

**PERSPECTIVES ON ARTIFICIAL INTELLIGENCE:  
WHERE WE ARE AND THE NEXT  
FRONTIER IN FINANCIAL SERVICES**

---

---

**HEARING**  
BEFORE THE  
TASK FORCE ON ARTIFICIAL INTELLIGENCE  
OF THE  
COMMITTEE ON FINANCIAL SERVICES  
U.S. HOUSE OF REPRESENTATIVES  
ONE HUNDRED SIXTEENTH CONGRESS  
FIRST SESSION

—————  
JUNE 26, 2019  
—————

Printed for the use of the Committee on Financial Services

**Serial No. 116-37**



—————  
U.S. GOVERNMENT PUBLISHING OFFICE

39-737 PDF

WASHINGTON : 2020

HOUSE COMMITTEE ON FINANCIAL SERVICES

MAXINE WATERS, California, *Chairwoman*

CAROLYN B. MALONEY, New York	PATRICK McHENRY, North Carolina,
NYDIA M. VELAZQUEZ, New York	<i>Ranking Member</i>
BRAD SHERMAN, California	PETER T. KING, New York
GREGORY W. MEEKS, New York	FRANK D. LUCAS, Oklahoma
WM. LACY CLAY, Missouri	BILL POSEY, Florida
DAVID SCOTT, Georgia	BLAINE LUETKEMEYER, Missouri
AL GREEN, Texas	BILL HUIZENGA, Michigan
EMANUEL CLEAVER, Missouri	SEAN P. DUFFY, Wisconsin
ED PERLMUTTER, Colorado	STEVE STIVERS, Ohio
JIM A. HIMES, Connecticut	ANN WAGNER, Missouri
BILL FOSTER, Illinois	ANDY BARR, Kentucky
JOYCE BEATTY, Ohio	SCOTT TIPTON, Colorado
DENNY HECK, Washington	ROGER WILLIAMS, Texas
JUAN VARGAS, California	FRENCH HILL, Arkansas
JOSH GOTTHEIMER, New Jersey	TOM EMMER, Minnesota
VICENTE GONZALEZ, Texas	LEE M. ZELDIN, New York
AL LAWSON, Florida	BARRY LOUDERMILK, Georgia
MICHAEL SAN NICOLAS, Guam	ALEXANDER X. MOONEY, West Virginia
RASHIDA TLAIB, Michigan	WARREN DAVIDSON, Ohio
KATIE PORTER, California	TED BUDD, North Carolina
CINDY AXNE, Iowa	DAVID KUSTOFF, Tennessee
SEAN CASTEN, Illinois	TREY HOLLINGSWORTH, Indiana
AYANNA PRESSLEY, Massachusetts	ANTHONY GONZALEZ, Ohio
BEN McADAMS, Utah	JOHN ROSE, Tennessee
ALEXANDRIA OCASIO-CORTEZ, New York	BRYAN STELL, Wisconsin
JENNIFER WEXTON, Virginia	LANCE GOODEN, Texas
STEPHEN F. LYNCH, Massachusetts	DENVER RIGGLEMAN, Virginia
TULSI GABBARD, Hawaii	
ALMA ADAMS, North Carolina	
MADELEINE DEAN, Pennsylvania	
JESÚS "CHUY" GARCIA, Illinois	
SYLVIA GARCIA, Texas	
DEAN PHILLIPS, Minnesota	

CHARLA OUERTATANI, *Staff Director*

TASK FORCE ON ARTIFICIAL INTELLIGENCE

BILL FOSTER, Illinois, *Chairman*

EMANUEL CLEAVER, Missouri  
KATIE PORTER, California  
SEAN CASTEN, Illinois  
ALMA ADAMS, North Carolina  
SYLVIA GARCIA, Texas  
DEAN PHILLIPS, Minnesota

FRENCH HILL, ARKANSAS, *Ranking  
Member*  
BARRY LOUDERMILK, Georgia,  
TED BUDD, North Carolina  
TREY HOLLINGSWORTH, Indiana  
ANTHONY GONZALEZ, Ohio  
DENVER RIGGLEMAN, Virginia



# CONTENTS

---

	Page
Hearing held on:	
June 26, 2019 .....	1
Appendix:	
June 26, 2019 .....	33

## WITNESSES

WEDNESDAY, JUNE 26, 2019

Buchanan, Bonnie, Head of Department of Finance and Accounting, Full Professor of Finance, Surrey Business School, The University of Surrey .....	6
McWaters, R. Jesse, Financial Innovation Lead, World Economic Forum .....	10
Merrill, Douglas, Founder and CEO, ZestFinance .....	8
Turner-Lee, Nicol, Fellow, Center for Technology Innovation, Brookings Institution .....	4

## APPENDIX

Prepared statements:	
Buchanan, Bonnie .....	34
McWaters, R. Jesse .....	46
Merrill, Douglas .....	54
Turner-Lee, Nicol .....	109

## ADDITIONAL MATERIAL SUBMITTED FOR THE RECORD

Budd, Hon. Ted:	
GAO report entitled, “Insurance Markets: Benefits and Challenges Presented by Innovative Uses of Technology,” dated June 2019 .....	127
Hill, Hon. French:	
ZestFinance article entitled, “Clarifying Why SHAP Shouldn’t Be Used Alone” .....	170



**PERSPECTIVES ON ARTIFICIAL  
INTELLIGENCE: WHERE WE ARE  
AND THE NEXT FRONTIER IN  
FINANCIAL SERVICES**

---

**Wednesday, June 26, 2019**

U.S. HOUSE OF REPRESENTATIVES,  
TASK FORCE ON ARTIFICIAL INTELLIGENCE,  
COMMITTEE ON FINANCIAL SERVICES,  
*Washington, D.C.*

The task force met, pursuant to notice, at 10 a.m., in room 2128, Rayburn House Office Building, Hon. Bill Foster [chairman of the task force] presiding.

Members present: Representatives Foster, Casten, Adams, Garcia of Texas, Phillips; Hill, Loudermilk, Budd, Hollingsworth, Gonzalez of Ohio, and Riggleman.

Also present: Representative Himes.

Chairman FOSTER. The Task Force on Artificial Intelligence will now come to order.

Without objection, the Chair is authorized to declare a recess of the task force at any time.

Also, without objection, members of the full Financial Services Committee who are not members of this task force are authorized to participate in today's hearing, consistent with the committee's practice.

Today's hearing is entitled, "Perspectives on Artificial Intelligence: Where We Are and the Next Frontier in Financial Services."

The Chair will now recognize himself for 5 minutes for an opening statement.

Thank you, everyone, for joining us today at the first hearing of the House Financial Services Committee's Task Force on Artificial Intelligence. And I would like to begin by thanking Chairwoman Waters and Ranking Member McHenry for working to establish this important task force and reaffirming this committee's commitment to understanding technological innovation in the financial services sector.

It is an exciting time to be on this committee. Today, the financial services sector is facing a period of rapid disruption and innovation, and artificial intelligence (AI) is at the heart of these changes.

