result, enabling the machine program to "learn" by making its own connections within the available data.

One goal of an algorithmic system is to eliminate the subjectivity and cognitive biases inherent in human decision-making. Computer scientists have long understood the effects of source data: The maxim "garbage in, garbage out" reflects the notion that biased or erroneous outputs often result from bias or errors in the inputs. In an instructional algorithm, bias in the data and programming is relatively easy to identify, provided the developer is looking for it. But smart algorithms are capable of functioning autonomously, and how they select and analyze variables from within large pools of data is not always clear, even to a program's developers. This lack of algorithmic transparency makes determining where and how bias enters the system difficult.

In an algorithmic system, there are three main sources of bias that could lead to biased or discriminatory outcomes: input, training and programming. Input bias could occur when the source data itself is biased because it lacks certain types of information, is not representative or reflects historical biases.

Training bias could appear in either the categorization of the baseline data or the assessment of whether the output matches the desired result. Programming bias could occur in the original design or when a smart algorithm is allowed to learn and modify itself through successive contacts with human users, the assimilation of existing data, or the introduction of new data. Algorithms that use Big Data techniques for underwriting consumer credit can be vulnerable to all three of these types of bias risks.



CONSUMER FINANCE AND BIG DATA

When assessing potential borrowers, lenders have historically focused on limited types of data that directly relate to the likelihood of repayment, such as debt-to-income and loan-to-value ratios and individuals' payment and credit histories. In recent years, however, the emergence of Big Data analytics has prompted many lenders to consider nontraditional types of data that are less obviously related to creditworthiness.

Nontraditional data can be collected from a variety of sources, including databases containing internet search histories, shopping patterns, social media activity and various other consumer-related inputs. In theory, this type of information can be fed into algorithms that enable lenders to assess the creditworthiness of people who lack sufficient financial records or credit histories to be "scorable" under traditional models. Although this approach to underwriting has the potential to expand access to credit for borrowers who would not have been considered creditworthy, it can also produce unfair or discriminatory lending decisions if not appropriately implemented and monitored.

Complicating the picture, nontraditional data is typically collected without the cooperation of borrowers—borrowers may not even be aware of the types of data being used to assess their creditworthiness. Consumers can ensure that they provide accurate responses on credit applications, and they can check if their credit reports contain false information. But consumers cannot easily verify the myriad forms of nontraditional data that could be fed into a credit-assessment algorithm. Consumers may not know whether an algorithm has denied them credit based on erroneous data from sources not even included in their credit reports.

Without good visibility into the nontraditional data driving the approval or rejection of their loan applications, consumers are not well positioned (and, regardless of visibility, may still be unable) to correct errors or explain what sometimes may be meaningless aberrations in this kind of data.

Although creditors must explain the basis for denials of credit, disclosing such denial reasons in ways that are accurate, easily understood by the consumer and formulaic enough to work at scale can present a significant challenge. This challenge will be magnified when the basis for denial is the output from an opaque algorithm analyzing nontraditional data. Borrowers' inability to understand credit decision explanations could be viewed as frustrating the purpose of existing adverse action notice and credit-reporting legal requirements.

Companies that attempt to comply with the law by providing notice of adverse actions and reporting credit data may face unique and complicated challenges in translating algorithmic decisions into messages that satisfy regulators and can be operationalized, especially where large swaths of potential borrowers are denied credit.

This challenge will be magnified when the basis for denial is the output from an opaque algorithm analyzing nontraditional data.

As machine learning becomes more powerful and pervasive, its complexity—as well as its potential for harm—will increase.

HOW ALGORITHMS INCORPORATE BIAS

Biased outcomes often arise when data that reflects existing biases is used as input for an algorithm that then incorporates and perpetuates those biases. Consider a lending algorithm that is programmed to favor applicants who graduated from highly selective colleges. If the admissions process for those colleges happens to be biased against particular classes of people, the algorithm may incorporate and apply the existing bias in rendering credit decisions. Using variables such as an applicant's alma mater is now easier and more attractive because of Big Data, but as the use of algorithms increases and as the variables included become more attenuated, the biases will become more difficult for lenders to identify and exclude.

