For release on delivery
9:35 a.m. EST
January 12, 2021

Supporting Responsible Use of AI and Equitable Outcomes in Financial Services

Remarks by

Lael Brainard

Member

Board of Governors of the Federal Reserve System

at the

AI Academic Symposium
hosted by the Board of Governors of the Federal Reserve System

Washington, D.C.

Virtual Event

January 12, 2021

Today's symposium on the use of artificial intelligence (AI) in financial services is part of the Federal Reserve's broader effort to understand AI's application to financial services, assess methods for managing risks arising from this technology, and determine where banking regulators can support responsible use of AI and equitable outcomes by improving supervisory clarity.[1]

The potential scope of AI applications is wide ranging. For instance, researchers are turning to AI to help analyze climate change, one of the central challenges of our time. With nonlinearities and tipping points, climate change is highly complex, and quantification for risk assessments requires the analysis of vast amounts of data, a task for which the AI field of machine learning is particularly well-suited.[2] The journal *Nature* recently reported the development of an AI network which could "vastly accelerate efforts to understand the building blocks of cells and enable quicker and more advanced drug discovery" by accurately predicting a protein's 3-D shape from its amino acid sequence.[3]

Application of AI in Financial Services

In November 2018, I shared some early observations on the use of AI in financial services.[4] Since then, the technology has advanced rapidly, and its potential implications have come into sharper focus. Financial firms are using or starting to use AI for operational risk

---

[1] I am grateful to Kavita Jain, Jeff Ernst, Carol Evans, and Molly Mahar of the Federal Reserve Board for their assistance in preparing this text. These remarks represent my own views, which do not necessarily represent those of the Federal Reserve Board or the Federal Open Market Committee.

[2] David Rolnick, et al., "Tackling Climate Change with Machine Learning," https://arxiv.org/pdf/1906.05433; Sarah Castellanos, "Climate Researchers Enlist Big Cloud Providers for Big Data Challenges," *The Wall Street Journal*, November 25, 2020, https://www.wsj.com/articles/climate-researchers-enlist-big-cloud-providers-for-big-data-challenges-11606300202.

[3] Ewen Callaway, "'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures," *Nature* 588 (November 30, 2020): 203–204, https://www.nature.com/articles/d41586-020-03348-4.

[4] Lael Brainard, "What Are We Learning about Artificial Intelligence in Financial Services?" (remarks at Fintech and the New Financial Landscape, Philadelphia, Pennsylvania, November 13, 2018), https://www.federalreserve.gov/newsevents/speech/brainard20181113a.htm.

management as well as for customer-facing applications. Interest is growing in AI to prevent fraud and increase security. Every year, consumers bear significant losses from frauds such as identity theft and imposter scams. According to the Federal Trade Commission, in 2019 alone, "people reported losing more than $1.9 billion to fraud," which represents a mere fraction of all fraudulent activity banks encounter.[5] AI-based tools may play an important role in monitoring, detecting, and preventing such fraud, particularly as financial services become more digitized and shift to web-based platforms. Machine learning-based fraud detection tools have the potential to parse through troves of data—both structured and unstructured—to identify suspicious activity with greater accuracy and speed, and potentially enable firms to respond in real time.

Machine learning models are being used to analyze traditional and alternative data in the areas of credit decisionmaking and credit risk analysis, in order to gain insights that may not be available from traditional credit assessment methods and to evaluate the creditworthiness of consumers who may lack traditional credit histories.[6] The Consumer Financial Protection Bureau has found that approximately 26 million Americans are credit invisible, which means that they do not have a credit record, and another 19.4 million do not have sufficient recent credit data to generate a credit score. Black and Hispanic consumers are notably more likely to be credit invisible or to have an unscored record than White consumers.[7] The Federal Reserve's

[5] Federal Trade Commission, Consumer Sentinel Network, Data Book 2019, (Washington: Federal Trade Commission, January 2019), https://www.ftc.gov/system/files/documents/reports/consumer-sentinel-network-data-book-2019/consumer_sentinel_network_data_book_2019.pdf.

[6] See Board of Governors of the Federal Reserve System et al., "Interagency Statement on the Use of Alternative Data in Credit Underwriting," https://www.federalreserve.gov/supervisionreg/caletters/CA%2019-11%20Letter%20Attachement%20Interagency%20Statement%20on%20the%20Use%20of%20Alternative%20Data%20in%20Credit%20Underwriting.pdf.

[7] Kenneth P. Brevoort, Philipp Grimm, and Michelle Kambara, *Data Point: Credit Invisibles* (Washington: Consumer Financial Protection Bureau, May 2015), https://files.consumerfinance.gov/f/201505_cfpb_data-point-credit-invisibles.pdf.