AI is transforming the way Americans live, work, and interact with each other. As members of this committee, it is incumbent upon us to engage with and to understand more deeply how it

works, how it is designed and operated, and how it affects and may affect consumers.

When done right, AI can mean innovative underwriting models that allow millions more people access to credit and financial services. And at a time when there are still over 50 million unbanked or underbanked Americans, this is a big deal. Companies are also using AI to execute trades, manage portfolios, and provide personalized services to customers.

AI can be used to better detect fraud and money laundering, and regulators are using AI to improve market surveillance and policing of bad actors. This is important, because AI is also giving criminals more ways to impersonate customers and steal their assets and sensitive financial information.

Last year, there were almost 15 million victims of identity fraud, costing Americans billions of dollars. Social security numbers, credit card numbers, and other personal identity factors can be stolen and sold on the dark web or used by criminals for quick and easy profit.

That is why it is imperative that we come up with better ways of protecting and securing our digital identities online. In fact, I was just, in the last hour, giving a keynote speech at the Identiverse conference, where thousands of people come together each year to understand what technologies can be applied to allow both individuals and organizations to protect themselves from often AI-enabled identity fraud.

And now, as the name of this hearing suggests, the other part of this equation that we need to explore is, where is this technology going and what are the next frontiers?

To truly reach its potential to change the face of financial services, there are some questions we need to address. First, how can we be sure that AI credit underwriting models are not biased? Second, who is accountable if AI algorithms are just a black box that nobody can explain when it makes a decision? And third, AI runs on an enormous amount of data. Where does this data come from? How is it protected? Do customers know where it is being held, under what legal regime?

Also, AI works far better with large datasets. Will these large datasets be one more factor driving the consolidation of financial services sectors? I worry frequently that small community banks may end up going the way of small community newspapers.

Another thing we will be looking at is, how many and what kind of financial services jobs will AI displace? A recent study by Deloitte indicated that 75 percent of financial firms are planning to displace humans with technology, and this is probably not a trend that will slow down. And it is not only going to apply to bank tellers and entry-level people; it will apply to some of the very highest salaried positions.

And as I mentioned, just the question about whether small banks and startups will be able to compete with the big tech firms, particularly when everyone is going to need access to these very large, personally identifiable datasets.

Over the next 6 months, we will begin to examine these questions to gain a deeper understanding of how this technology is being used in the financial services industry. It is my hope that to-



day's dialogue between our diverse and bipartisan group of Members and the expert panel of witnesses joining us will lead to a better understanding of how AI is changing the industry, how it can lead to innovative and inclusive products and more personalized customer experience, and how this technology will shape the questions that policymakers will have to grapple with in the coming years.

And so at this time, I would like to recognize the ranking member of the task force, my colleague, Mr. Hill from Arkansas, who has been a valuable asset and a trusted bipartisan partner as we begin this important endeavor.

Mr. HILL. I thank the chairman. I appreciate you convening the hearing today and selecting this excellent panel before us. And I, too, want to thank our mutual leaders, Chairwoman Waters and Ranking Member McHenry, for their partnership in creating this task force.

Over the next few months, I look forward to working with you and our colleagues on both sides of the aisle to find ways to foster innovation through the use of artificial intelligence for both disruptive innovators and for our incumbent financial players, both small and large, as well as finding ways to use AI successfully to enhance our compliance obligations among our regulatory agencies.

The use of AI has grown exponentially in the last few years. AI has the potential to improve human life, economic competitiveness, and societal challenges.

Recent GAO testimony identified four high-consequence sectors where leveraging AI will bring significant benefits: cybersecurity; automated vehicles; criminal justice; and financial services. And today's timely hearing will discuss how AI is impacting and influencing financial services.

Artificial intelligence can be used to gather enormous amounts of data, detect abnormalities, and solve complex problems. Financial institutions are already experimenting extensively with AI strategies to enhance and streamline financial institutions, BSA and AML compliance, CRA requirements, fraud detection, and real estate valuations, all while reducing cost levels.

Also, AI can create better efficiencies for underwriting and reaching underbanked communities. Algorithmic-driven lending is proliferating online and transforming everything from personal loans to small business credit extension. A recent National Bureau of Economic Research working paper found that online financial companies discriminate 40 percent less than loan officers who make decisions face-to-face.

I know Dr. Merrill of ZestFinance, who grew up in my district in Arkansas, has been doing some interesting things in regard to AI and underwriting, and I look forward to hearing more from him today.

All that to say that the use of artificial intelligence and machine learning is not without challenges and questions, just like any other technology.

Dr. Henry Kissinger published an interesting article in *The Atlantic* recently outlining concerns about the rise of artificial intelligence. Dr. Kissinger argues that we are in the midst of a technological revolution that could culminate in a world "relying on ma-

chines powered by data and algorithms and ungoverned by ethical or philosophical norms.” He goes on to say that, “Truth becomes relative and information threatens to overwhelm wisdom.” Well, we are not into overwhelming wisdom in anything we do on Capitol Hill.

While it remains to be seen whether Dr. Kissinger’s concerns are fully proved, I think we should heed his advice. As policymakers, we need to ensure that we are asking the right questions about appropriate testing and evaluating of new technology, so that the ultimate benefits are, in the end, benefiting consumers.

We need to ensure that AI does not create biases in lending toward discrimination and that prudential regulators and market participants have an understanding of the underlying technology, model validation, and how algorithmic decisions are being made and the manner of the audit trail. These questions must be analyzed.

Lastly, I would be remiss if I didn’t mention the potential of job losses connected with the advent of artificial intelligence. I am sure this topic will arise throughout our hearings during the Congress.

The World Economic Forum argues that machines and algorithms in the workplace are expected to create 130 million new roles in work, but cost about 75 million jobs to be displaced by 2022, which means net 58 million jobs might be created. In my view, this will contribute positively on the economy and the future of work in the long run.

People might be putting the cart before the horse on the number of net displacements. I start this journey in the “cup half full” camp, and I am optimistic about our future.

I look forward to continuing to seek out answers throughout our work on the task force. I thank my good friend, Dr. Foster, for his partnership. And I look forward to finding bipartisan solutions to these many interesting and challenging questions in financial services.

I yield back.

Chairman FOSTER. Thank you.

Today, we welcome the testimony of Dr. Nicol Turner-Lee, fellow at the Center for Technology Innovation, Brookings Institution; Dr. Bonnie Buchanan, head of the School of Finance and Accounting and full professor of finance at the Surrey Business School, University of Surrey; Dr. Douglas Merrill, founder and CEO of ZestFinance; and Mr. Jesse McWaters, financial innovation lead at the World Economic Forum.

Witnesses are reminded that your oral testimony will be limited to 5 minutes, and without objection, your written statements will be made a part of the record.

So, Dr. Turner-Lee, you are now recognized for 5 minutes to give an oral presentation of your testimony.