Algorithms often do not distinguish causation from correlation, or know when it is necessary to gather additional data to form a sound conclusion. Data from social media, such as the average credit score of an applicant's "friends," may be viewed as a useful predictor of default. However, such an approach could ignore or obscure other important (and more relevant) factors unique to individuals, such as which connections are genuine and not superficial.

Analyses that account for other attributes could reveal that certain social media metrics are better than others at predicting individuals' creditworthiness, but an algorithm may not be able to determine when data is missing or what other data to include in order to arrive at an unbiased decision.

Finally, and most importantly, an algorithm that assumes financially responsible people socialize with other financially responsible people may incorporate systemic biases, and deny loans to individuals who are themselves creditworthy but lack creditworthy connections.

Determining every factor that should be included in a predictive algorithm is challenging. A compelling aspect of AI and machine learning is the capacity to learn which factors are truly relevant and when circumstances exist to override an otherwise important indicator.

Algorithms that predict creditworthiness rely on advanced versions of what are called "unobtrusive measures." The classic example of an unobtrusive measure comes from a 1966 social science textbook that described how wear patterns on the floor of a museum could be used to determine which exhibits are most popular. This type of analysis uses easily observed behaviors, without direct participation by individuals, in order to measure or predict some related variable.

However, systems based on unobtrusive measures may have a large inference gap, meaning that there can be a significant mismatch or distance between the system's ability to observe variables and its ability to understand them (based on factors such as the range and depth of background knowledge and the context provided). In a 2012 paper that modeled the spread of diseases based on social network postings, researchers from the University of Rochester demonstrated that machine learning can be applied to unobtrusive measures of text to identify phrases that are the strongest predictors of illness.[1] In part, the algorithm developed strategies to minimize the inference gap and deliver more accurate results.

As machine learning becomes more powerful and pervasive, its complexity—as well as its potential for harm—will increase. Code aided by AI will increasingly enable computer systems to write and incorporate new algorithms autonomously, with results that could be discriminatory. Even the developers who initially set these new algorithms in motion may not be able to understand how they will work as the AI evolves and modifies the features and capabilities of the program.

Consider an AI lending algorithm with machine learning capabilities that evaluates grammatical habits when making a credit decision. If the algorithm "learns" that people with a propensity to type in capital letters, use vernacular English or commit typos have higher rates of default, it will avoid qualifying those individuals, even though such habits may have no direct connection to an individual's ability to pay his or her bills.

From a risk standpoint, using language skills as a creditworthiness criterion could be interpreted as a proxy for an applicant's education level, which in turn could implicate systemic discriminatory bias. Reference to certain language skills or habits, while seemingly relevant, could expose a lender to significant bias accusations. The lender may have no advance notice that the algorithm incorporated such criteria when evaluating potential borrowers, and therefore cannot avert the discriminatory practice before it causes consumer harm.

1 "Predicting Disease Transmission from Geo-Tagged Micro-Blog Data," University of Rochester, July 2012.

This scenario clearly sets up the distinct possibility of not only a bad customer experience, but also the potential for reputational risk to a lender that fails to disclose in advance the factors for making a credit decision—and perhaps similar risk if disclosure calls attention to a factor that may be hard to explain from a public relations standpoint.

An algorithm learning the wrong lessons or formulating responses based on an incomplete picture, and the lack of transparency into what criteria are reflected in a decision model, are especially problematic when identified correlations function as inadvertent proxies for excluding or discriminating against protected classes of people. Consider an algorithm programmed to examine and incorporate certain shopping patterns into its decision model. It may reject all loan applicants who shop primarily at a particular chain of grocery stores because an algorithm "learned" that shopping at those stores is correlated with a higher risk of default. But if those stores are disproportionately located in minority communities, the algorithm could have an adverse effect on minority applicants who are otherwise creditworthy.

While humans may be able to identify and prevent this type of biased outcome, smart algorithms, unless they are programmed to account for the unique characteristics of data inputs, may not.

To avoid the risk of propagating decisions that disparately impact certain classes of individuals, lenders must incorporate visualization tools that empower them to understand which concepts an algorithm has learned and how they are influencing decisions and outcomes.