Federal Advisory Council, which includes a range of banking institutions from across the country, recently noted that nontraditional data and the application of AI have the potential "to improve the accuracy and fairness of credit decisions while also improving overall credit availability."[8]

To harness the promise of machine learning to expand access to credit, especially to underserved consumers and businesses that may lack traditional credit histories, it is important to be keenly alert to potential risks around bias and inequitable outcomes. For example, if AI models are built on historical data that reflect racial bias or are optimized to replicate past decisions that may reflect bias, the models may amplify rather than ameliorate racial gaps in access to credit. Along those same lines, the opaque and complex data interactions relied upon by AI could result in discrimination by race, or even lead to digital redlining, if not intentionally designed to address this risk. It is our collective responsibility to ensure that as we innovate, we build appropriate guardrails and protections to prevent such bias and ensure that AI is designed to promote equitable outcomes. As Rayid Ghani notes, "…[A]ny AI (or otherwise developed) system that is affecting people's lives has to be explicitly built to focus on increasing equity and not just optimizing for efficiency…[W]e need to make sure that we put guidelines in place to maximize the chances of the positive impact while protecting people who have been traditionally marginalized in society and may be affected negatively by the new AI systems."[9]

[8] Federal Advisory Council (FAC) Record of Meeting, (December 3, 2020), https://www.federalreserve.gov/aboutthefed/files/fac-20201203.pdf.

[9] Rayid Ghani, "Equitable Algorithms: Examining Ways to Reduce AI Bias in Financial Services" (testimony before the House Committee on Financial Services Task Force on Artificial Intelligence Hearing on February 12, 2020), https://www.congress.gov/116/meeting/house/110499/witnesses/HHRG-116-BA00-Wstate-GhaniR-20200212-U1.pdf.

Black Box Problems

Recognizing the potential and the pitfalls of AI, let us turn to one of the central challenges to using AI in financial services—the lack of model transparency. Some of the more complex machine learning models, such as certain neural networks, operate at a level of complexity that offers limited or no insight into how the model works. This is often referred to as the "black box problem," because we can observe the inputs the models take in, and examine the predictions or classifications the model makes based on those inputs, but the process for getting from inputs to outputs is obscured from view or very hard to understand.

There are generally two reasons machine learning models tend toward opacity. The first is that an algorithm rather than a human being "builds" the model. Developers write the initial algorithm and feed it with the relevant data, but do not specify how to solve the problem at hand. The algorithm uses the input data to estimate a potentially complex model specification, which in turn make predictions or classifications. As Michael Tyka puts it, "[t]he problem is that the knowledge gets baked into the network, rather than into us. Have we really understood anything? Not really—the network has."[10] This is somewhat different from traditional econometric or other statistical models, which are designed and specified by humans.

The second is that some machine learning models can take into account more complex nonlinear interactions than most traditional models in ways that human beings would likely not be able to identify on their own.[11] The ability to identify subtle and complex patterns is what makes machine learning such a powerful tool, but that complexity often makes the model

---

[10] Davide Castelvevchi, "Can we open the black box of AI?" *Nature* 538 (October 5, 2016): 20–23, https://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731.

[11] Cynthia Rudin, "Please Stop Explaining Black Box Models for High-Stakes Decisions" (paper presented at 32nd Conference on Neural Information Processing Systems, Montreal, Canada, November 2018), https://www.researchgate.net/publication/329206654_Please_Stop_Explaining_Black_Box_Models_for_High_Stakes_Decisions/fulltext/5bfcc080458515b41d107a0a/Please-Stop-Explaining-Black-Box-Models-for-High-Stakes-Decisions.pdf?origin=publication_detail.

inscrutable and unintuitive. Hod Lipson likens it to "meeting an intelligent species whose eyes have receptors [not] just for the primary colors red, green, and blue, but also for a fourth color. It would be very difficult for humans to understand how the alien sees the world, and for the alien to explain it to us."[12]

The Importance of Context

While the black box problem is formidable, it is not, in many cases, insurmountable. The AI research community has made notable strides in explaining complex machine learning models—indeed, some of our symposium panelists have made major contributions to that effort. One important conclusion of that work is that there need not be a single principle or one-size-fits-all approach for explaining machine learning models. Explanations serve a variety of purposes, and what makes a good explanation depends on the context. In particular, for an explanation to "solve" the black box problem, it must take into account who is asking the question and what the model is predicting.

So what do banks need from machine learning explanations? The requisite level and type of explainability will depend, in part, on the role of the individual using the model. The bank employees that interact with machine learning models will naturally have varying roles and varying levels of technical knowledge. An explanation that requires the knowledge of a PhD in math or computer science may be suitable for model developers, but may be of little use to a compliance officer, who is responsible for overseeing risk management across a wide swath of bank operations.

The level and type of explainability also depends on the model's use. In the consumer protection context, consumers' needs and fairness may define the parameters of the explanation.