**STATEMENT OF NICOL TURNER-LEE, FELLOW, CENTER FOR TECHNOLOGY INNOVATION, BROOKINGS INSTITUTION**

Ms. TURNER-LEE. Thank you very much, distinguished members of the task force, and thank you for this opportunity to speak before you on artificial intelligence and the application of autonomous systems in the financial services sector.

With a history of over 100 years, we at Brookings are committed to evidence-based nonpartisan research in this area, and my particular area of focus is on algorithmic bias. So, I appreciate the opportunity to speak before you.

Increasingly, the public and private sectors are turning to AI and machine-learning algorithms to automate simple and complex decision-making processes. The mass scale digitization of data and the emerging technologies that use them are disrupting most economic sectors, including transportation, retail, advertising, financial services, and energy.

These massive datasets have made it easy to derive new insights through computers, and as a result, machine-learning algorithms, which are step-by-step instructions that computers follow to perform a task, have become more sophisticated and pervasive tools for automated decision-making.

While many of us are aware of the context in which they are used, from making recommendations about movies, to credit products, these models make inferences from data about people including their identities, their demographic attributes, their preferences, and their likely future behaviors, as well as the objects related to them. And from that data, it learns a model which then can be applied to other people and objects, making what they believe to be accurate predictions.

But because machines can treat similarly situated people and objects differently, we are starting to reveal, much like has been said, some troubling examples in which the reality of algorithmic decision-making falls short of our expectations or is simply wrong.

In the case of credit, we are seeing people denied credit due to the factoring of digital composite profiles, which include their web browsing histories, social media profiles, and other inferential characteristics in the factoring of credit models, and these biases are systematically finding themselves with less favor to individuals within particular groups where there is no relevant difference between those groups which justifies those harms.

While my written testimony goes into more detail about this, I would just like to share in my remaining few minutes how we can create more fair, ethical, and just algorithmic models. From this perspective, if we do not do such at this time, we have the potential to replicate and amplify stereotypes historically prescribed to people of color and other vulnerable populations.

Let me start with an initial truth about emerging technologies: Despite their greater facilitation of efficiency and cognition, the online economy has not resolved the issue of racial bias. And we see that in terms of search inquiries that have classified African Americans as primates in the past.

These controversies are primarily due to the microtargeting of certain populations that go awry, even when they are not deliberate. Some of it can happen on an explicit level, where the algorithm may not start out being discriminatory in intent but adapts to the societal stereotypes and unfair profiling. In the case of credit, Latanya Sweeney at Harvard University has said that African Americans may find themselves the subject of higher-interest credit cards and other financial products simply because the computer has inferred their race.

In the issue of implicit or unconscious bias, we simply do not have enough people working in this field to help us make the right decisions, which goes back to the inclusivity and the diversity and design of these models.

Given this—and, again, in my written testimony I speak to the ways and the reasons of these biases, whether it is skewed training data, whether it is the fact that we have less counterfactual data that is actually going into training the algorithm—these issues are nonetheless troubling and dangerous, particularly for vulnerable populations like African Americans and Latinos, who have been ill-served within the financial services market. Most of these populations tend to be unbanked compared to whites, underbanked, and lack access to home ownership.

If you think about the physical redlining that happens oftentimes offline, what does it mean, as Frank Pasquale has called weblining or applications discrimination, when we begin to look at the algorithmic economy?

What do we do about this so that we avoid unfair credit rationing, exclusionary filtering, digital redlining? I would just like to offer just three recommendations that I would love to answer additional questions around that may be helpful.

First and foremost, Congress must modernize civil rights laws and other consumer protections to safeguard protected classes from online discrimination. We have laws like the Equal Credit Opportunity Act, the Fair Housing Act, and other laws, which I feel have to be modernized in the digital age to ensure equity and fairness.

We also need companies to exercise self-regulatory behaviors, whether it is looking at the auditing of their algorithms, bringing in more human content moderators, or finding ways to advance exclusivity.

And finally—and I will save this again for questions—I think it is important that we are more deliberate in bringing in diverse populations, partnering with Historically Black Colleges and Universities (HBCUs) and other minority-serving institutions, to ensure that we have more people at the table in the design of these models.

Thank you very much, and I look forward to questions.

[The prepared statement of Dr. Turner-Lee can be found on page 109 of the appendix.]

Chairman FOSTER. Thank you.

Dr. Buchanan, you are now recognized for 5 minutes to give an oral presentation of your testimony.

**STATEMENT OF BONNIE BUCHANAN, HEAD OF DEPARTMENT OF FINANCE AND ACCOUNTING, FULL PROFESSOR OF FINANCE, SURREY BUSINESS SCHOOL, THE UNIVERSITY OF SURREY**

Ms. BUCHANAN. Thank you, Chairman Foster.

Distinguished members of the task force, thank you for the opportunity to appear before you and provide testimony to help inform discussion about artificial intelligence in the financial services industry.

I am Dr. Bonnie Buchanan, professor of finance at the University of Surrey Business School, and I will provide some insights on arti-

ficial intelligence, its applications in financial services, as well as its challenges and opportunities. And I hope we can all work together to address those challenges and opportunities.

Artificial intelligence is rapidly impacting the financial services industry in a profound way, through banking, insurance, wealth management, personal financial planning, and regulation. It can be broadly thought of as a group of related technologies, including machine learning and deep learning.

Machine learning deals with general pattern recognition and universal approximation of relationships. One such example details teaching an algorithm to learn from past regulatory breaches and to predict new breaches, such as insider trading or cartels.

Regulators use clustering algorithms to better understand trades and categorize bank business models in advance of regulatory examinations. Chatbots, powered by natural language processing algorithms, have become powerful tools which provide a personalized and conversational experience to users.

Deep-learning algorithms automate routine tasks, mitigate risk, and help prevent fraud. It is based on neural networks, which are based on mimicking the way the multiple layers of the brain's neurons work. And neural networks have been used in financial distress models.

Artificial intelligence offers the possibility of greater financial inclusion, but its rapid growth and an already very complex financial system presents major challenges regarding regulation and policy-making, and risk management, as well as ethical, economic, and social hurdles. For one, the financial services workplace is going to look very different in the short and long term, with artificial intelligence augmenting many positions.

Machine-learning algorithms can also potentially introduce bias and discrimination. Deep learning provides predictions, but it does lack insight as to how the variables are being used to reach these predictions. Hiring and credit-scoring algorithms can exacerbate inequities due to biased data. Policymakers need to be concerned about the explainability of artificial intelligence models, and we should avoid black-box modeling where humans cannot determine the underlying process or outcomes of the machine-learning or deep-learning algorithms.

And resolving such issues as discrimination and bias requires being grounded in ethics and understanding what causes the bias in the algorithm in the first place. When it comes to artificial intelligence in financial services and a fairer future, policymakers need to be concerned about explainability, accountability, and, indeed, even auditability of artificial intelligence modeling.

Many artificial intelligence techniques remain untested in a financial crisis scenario. My written testimony discusses several instances where algorithms implemented by financial firms appeared to act in ways quite unforeseen by their developers, leading to errors and flash crashes.