AI AND THE ALGORITHMIC VANGUARD

The subjective credit evaluation process is an ideal target for AI and machine learning development. Lending decisions, by their nature, are based on probabilities derived from patterns and previous experience.

Research has demonstrated that subjective evaluations with similar characteristics can be effectively handled by purpose-built AI. The key is the ability of AI to reprogram itself based on categorized data provided by the developers, a process called supervised learning. While effective, supervised learning can inject unintended biases if the inputs and outputs are not monitored as an AI program evolves.

In one of the earliest attempted uses of supervised learning for photo identification, a detection system designed for the military was able to correctly distinguish pictures of tanks hiding among trees from pictures that had trees but no tanks in them. Although the system was 100 percent accurate in the lab, it failed in the field. Subsequent analysis revealed that the groups of pictures used for training were taken on different days with different weather and the system had merely learned to distinguish pictures based on the color of the sky.

The people responsible for training the tank-detection program had unintentionally incorporated a brightness bias, but the source of that bias was easy to identify. How will the humans training future credit-evaluation algorithms avoid passing along such subconscious biases and other biases of unknown origin?

DISCRIMINATION NEED NOT BE INTENTIONAL

For years, fair lending claims were premised mainly on allegations that an institution intentionally treated a protected class of individuals less favorably than other individuals. Institutions often could avoid liability by showing that the practice or practices giving rise to the claim furthered a legitimate, non-discriminatory purpose and that any harmful discriminatory outcome was unintentional.

But recently the government and other plaintiffs have advanced disparate impact claims that focus much more aggressively on the effect, not intention, of lending policies. A 2015 Supreme Court ruling in a case captioned *Texas Department of Housing and Community Affairs v. Inclusive Communities Project* appears likely to increase the ability and willingness of plaintiffs (and perhaps the government) to advance disparate impact claims.

In the case, a nonprofit organization sued the Texas agency that allocates federal low-income housing tax credits for allegedly perpetuating segregated housing patterns by allocating too few credits to housing in suburban neighborhoods relative to inner-city neighborhoods. The Court, for the first time, held that a disparate impact theory of liability was available for claims under the Fair Housing Act (FHA), stating that plaintiffs need only show that a policy had a discriminatory impact on a protected class, and not that the discrimination was intentional.[2]

2 "Symposium: The Supreme Court recognizes but limits disparate impact in its Fair Housing Act decision," Scotus Blog, June 26, 2015.

Although the Court endorsed the disparate impact theory of liability under the FHA, it nevertheless also imposed certain safeguards designed to protect defendants from being held liable for discriminatory effects that they did not create. Chief among those protections is the requirement that a plaintiff must show through statistical evidence or other facts "a robust causal" connection between a discriminatory effect and the alleged facially neutral policy or practice.[3]

Notwithstanding such safeguards, the fundamental validation of disparate impact theory by the Court in the *Inclusive Communities* case remains a particularly sobering result for technology and compliance managers in financial services and fintech companies. An algorithm that inadvertently disadvantages a protected class now has the potential to create expensive and embarrassing fair lending claims, as well as attendant reputational risk.

WHAT LENDERS CAN DO TO MANAGE THE RISK

Financial institutions may have to forge new approaches to manage the risk of bias as technologies advance faster than their ability to adapt to such changes and to the rules governing their use. In managing these risks, lenders should consider following four broad guidelines:

Closely monitor evolving attitudes and regulatory developments

The Federal Trade Commission, the US Department of the Treasury and the White House all published reports in 2016 addressing concerns about bias in algorithms, especially in programs used to determine access to credit.[4] Each report describes scenarios in which relying on a seemingly neutral algorithm could lead to unintended and illegal discrimination.

An algorithm that inadvertently disadvantages a protected class now has the potential to create expensive and embarrassing fair lending claims, as well as attendant reputational risk.

Also in 2016, the Office of the Comptroller of the Currency (OCC) announced in a whitepaper that it is considering a special-purpose national bank charter for fintech companies.[5] The OCC paper makes it clear that any company issued a fintech charter would be expected to comply with applicable fair lending laws. In a speech focused on marketplace lending, OCC Comptroller Thomas Curry questioned whether fintech, specifically credit-scoring algorithms, could create a disparate impact on a particular protected class.[6] He also stressed that existing laws apply to all creditors, even those that are not banks.