[12] Castelvevchi, "Can we open," 20–23.

Importantly, consumer protection laws require lenders who decline to offer a consumer credit—or offer credit on materially worse terms than offered to others—to provide the consumer with an explanation of the reasons for the decision. That explanation serves the important purposes of helping the consumer to understand the basis of the determination as well as the steps the consumer could take to improve his or her credit profile.[13]

Additionally, to ensure that the model comports with fair lending laws that prohibit discrimination, as well as the prohibition against unfair or deceptive practices, firms need to understand the basis on which a machine learning model determines creditworthiness. Unfortunately, we have seen the potential for AI models to operate in unanticipated ways and reflect or amplify bias in society. There have been several reported instances of AI models perpetuating biases in areas ranging from lending and hiring to facial recognition and even healthcare. For example, a 2019 study by Science revealed that an AI risk-prediction model used by the U.S. healthcare system was fraught with racial bias. The model, designed to identify patients that would likely need high-risk care management in the future, used patients' historical medical spending to determine future levels of medical needs. However, the historical spending data did not serve as a fair proxy, because "less money is spent on Black patients who have the same level of need, and the algorithm thus falsely concludes that Black patients are healthier than equally sick White patients."[14] Thus, it is critical to be vigilant for the racial and other biases that may be embedded in data sources.

It is also possible for the complex data interactions that are emblematic of AI—a key strength when properly managed—to create proxies for race or other protected characteristics,

[13] Among other things, the explanation can also make consumers aware of any erroneous information that drove the denial of credit.

[14] Ziad Obermeyer et al., "Dissecting racial bias in an algorithm used to manage the health of populations," *Science* 366 (October 25, 2019): 447–453, https://science.sciencemag.org/content/366/6464/447.

leading to biased algorithms that discriminate. For example, when consumers obtain information about credit products online, the complex algorithms that target ads based on vast amounts of data, such as where one went to school, consumer likes, and online browsing habits, may be combined in ways that indicate race, gender, and other protected characteristics.[15] Even after one online platform implemented new safeguards pursuant to a settlement to address the potential exclusion of consumers from seeing ads for credit products based on race, gender, or other protected characteristics, Professor Alan Mislove and his collaborators have found that the complex algorithms may still result in bias and exclusion.[16] Therefore, it is important to understand how complex data interactions may skew the outcomes of algorithms in ways that undermine fairness and transparency.

Makada Henry-Nickie, notes that "…[I]t is of paramount importance that policymakers, regulators, financial institutions, and technologists critically examine the benefits, risks, and limitations of AI and proactively design safeguards against algorithmic harm, in keeping with societal standards, expectations, and legal protections."[17] I am pleased that the symposium includes talks from scholars who are studying how we can design AI models that avoid bias and promote financial inclusion. No doubt everyone here today who is exploring AI wants to

[15] Carol A. Evans and Westra Miller, "From Catalogs to Clicks: The Fair Lending Implications of Targeted, Internet Marketing," Consumer Compliance Outlook, third issue, 2019, https://www.consumercomplianceoutlook.org/2019/third-issue/from-catalogs-to-clicks-the-fair-lending-implications-of-targeted-internet-marketing/.

[16] Piotr Szapiezynski et al., "Algorithms That 'Don't See Color': Comparing Biases in Lookalike and Special Ad Audiences," (2019), https://sapiezynski.com/papers/sapiezynski2019algorithms.pdf; Till Speicher, et al., "Potential for Discrimination in Online Targeted Advertising," Proceedings of Machine Learning Research 81:1–15, 2018 Conference on Fairness, Accountability, and Transparency, http://proceedings.mlr.press/v81/speicher18a/speicher18a.pdf.

[17] Makada Henry-Nickie, "Equitable Algorithms: Examining Ways to Reduce AI Bias in Financial Services" (testimony before the House Committee on Financial Services Task Force on Artificial Intelligence Hearing on February 12, 2020), https://www.congress.gov/116/meeting/house/110499/witnesses/HHRG-116-BA00-Wstate-Henry-NickieM-20200212-U1.pdf.

promote financial inclusion and more equitable outcomes and ensure that it complies with fair lending and other laws designed to protect consumers.

In the safety and soundness context, bank management needs to be able to rely on models' predictions and classifications to manage risk. They need to have confidence that a model used for crucial tasks such as anticipating liquidity needs or trading opportunities is robust and will not suddenly become erratic. For example, they need to be sure that the model would not make grossly inaccurate predictions when it confronts inputs from the real world either that differ in some subtle way from the training data or that are based on a highly complex interaction of the data features. In short, they need to be able to have confidence that their models are robust. Explanations can be an important tool in providing that confidence.