Cybercrime costs the global economy over \$400 billion, but many banks have started to successfully turn to artificial intelligence techniques to address fraud through AI-based voice phishing detection apps.

Artificial intelligence and machine learning's rapid development are to such an extent where it is almost outstripping the current regulatory framework. But if we look overseas, we have in the United Kingdom the introduction of open banking, which gives consumers the ability to compare product offerings and exchange data between providers in a secure way.

Under the General Data Protection Rules (GDPR), EU citizens have the right to receive an explanation for decisions based solely on automatic processing. Furthermore, GDPR stipulates that companies must first obtain consent from an EU citizen before using their data, and failure to comply with GDPR rules can result in substantial fines.

The European Market in Financial Instruments Directive Part II requires that firms that apply artificial intelligence and algorithmic models have a robust development plan in place.

As big data and computing power increases, artificial intelligence needs to be technically robust, secure, protect privacy, and be ethically sound and regulation-compliant. We must not forget the importance of better digital and financial literacy, and ultimately, it needs to emphasize financial inclusion.

Thank you very much for your time today, and I appreciate the opportunity to share my thoughts with you later. Thank you.

[The prepared statement of Dr. Buchanan can be found on page 34 of the appendix.]

Chairman FOSTER. Thank you.

Dr. Merrill, you are now recognized for 5 minutes to give an oral presentation of your testimony.

**STATEMENT OF DOUGLAS MERRILL, FOUNDER AND CEO,  
ZESTFINANCE**

Mr. MERRILL. Chairman Foster, Ranking Member Hill, and members of the task force, thank you for the opportunity to appear before you to discuss the use of artificial intelligence in financial services.

My name is Douglas Merrill. I am the CEO of ZestFinance, which I founded 10 years ago with a mission to make fair and transparent credit available to everyone.

Lenders use our software to increase approval rates, lower defaults, and to make their lending fairer. Before ZestFinance, I was the chief information officer at Google. I have a Ph.D. in artificial intelligence from Princeton University.

The use of artificial intelligence in the financial industry is growing. Today, I will discuss a type of AI, machine learning, also known as ML, that discovers relationships between many variables in a dataset to make better predictions.

Because ML-powered credit scores substantially outperform traditional credit scores, companies will increasingly use ML to make more accurate decisions. For example, customers using our ML underwriting tools to predict creditworthiness have seen a 10 percent approval rate increase for credit card applications, a 15 percent approval rate increase for auto loans, and a 51 percent increase in approval rates for personal loans, each with no increase in defaults.

Overall, this is good news and should be encouraged. Machine learning increases access to credit, especially for low-income and

minority borrowers. Regulators understand these benefits and, in our experience, want to facilitate, not hinder, the use of ML.

But at the same time, ML raises serious risks for institutions and consumers. ML models are opaque and inherently biased. Lenders put themselves, consumers, and the safety and soundness of our entire financial system at risk if they do not appropriately validate and monitor ML models.

Getting this mix right, enjoying ML's benefits while employing responsible safeguards, is very difficult. Specifically, ML models have a black-box problem. Lenders know only that an ML algorithm made a decision, not why it made that decision.

Without understanding why a model made a decision, bad outcomes will occur. For example, a used car lender we work with had two seemingly benign signals in their model. One signal was that higher-mileage cars tend to yield higher-risk loans. Another was that borrowers from a particular State were slightly less risky than those from other States. Neither of these signals raised compliance concerns.

However, our ML tools noted that, taken together, these signals predicted a borrower to be African American and more likely to be denied.

Without visibility into how seemingly fair signals interact, lenders will make decisions which tend to adversely affect minority borrowers.

There are purported to be a variety of methods for understanding how ML models make decisions. Most don't actually work. As explained in our white paper and a recent essay on a technique called SHAP, both of which I have submitted for the record, many explainability techniques are inconsistent, inaccurate, computationally expensive, or fail to spot discriminatory outcomes.

At ZestFinance, we have developed explainability methods that render ML models truly transparent. As a result, we can assess disparities in outcomes and create less discriminatory models. This means we can identify approval rate gaps in protected classes such as race, national origin, and gender, and then minimize or eliminate those gaps. In this way, ZestFinance's tools decrease disparate impacts across protected groups and ensure that the use of machine learning-based underwriting mitigates rather than exacerbates bias in lending.

Congress could regulate the entirety of ML in finance to avoid bad outcomes, but it need not do so. Regulators have the authority necessary to balance the risks and benefits of ML underwriting.

In 2011, the Federal Reserve, the OCC, and the FDIC published guidance on effective model risk management. ML was not commonly in use in 2011, so the guidance does not directly address best practices in ML model development, validation, and monitoring.

We have recently produced a short FAQ, which we have also submitted for the record, that suggests updates to bring the guidance into the ML era. Congress must encourage regulators to set high standards for ML model development, validation, and monitoring.

We stand upon the brink of a new age of credit, an age that is fairer and more inclusive, enabled by this new technology of machine learning. However, "brink" can also imply the edge of a cliff.

Without rigorous standards for understanding why models work, ML will surely drive us over the edge. Every day that we wait to responsibly implement ML keeps tens of millions of Americans out of the credit system or poorly treated by it.

Thank you so much for your time.

[The prepared statement of Dr. Merrill can be found on page 54 of the appendix.]

Chairman FOSTER. Thank you.

And, Mr. McWaters, you are now recognized for 5 minutes to give an oral presentation of your testimony.

**STATEMENT OF R. JESSE MCWATERS, FINANCIAL  
INNOVATION LEAD, WORLD ECONOMIC FORUM**

Mr. MCWATERS. Thank you.

Chairman Foster, Ranking Member Hill, distinguished members of this task force, I am honored to be invited to appear before you today to discuss this important topic.

I would like to share with you in a personal capacity key insights from an ongoing research initiative that I lead at the World Economic Forum. These findings are drawn from 18 months of interviews and workshops with leading thinkers from large financial institutions, fintech innovators, large technology firms, and regulatory authorities, from all around the world.

It is manifestly clear that artificial intelligence is transforming the operating models of financial institutions. It is being deployed to improve the speed and efficiency of financial processes, to improve the accuracy of financial predictions, to create more accessible and personalized advisory capabilities, and to establish entirely new business offerings.

Less visible, but even more important, are the potential long-term impacts of AI on the competitive dynamics of the financial ecosystem. As AI becomes more central to the differentiation strategies of financial institutions, their appetite for deeper and broader datasets will increase, making access to this data a competitive imperative for all financial institutions.

Over time, artificial intelligence may even redraw the map of what we consider the financial sector. For example, small and mid-sized financial institutions which are unable to invest in becoming AI leaders may instead choose to employ the AI capabilities of third parties on an “as a service” basis. The providers of these services could be large technology firms, they could be specialized fintechs, or even competing financial institutions.

Moreover, the tendency of AI businesses to rapidly scale via the so-called AI “flywheel effect” means that successful service providers of this kind could rapidly become central to the operations of many financial institutions, resulting in a deep change to the systemic structure of the financial system.