If and when the OCC begins to issue fintech charters, the agency may provide guidance to help newly supervised companies manage algorithms in ways that reduce their exposure to bias claims. The OCC supervisory framework developed for fintech banks may also be instructive to other financial services firms that use algorithms for credit decisions.

Meanwhile, companies should ensure individuals responsible for developing machine learning programs receive training on applicable fair lending and anti-discrimination laws and are able to identify discriminatory outcomes and be prepared to address them.

Pretest, test and retest for potential bias

Under protection of attorney-client privilege, companies should continuously monitor the outcomes of their algorithmic programs to identify potential problems. As noted in a recent White House report on AI, companies should conduct extensive testing to minimize the risk of unintended consequences. Such testing could involve running scenarios to identify unwanted outcomes, and developing and building controls into defective algorithms to prevent adverse outcomes from occurring or recurring.[7]

Analyzing data inputs to identify potential selection bias or the incorporation of systemic bias will minimize the risk that algorithms will generate discriminatory outputs. The White House report suggests that companies developing AI could publish technical details of a system's design or limited data sets to be reviewed and tested for potential discrimination or discriminatory outcomes.

Other possible approaches include creating an independent body to review companies' proposed data sets or creating best practices guidelines for data inputs and the development of nondiscriminatory AI systems, following the

3 "Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc.," United States Supreme Court, June 25, 2015.

4 "Big Data: A Tool for Inclusion or Exclusion?," Federal Trade Commission, January 2016; "Opportunities and Challenges in Online Marketplace Lending," U.S. Department of the Treasury, May 10, 2016; "Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights," Executive Office of the President, May 2016; and "Preparing for the Future of Artificial Intelligence," Executive Office of the President, October 2016.

5 "Exploring Special Purpose National Bank Charters for Fintech Companies," Office of the Comptroller of the Currency, December 2016.

6 "Remarks Before the Marketplace Lending Policy Summit 2016," Thomas J. Curry, September 13, 2016.

7 "Preparing for the Future of Artificial Intelligence," Executive Office of the President, October 2016.

self-regulatory organization model that has been successful for the Payment Card Industry Security Standards Council.

Promising technological solutions are already emerging to help companies test and correct for bias in their algorithmic systems. Researchers at Carnegie Mellon University have developed a method called Quantitative Input Influence (QII) that can detect the potential for bias in an opaque algorithm.[8] QII works by repeatedly running an algorithm with a range of variations in each possible input. The QII system then determines which inputs have the greatest effect on the output.

Impressively, QII is able to account for the potential correlation among variables to identify which independent variables have a causal relationship with the output. Using QII on a credit-scoring algorithm could help a lender understand how specific variables are weighted. Indeed, QII could beget a line of tools that will enable financial services companies to root out bias in their applications, and perhaps minimize liability by demonstrating due diligence in connection with efforts to prevent bias.

Researchers at Boston University and Microsoft Research have developed a method by which human reviewers can use known biases in an algorithm's results to identify and offset biases from the input data.[9] The researchers used a linguistic data set that produces gender-biased results when an AI program is asked to create analogies based on occupations or traits that should be gender-neutral. The team had the program generate numerous pairs of analogies and used humans to identify which relationships were biased. By comparing gender-biased pairs to those that should exhibit differences (such as "she" and "he"), researchers constructed a mathematical model of the bias, which enables the creation of an anti-bias vaccine, a mirror image of the bias in the data set. When the mirror image is merged with the original data set, the identified biases are effectively nullified, creating a new and less-biased data set.

THE RISKS OF ALGORITHMIC BIAS ARE GLOBAL

While legal frameworks differ, the anti-discrimination principles embedded in US fair lending laws have non-US analogues. For example, the UK requires financial institutions to show proactively that fairness to consumers undergirds product offerings, suitability assessments and credit decisions.