Not all contexts require the same level of understanding of how machine learning models work. Users may, for example, have a much greater tolerance for opacity in a model that is used as a "challenger" to existing models and simply prompts additional questions for a bank employee to consider relative to a model that automatically triggers bank decisions. For instance, in liquidity or credit risk management, where AI may be used to test the outcomes of a traditional model, banks may appropriately opt to use less transparent machine learning systems.

Forms of Explanations

Researchers have developed various approaches to explaining machine learning models. Often, these approaches vary in terms of the type of information they can provide about a model. As banks contemplate using these tools, they should consider what they need to understand about their models relative to the context, in order to determine whether there is sufficient transparency in how the model works to properly manage the risk at issue.

Not all machine learning models are a black box. In fact, some machine learning models are fully "interpretable" and therefore may lend themselves to a broader array of use cases. By "interpretable" I mean that developers can "look under the hood" to see how those models make their predictions or classifications, similar to traditional models. They can examine how much weight the model gives to each data feature, and how it plays into a given result. Interpretable machine learning models are intrinsically explainable.

In the case of machine learning models that are opaque, and not directly interpretable, researchers have developed techniques to probe these models' decisions based on how they behave. These techniques are often referred to as model agnostic methods, because they can be used on any model, regardless of the level of explainability. Model agnostic methods do not access the inner workings of the AI model being explained. Instead, they derive their explanations *post hoc* based on the model's behavior: essentially, they vary inputs to the AI model, and analyze how the changes affect the AI model's outputs.[18] In effect, a model agnostic method uses this testing as data to create a model of the AI model.[19]

While post hoc explanations generated by model agnostic methods can allow inferences to be drawn in certain circumstances, they may not always be accurate or reliable, unlike intrinsic explanations offered by interpretable models. Basing an explanation on a model's behavior rather than its underlying logic in this way may raise questions about the explanation's accuracy, as compared to the explanations of interpretable models. Still, such explanations may

[18] See Marco Tulio Ribeiro et al., "Model-Agnostic Interpretability of Machine Learning" (presented at 2016 ICML Workshop on Human Interpretability in Machine Learning, New York, New York, 2016), https://arxiv.org/abs/1606.05386; Zachary C. Lipton, "The Mythos of Interpretability" (presented at 2016 ICML Workshop on Human Interpretability in Machine Learning, New York, New York, 2016), https://arxiv.org/abs/1606.03490.

[19] See Cynthia Rudin, "Please Stop Explaining" and Christoph Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (Christoph Molnar), https://christophm.github.io/interpretable-ml-book/.

be suitable in certain contexts. Thus, one of the key questions banks will face is when a post hoc explanation of "black box" model is acceptable versus when an interpretable model is necessary.

To be sure, having an accurate explanation for how a machine learning model works does not by itself guarantee that the model is reliable or fosters financial inclusion. Time and experience are also significant factors in determining whether models are fit to be used. The boom-bust cycle that has defined finance for centuries should make us cautious in relying fully for highly consequential decisions on any models that have not been tested over time or on source data with limited history, even if in the age of big data, these data sets are broad in scope.

Expectations for Banks

Recognizing that AI presents promise and pitfalls, as a banking regulator, the Federal Reserve is committed to supporting banks' efforts to develop and use AI responsibly to promote a safe, fair, and transparent financial services marketplace. As regulators, we are also exploring and understanding the use of AI and machine learning for supervisory purposes, and therefore, we too need to understand the different forms of explainability tools that are available and their implications. To ensure that society benefits from the application of AI to financial services, we must understand the potential benefits and risks, and make clear our expectations for how the risks can be managed effectively by banks. Regulators must provide appropriate expectations and adjust those expectations as the use of AI in financial services and our understanding of its potential and risks evolve.[20]

[20] The Federal Reserve's Model Risk Management guidance (SR 11-7) establishes an expectation that models used in banking are conceptually sound or "fit for purpose." SR 11-7 instructs that when evaluating a model, supervised institutions should consider the "[t]he design, theory, and logic underlying the model." The Model Risk Management guidance discusses in detail the tools banks rely on to help establish the soundness of their models, such as back-testing and benchmarking and other outcomes-based tests.

To that end, we are exploring whether additional supervisory clarity is needed to facilitate responsible adoption of AI. It is important that we hear from a wide range of stakeholders—including financial services firms, technology companies, consumer advocates, civil rights groups, merchants and other businesses, and the public. The Federal Reserve has been working with the other banking agencies on a possible interagency request for information on the risk management of AI applications in financial services. Today's symposium serves to introduce a period of seeking input and hearing feedback from a range of external stakeholders on this topic. It is appropriate to be starting with the academic community that has played a central role in developing and scrutinizing AI technologies. I look forward to hearing our distinguished speakers' insights on how banks and regulators should think about the opportunities and challenges posed by AI.