These seismic shifts in the landscape of financial services obviously create new risks. The enormous complexity of some advanced AI systems can make them opaque, challenging traditional models of regulation and compliance.

The use of ever broader datasets introduces risks to user privacy, as well as to the introduction of unintended bias into financial decision-making. Furthermore, an inherently specialized and inter-



connected financial system creates new vectors for both the accumulation and the propagation of systemic risk.

However, while these threats are very real and should be taken seriously, it is critical that we avoid knee-jerk reactions informed by fear.

In my view, the advent of AI does not call into question the fundamental principles that inform our regulatory framework. Rather, it demands that we be open to using both existing and emerging techniques to ensure that we remain aligned to these principles, even against a backdrop of rapid technological change.

Moreover, AI's risks must be considered alongside the opportunities that it creates. AI has the potential to help motorists get the money that they need from an insurance claim more quickly after an accident, to help immigrants without an established credit history access financing, and to make high-quality financial advice, so needed, more accessible for everyday Americans.

Moreover, the ability to outsource selected functions to specialized third parties has the potential to help smaller community banks remain digitally relevant to their customers.

Ultimately, AI is a tool. As with all powerful tools, preventing misuse is of the utmost importance. But with the right governance and oversight, I believe that AI has the potential to do enormous good for the financial sector.

Thank you.

[The prepared statement of Mr. McWaters can be found on page 46 of the appendix.]

Chairman FOSTER. Thank you.

And I now recognize myself for 5 minutes for questions.

Dr. Turner-Lee and Dr. Merrill, the National Bureau of Economic Research Working Paper recently published by UC Berkeley found that the algorithmic lending models discriminate in their case 40 percent less than face-to-face lenders for mortgage and refinancing loans.

If that sort of result proves generally true, it is positive news for consumers, especially African-American and Latino consumers, who pay \$765 million in additional interest costs each year.

And it highlights the fact that the artificial intelligence algorithms don't have to be perfect as long as they are significantly better than the current procedures. That is obviously a moving target, because as our underwriting gets better and more fair over time, I think we have to continue to ask machine-learning techniques to continually up their game as well.

And so my question is, to what extent companies should be required to audit these algorithms so that they don't unfairly discriminate? Who should determine the standards for that? What is the current understanding of best practices?

Ms. TURNER-LEE. Thank you, Mr. Chairman, for that question.

I am actually also delighted to see that we are seeing research that is actually saying that we are leveraging some of the disparities when it comes to the use of AI. But I, too, am cautious, because I think the institution of auditing practices are really what is needed to ensure that we are not seeing these unintended consequences of racial or ethnic bias against different economic classes actually happening.

I would say to you that we are seeing more self-regulatory models where companies are actually coming in and engaging in auditing. I would also recommend, as I said earlier, that we see developers look at how the algorithm is in compliance with some of the nondiscrimination laws prior to the development of the algorithm, which would also help to audit out some bias at the onset.

A paper that we recently released also combines auditing with a bias impact statement. There is a lot more proactive conversation prior to the launch of the product into the public domain.

Chairman FOSTER. How close are we to having generally agreed-upon metrics for things like fairness? I remember encountering a paper that claimed to have 15 different definitions of fairness.

Ms. TURNER-LEE. Right.

Chairman FOSTER. So, how do we decide which one of those is most applicable?

Ms. TURNER-LEE. That is a question with which I think all of us on this panel today struggle. How do you look at fairness and equity tradeoffs? Where do you find that there is a product that is not creating more discrimination versus less? And how do you document what those models are?

I think at this stage, our discussion around explainability and accountability is one part of it. But I think, to your point, getting companies as well as consumers engaged, creating more feedback loops so that we actually go into this together, I think is a much more proactive approach than trying to figure out ways to clean up the mess and the chaos at the end where we are discriminating against more people, we are incarcerating more people, and we are denying credit to more people. We have to figure out how to get ahead of this game.

Chairman FOSTER. Dr. Merrill?

Mr. MERRILL. I think it is quite clear that machine-learning models are biased. They are biased for three primary reasons.

First, they are biased because historically, white men have dominated the credit roles in the past, so that back data is a bad representation of the world.

Second, they are biased because machine-learning models tend to use a large number of signals of variables and there has to date been relatively little best practice around, how do you analyze those variables, because many times one or more of them will covary to yield a protected class.

And third, they are biased because most ML models are produced by the proverbial “white guy in a hoodie.” I, by the way, own a hoodie, but I try really hard not to be biased.

I think, absolutely, we must have an audit requirement, and I actually think a creation up-front requirement, in the way that we today have build requirements for financial services. FCRA produces quite striking, quite clear laws on what we are allowed to do.

I would hope that either through congressional intervention or regulatory intervention, we would come to a world in which there would be a language to describe what is acceptable before you build models and then an agreed-upon language at the end of models to show if, in fact, you have a bias problem, because again, the odds are good you are going to.

Chairman FOSTER. Mr. McWaters and Dr. Buchanan, both of you have worked on the issue of whether or not the access to large datasets is going to drive consolidation. Dr. Buchanan, you have written on China, where they have simply let things consolidate and let the access to enormous amounts of data result in a very small number of very large players.

Are there policy options that we can do to lean against that consolidation, in my negative 2 seconds? If you could just say one sentence, like, read my testimony or something?

Ms. BUCHANAN. I do talk about this in my written testimony and also my Turing report, Chairman Foster.

But I think we also have to understand what makes China so different, too. Its supply of data, its online population is twice the size of the United States. WeChat hosts over a billion users. And they have also—

Chairman FOSTER. Okay. Now, I will have to; I am going to use my power of the gavel on myself.

Ms. BUCHANAN. Yes, there are. We can, yes.

Chairman FOSTER. All right.

Now, I am happy to yield 5 minutes to Ranking Member Hill.

Mr. HILL. Thank you, Mr. Chairman.

This is a really good discussion, and I think that it is exactly why we have this task force, to talk through these issues.

And also, we invite our regulators to be full participants. All of you have made that suggestion. And I think we saw yesterday that they are eager to do that as they appoint their own innovation officers, their own legal teams who are thinking through this set of issues.

We are talking about innovation, we are talking about small and large, and then we are also talking about pursuing innovation, yet, obviously, complying with all the laws that we have in the country. And these are doable things, right?

Nobody seeks to create a model with bias in it. In fact, they have a legal obligation not to do that. So, there is no group of people, hoodies or no hoodies, who are out there seeking to generate a credit model that has bias in it.

But, Dr. Merrill, you make good points about this.

This is a problem in government, too. Let's talk about the Consumer Financial Protection Bureau (CFPB), just a few years ago in their settlements with Honda and Toyota, where they used big data to estimate somebody who might have been a source of bias in auto finance—using big data, not real customer data, and just assumed that if your name is “Hill” and you are from “72207”, you might have a chance of getting a reimbursement from one of these settlements, based on bias. It was fallacious, and I think this committee was stunned by that a few years ago.