Many jurisdictions have (or are considering) laws requiring institutions, including lenders, to allow individuals to opt out of "automated decisions" based on their personal data. Individuals who do not opt out must be notified of any such decision and be permitted to request reconsideration. Such automated decision-taking rights, which would likely apply to algorithmic creditworthiness models, are found in the UK Data Protection Act 1998 and EU General Data Protection Regulation of 2016. Australia may adopt similar regulations following the Productivity Commission's October 2016 Draft Report on Data Availability and Use.

There is little doubt that regulators and academics worldwide are focused on the potential bias and discrimination risks that algorithms pose. In a November 2016 report on artificial intelligence, for example, the UK Government Office for Science expressed concern that algorithmic bias may contribute to the risk of stereotyping, noting that regulators should concentrate on preventing biased outcomes rather than proscribing any particular algorithmic design. Likewise, UK and EU researchers are working to advance regtech approaches to manage the risks of potential algorithmic bias. For instance, researchers at the Helsinki Institute for Information Technology have created discrimination-aware algorithms that can remove historical biases from data sets.

Document the rationale for algorithmic features

As part of any proactive risk mitigation strategy, institutions should prepare defenses for discrimination claims before they arise. Whenever an institution decides to use an attribute as input for an algorithm that may have disparate impact risk, counsel should prepare detailed business justifications for using the particular attribute, and document the business reasons for not using alternatives. The file should also demonstrate due diligence by showing the testing history for the algorithm, with results that should support and justify confidence in its use.

Institutions using algorithmic solutions in credit transactions should consider how best to comply with legal requirements for providing statements of specific reasons for any adverse actions, as well as requirements for responding to requests for information and record retention.

8 "Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems," Carnegie Mellon University, April 2016.

9 "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings," Boston University and Microsoft Research, arXiv, July 2016.

Develop fintech that regulates fintech

Regulatory technology (regtech), the branch of fintech focused on improving the compliance systems of financial services companies,[10] is showing significant promise and already yielding sound approaches to managing the risks of algorithmic bias. Numerous financial services companies are expected to develop or leverage third-party regtech algorithms to test and monitor the smart algorithms they already deploy for credit transactions.

QII and the mirror-image method discussed above are only the first of many potential algorithm-based tools that will be incorporated into regtech to prevent algorithm-based discrimination. For example, a paper presented at the Neural Information Processing Systems Conference shows how predictive algorithms could be adjusted to remove discrimination against identified protected attributes.[11]

The operational challenges of implementing two algorithmic systems that continually feed into each other are no doubt complex. But AI has proven adept at navigating precisely these types of complex challenges.

Operational hurdles aside, institutions seeking to leverage AI-based regtech solutions to validate and monitor algorithmic lending will have the added challenge of getting financial regulators comfortable with such an approach. Although it may take some time before regtech solutions focused on fair lending achieve broad regulatory acceptance, regulators are increasingly recognizing the valuable role such solutions can play as part of a robust compliance management system.

Regulators are also becoming active consumers of regtech solutions, creating the potential for better coordination among regulators and unprecedented opportunities for regtech coordination between regulators and regulated institutions.

* * *

No financial services company wants to find itself apologizing to the public and regulators for discriminatory effects caused by its own technology, much less paying damages in the context of government enforcement or private litigation. To use smart algorithms responsibly, companies—particularly financial services firms—must identify potential problems early and have a well-conceived plan for addressing and removing unintended bias before it leads to discrimination in their lending practices, as well as potential discriminatory biases that may reach beyond lending and affect other aspects of a company's operations. ■

> **Promising technological solutions are already emerging to help companies test and correct for bias in their algorithmic systems.**

10 "Regtech rising: Automating regulation for financial institutions," White & Case LLP, September 2016.

11 "Equality of Opportunity in Supervised Learning," Neural Information Processing Systems Conference, December 2016.

NY0117/TL/R/289736_19

WHITE & CASE

Kevin Petrasic
Partner, Washington, DC
T +1 202 626 3671
E kpetrasic@whitecase.com

Benjamin Saul
Partner, Washington, DC
T +1 202 626 3665
E bsaul@whitecase.com

James Greig
Partner, London
T +44 20 7532 1759
E jgreig@whitecase.com

whitecase.com