We know in government and the private sector, this is a real challenge.

Dr. Merrill, you talked about the model development and updating the regulatory guidance, and you have shared your work. How do we invite those regulators to put out for a rulemaking on updating that 2011 guidance? How would you propose that we encourage that?

Mr. MERRILL. My team who built the updates and I have spent a long time meeting with essentially all of the regulators, prudential and non-prudential. And one of the things that we have found is, I think if you wandered around Silicon Valley and asked, people would say, oh, regulators are against innovation. And that has not been my experience at all. The question has been, how do they do the changes in a way which serves them well, their regulated institutions well, and Congress well?

For me, I think the single most important element moving forward is regulatory certainty. And I think it is impertinent of me to suggest what Congress should do, although I am "72032", so—

Mr. HILL. There we go.

Mr. MERRILL. —slightly different.

But even a small push to the regulators to say, we believe ML is coming and we believe your methods of ensuring fairness, of validating for FCRA and ECOA, and of making the promise of ML win, would be a substantial step forward.

Mr. HILL. That is why I support the sandbox idea. I think you all do, because you learn by doing. Of course, we are alleging the machines are learning by doing too. So, it is a way to backtest the reality, and I think sandboxes are useful. We would like to see sandbox uniformity among the agencies and a process that is open and not just—although I like the regulatory competition. In our society, it seems to be good. But we need to press on with that.

Also, I was comforted in a recent meeting with one of the Federal Reserve district banks that, don't forget, we have a lot of depository institutions that are buying credit that is originated in this way on their books. This is a good market test right now because we are looking at that data, we are doing our HMDA, our fair lending analysis against those purchase loans. And that is a way to get grassroots data as well.

Mr. McWaters, with 18 months of research focusing around the world on this, could you expand a little bit on why you are in the cup-half-full camp as well on long-term employment trends that we need? Give us some examples of these jobs that are being created that may see roles changed.

Mr. McWATERS. I can't speak to the specific methodology of the report that you mentioned. However, I think that it is actually quite useful when we think about this to reflect on history. The ATM was first introduced into the financial sector in the late 1960s, and there were some who predicted that we would no longer have branches, we would no longer have people in those branches.

What has happened instead is that the role that the individuals in those branches perform is markedly different than it was 20 or 30 years ago. It is no longer focused on basic transaction processing, but instead focused on advice and new sales origination.

And I think there are many examples of where we will see the fundamental activities of a job, the things people spend their time on, change and shift. That will likely require re-skilling and re-training. But we won't see the job in and of itself removed.

Mr. HILL. Thank you.

Mr. Chairman, thank you, and I yield back my time.

Chairman FOSTER. Thank you.

The gentleman from Illinois, Mr. Casten, is recognized for 5 minutes.

Mr. CASTEN. Thank you very much. And thank you so much to Representative Foster for your leadership on this issue. It is truly a privilege to serve on this committee. And thank you to all the members.

I have to start with a story. I ran an energy company for a number of years, and we had about 60 customers. Our biggest source of budget variance every year was our inability to predict how much energy our customers were going to use.

And, nerd that I am, I built a big genetic algorithm. We tweaked it. And ultimately, we were able to massively cut the revenue variance in ways that scared the pants off my customers, because they had no idea how we did this, and neither did I.

I mention that because what we found—I designed this to solve for a question of how to get better accuracy in our revenue forecast, and it did that beautifully.

The more granular I got, the more inaccurate it was. If I asked what a specific customer was going to be, it was a little goofier. If I asked what a specific customer's consumption of chilled water would be, it would be goofier still. And if I said what a specific customer's chilled water consumption was in May, it was off the charts.

Now, we knew well enough not to use it to ask those latter questions. But, Dr. Turner-Lee, a lot of what you described is that we have these tools that we built to ask one set of questions, which are really good. How do we improve our credit evaluation? How do we improve our underwriting? But then we have unintended consequences when we dig down to say, what does this say about a specific individual? And I don't know how to decouple that in the underwriting realm.

But I guess my question for you is, do you see ways, computationally or regulatorily, to say, if we design this to do one set of things, let's use it for that thing and be aware to where the blind spots are, just because of the nature of the math? These could be totally unintended. But how do we constrain it in that way? Your thoughts?

Ms. TURNER-LEE. Yes, I think that is an interesting question. It is a regulatory question that we are looking at in the privacy discussion right now, the extent to which consumers give so much data that there are no start and stop points with the accumulation of that.

I would echo what the panelists have said about the opaque nature of algorithms. And to your point, Congressman, what we are seeing is once it goes deeper into the ocean, the inferences that come out of that data are what is troubling, and are what lead to those unintended consequences.

So, we have to find ways to cure that. Do we allow consumers to tell us when that start/stop is with regard to use of their data? And the comment earlier about regulatory sandboxes, do we permit for anti-bias experimentation the use of demographic information when we know it is actually going to help us curb bias in ways that would be detrimental to certain populations?

I think, as you are talking about, the more granular we get, the less accurate we are, because there are certain data blind spots, as you suggested, that we are just not getting at. And the way that the technology works with machine-learning algorithms is, it assumes because a person or subject or object has engaged in that way, that that is who they are.

And that is where we find ourselves replicating and amplifying the stereotypes externally, because it is not the algorithm that is saying to itself, "I am going to be biased today." It is who we are as a society and who is actually inputting that data to create what has been considered the "garbage-out" variables.

Mr. CASTEN. The second question is for Dr. Merrill or McWaters, you guys can arm wrestle over who gets to answer this one.

None of you mentioned algorithmic trading. Some friends and colleagues who are in that space have described it to me as being: number one, awesome; and number two, completely unhedgeable, because it is totally blind to black swan events, because of the conversations that you mentioned. It overweights recent data, it overweights success, and, therefore, is both blind to black swans and, as my friend who shall remain nameless said, potentially creates some really bizarre social outcomes. Because if you are managing a socially responsible fund, and all of a sudden your algorithm is trading on a bet that we are going to invade Crimea next week, you know, weird things happen.

How do you think we should be regulating algorithmic trading in terms of the underlying risk, how much can we let it penetrate the market, and what do you do with an algorithm that is trading in a way that people may not actually understand what the bet is?

Mr. MCWATERS. I think that this is an excellent point and one that requires further investigation. We have seen in this space a tendency for machine-to-machine interactions to lead to feedback loops that have damaging impacts.

We have also seen that the innate foreignness that you have referred to in terms of the way that an AI-enabled model thinks can create confusion between fast-moving AI and slow-moving individuals, where people effectively freeze in response to an unexpected event. And that freezing is then interpreted as a further negative signal by the AI, driving things to an even more difficult situation.

Core to addressing this, in my mind, is scenario-based modeling and the types of stress-testing approaches that we have used in the past.

Mr. CASTEN. I am out of time, so I thank you.

And I yield back.

Chairman FOSTER. The gentleman from Ohio, Mr. Gonzalez, is recognized for 5 minutes.

Mr. GONZALEZ OF OHIO. Thank you, Mr. Chairman.

And thank you, everybody, for being here.

I am really excited about the direction of this task force and the leadership on both sides of the aisle from Dr. Foster and my colleague French Hill, and just really excited. And thank you for convening this.

One of my big priorities here on the committee has always been finding ways to expand affordable credit to low- and moderate-income borrowers. I think that has been one of the more difficult

challenges that we have faced as a society, certainly in the financial services sector, for a very long time.

And part of why I am excited about machine learning is what, Dr. Merrill, you suggested, which is that we can do this. This is something that is attainable. But there are certainly questions.

In your testimony, you talked about how there are “explainability models” that aren’t really doing a great job, but at ZestFinance you have developed one or you have developed methods that render ML models truly transparent, to directly quote you.

My question is more on the technical side. Technically speaking, how difficult is it to create a proper explainability model, knowing that, from my time in tech—I used to work in tech, not at your level—an A-plus engineer is kind of worth about 10 midlevel engineers, if you will.

Talk to me about the technical side of this, if you would?

Mr. MERRILL. Thank you for that question.

I think the way to think about it is to just kind of draw some broad boundaries about the question at first. One of the techniques that differs in machine learning from traditional underwriting is you use a bunch more data, and data is sometimes called signals.

And when you are going to do explainability, conceptually, the hard part isn’t actually comparing the inputs and the outputs. The hard part is understanding what things inside the models moved together to produce that output.

That essentially means you have to compare all pairs of signals. If you have 100 signals in a model—which, by the way, would be a very small model—you would have to compare all 100 to all other 100, which sounds easy, except that turns out to be more computations than there are atoms in the universe, which is a bad outcome. Well, it is a bad outcome, if you want an answer.

The tricky part is you have to figure out how do you optimize that in a way which guarantees correctness, but doesn’t require you to be computing until the sun burns out. And what the mathematicians on our team have figured out a way to do is to make those optimizations, but to do it in a way that they can still prove the answer and we can demonstrably answer the question of are we, in fact, accidentally discriminating against African Americans or women.

And that is our view, is that the two things that an explainability model must do: one, it has to successfully optimize across the space; and two, it has to be directly inquirable as to what do you do with respect to whatever classes are relevant.

Mr. GONZALEZ OF OHIO. Thank you.

And then one thing we have talked about a lot is the data itself. But we haven’t covered as much about—Dr. Buchanan, you mentioned it—privacy and who ultimately owns the data. I think that is an outstanding question for sure.

And so I guess my question is for Dr. Buchanan and anybody else who wants to take a stab at this, how should we think about balancing the innovation that we all agree can have a positive impact on society if we are good about it, with protecting consumers and empowering consumers with their individual data?

Ms. BUCHANAN. Thank you, Congressman.

I absolutely agree with this. And I have been very encouraged by what I have seen in the European Union regarding consumer protection on data and the right to own the data and what happens with your data.

I think one thing I would like to stress to you throughout today is, I keep hearing the term “big data”, but I think, moving forward, what we also need to distinguish when we are getting down to that granular level is that big data is not the same as strong, robust data.

Mr. GONZALEZ OF OHIO. Right.

Ms. BUCHANAN. When we are thinking about privacy, we need to think about using strong, robust data.

And I think I would also draw your attention to my written report where I look at China. Look at what they have been doing with their Sesame Credit model with Ant Financial, which is not the same as the government social credit scoring model, where basically every data point ever collected about you goes into a model to measure what is called “trustworthiness.” Not creditworthiness, trustworthiness.

And my thoughts on this is, at the end of the day, if I am going to look at getting a loan for a house, the data I really want to use and protect is my loan repayment history, not my subway fare usage, for example.

Mr. GONZALEZ OF OHIO. Right. Thank you.

Ms. BUCHANAN. And context is very important, too.

Mr. GONZALEZ OF OHIO. Yes, ma’am. Thank you. We will follow up.

And I yield back.

Chairman FOSTER. The gentlewoman from North Carolina, Ms. Adams, is recognized for 5 minutes.

Ms. ADAMS. Thank you, Mr. Chairman.

First of all, let me, before I begin my questions, I want to thank you for the opportunity to serve on this task force. And I am looking forward to it, along with you, and my friend, Congressman Hill.

To the witnesses today, thank you so much for your testimony.

As technology becomes more and more commonplace, it is critical that we proactively address issues that could positively and negatively impact our constituents and our financial institutions.

Algorithms have become a part of everyday life, even though most Americans have limited awareness or understanding of these systems and their impact. Increasingly, public and private enterprises have turned to artificial intelligence software and machine-learning programs to help increase the effectiveness of the services rendered.

Let me begin by addressing this question to Dr. Turner-Lee. There have been concerns about bias in AI systems, such as the potential of historical biases in datasets to be perpetuated or amplified in AI systems. How do firms ensure that AI systems are not having a disparate impact on vulnerable communities? And what safeguards should regulators and Congress put in place to protect consumers?

Ms. TURNER-LEE. Thank you, Congresswoman, and thank you for that question.



I am going to just give three points that I think need to be injected into this debate.

One is diversity in the workforce. The developers who sit at the table in the design of algorithms are not representative of the colorful spectrum of people who actually are using these algorithms. And, as a result, I think that we miss opportunities to have a seat at the table to mitigate issues related to gender or race or even background. I am a sociologist sitting among computer scientists. We need more perspectives with regards to that.

And I think to push for inclusion, we also need diversity in design. We wrote a paper at Brookings that is really about sitting at the table and thinking through what may become the intended and unintended consequences of these models. How are they replicating stereotypes that we see? In what ways should companies be trying to put in best practices that avert those types of discriminatory actions?

People of color, in particular, have not come this far to have technology become one of the major elements of further discrimination and amplified bias. And so, we have to be proactive in increasing the number of data scientists who are engaged in this, who come from diverse backgrounds, and also creating, I think, a standard, particularly in the sensitive use cases like financial services, employment, and housing, where people of color have already been historically disadvantaged, that we have to ensure that these sensitive use cases are not open for business with regards to doing further damage.

Ms. ADAMS. Great. Thank you.

Dr. Buchanan, within the context of financial services, have you seen the potential for bias in the use of AI? And how are various countries handling this issue? What should policymakers do to ensure the use of AI doesn't discriminate against vulnerable communities?

Ms. BUCHANAN. Some of the more notable examples that I highlight in my report, Congresswoman, relate to how algorithms are used in the peer-to-peer lending industry. And so, just to follow on from Dr. Turner-Lee's comments, I can refer you to a paper where I found that peer-to-peer listings where African Americans provide their pictures on the lending site are roughly 3 percent less likely to be funded and receive a loan and are more likely to pay higher basis points than white people with similar credit profiles. The examples I detailed in my reports are particularly pertinent in the debt consolidation.

Ms. ADAMS. Okay. Let me ask a yes-or-no question: Would it be useful for Congress to fund algorithmic bias research through NSF, NIST, and other Federal agencies, to develop tools, methods, and programs to resolve bias in artificial intelligence systems? If I can get a yes or no?

Ms. BUCHANAN. Absolutely, yes.

Ms. ADAMS. Okay. Dr. Turner-Lee?

Ms. TURNER-LEE. Yes.

Ms. ADAMS. Dr. Merrill?

Mr. MERRILL. Yes.

Ms. ADAMS. Mr. McWaters?

Mr. MCWATERS. Yes.

Ms. ADAMS. Okay, very good. Thank you very much.

Dr. Merrill—and I know we don't have a lot of time—what steps should companies and policymakers take to address this concern? Can you give me one?

Mr. MERRILL. I think the most important thing that regulators and policymakers should do is provide clarity. Even clarity that is not perfect is better than uncertainty to get companies to innovate in a good way.

Ms. ADAMS. Great.

Thank you, Mr. Chairman. I yield back.

Chairman FOSTER. Thank you.

The gentleman from North Carolina, Mr. Budd, is recognized for 5 minutes.

Mr. BUDD. Thank you, Chairman Foster. I want to commend you and my friend Ranking Member Hill for all your work on this task force.

I am excited that you all are here today.

And I want to start my time by highlighting the potential impact that machine learning and AI can have in our insurance market for institutions and their customers. But before I do so, I want to ask permission, Mr. Chairman, to enter into the record this report from the GAO. It is entitled, "Insurance Markets: Benefits and Challenges Presented by Innovative Uses of Technology."

Chairman FOSTER. Without objection, it is so ordered.

Mr. BUDD. Thank you, Mr. Chairman.

This report highlights how AI and machine learning benefit insurance markets and the consumer. I am excited to explore how this technology can improve underwriting accuracy, facilitate stronger communication with customers, make the claims processes easier to navigate for the consumer, and combat insurance fraud, among many other things.

Let me just highlight one specific provision from the GAO report, that is found on page 11. The report highlights telematics, which is the combination of telecommunications and information processing to send, receive, and store information related to specific items such as automobiles and water heaters. And I happen to have one of those water heaters, and it never knows when the in-laws are coming and when all the kids are home from college.

Telematics allows sensors in an automobile to provide data on a driver's behavior such as speed, hard braking, and turning radius. Now, according to the GAO report, insurers can then use that information to determine the driver's risk profile and help determine the premium rate for that driver, if a driver so chooses.

So, I encourage my colleagues to read this report that was requested by Ranking Member McHenry as we move forward with this task force with any potential policy proposals. Thank you.

I am sure we all agree that the U.S. must stay at the forefront of this new technology in the financial sector, like artificial intelligence and machine learning.

And here is the question. It is for Mr. McWaters: What challenges are companies facing that inhibit them from achieving the full potential of these emerging technologies? How are overly burdensome regulations stunting growth in this area? And how can

our committee ensure that proper controls are in place to protect customers while also fostering growth in AI?

Mr. MCWATERS. Thank you very much.

I think that one of the most significant instances of where we see challenges to responding to this on the part of particularly incumbent financial institutions are the legacy IT systems that are in place.

Typically, data is heavily siloed, making it difficult for that data to be ingested and used by conventional machine-learning methods, and the systems themselves, while extremely robust and resilient, are not as adaptable as modern and particularly cloud-based computing methodologies.

Interestingly, one of the things that we have seen in this space—and this pertains to some degree to Chairman Foster’s question about consolidation—is that there is an opportunity for third-party service providers to play a helpful role in enabling financial institutions to leapfrog forward, in terms of their capabilities.

By plugging into specialized fintech or regtech firms, into large tech firms which might offer, for example, machine vision as a service, you might as an insurance entity be able to use that machine vision to accelerate the processing of minor automotive claims, for example.

I think that, in terms of the discussions that I have internationally, one of the perceptions of the United States in this space is that the regulatory environment is extremely complex to navigate and that the large number of regulatory entities creates challenges to deploying new innovations effectively.

I don’t have a specific remedy for that, but it certainly is one of the contributors to the challenge of deploying these technologies here in the United States.

Mr. BUDD. I appreciate that, Mr. McWaters. And continuing on with you, besides lower cost of financial products and services, what are some other ways in which a consumer stands to benefit from adoption of these technologies in the financial services?

Mr. MCWATERS. I think one of the particular items here is the opportunity to provide valuable advice and intervention for clients. So, if you pursue the example of insurers that you gave, telematics has an opportunity to, on one hand, support more accurate and more personalized underwriting, but it also increasingly has the potential to give drivers valuable feedback on how they might be safer drivers.

The water heater that you mentioned might be able to alert you if there was a leak, allowing you to minimize the damage to your home in a way that is beneficial both to you and to the insurer who has provided that cover.

Mr. BUDD. It sounds like a lot of opportunities.

With that, I yield back. Thank you.

Chairman FOSTER. Thank you.

And after consultation with the ranking member, I would like to inform Members that we are going to have time for a second round of questions, subject to the fact that we have to be done here by 11:30. So, we should at least have a partial second round here.

I now recognize the gentlewoman from Texas, Ms. Garcia, for 5 minutes.

Ms. GARCIA OF TEXAS. Thank you, Mr. Chairman, and thank you for having this hearing. And I thank Chairwoman Waters for really focusing on this issue, because it is so important as we move forward.

However, I think it is one that is kind of confused, and I wanted to just start with a question. I was trying to figure out which professor to ask, so I am going to go ahead and go with a woman. I, too, have some biases.

Dr. Buchanan, for those who are watching who are not in the financial industry, who don't know what artificial intelligence means, they hear the word, "intelligence", and they think it is some really super big-brother secret stuff. Can you in just plain English, in 25 words or less, tell the average viewer what the heck we are talking about?

Ms. BUCHANAN. First of all, there is no generally agreed upon definition of "artificial intelligence."

Ms. GARCIA OF TEXAS. You are using up your 25 words now. You are talking straight to the average consumer in the United States.

Ms. BUCHANAN. Okay. I would say it is a group of technologies and processes that can look at determining general pattern recognition, universal approximation of relationships, and trying to detect patterns from noisy data or sensory perception.

Ms. GARCIA OF TEXAS. I think that probably confused them more.

Ms. BUCHANAN. Sorry.

Ms. GARCIA OF TEXAS. With all due respect, but I think that is one of the challenges that we have. I wanted to do that, not to make light, but just to accentuate the problem that we are facing, because I think there is an idea that now all these robots are going to take over all the jobs and everybody is going to get into our information, this whole balance that one of my colleagues mentioned between privacy and the markets. So, I think it is important.

Ms. Turner-Lee, one of the things that would help us better understand it, I think, are some of the things you pointed out, in terms of diversity of the people at the table who are developing the software, the people who are the workforce involved.

If you could name the single one thing that Congress could do, I mean, we can't change attitudes. We probably can't change some of the criteria that the folks who are putting this together are looking at. What would you suggest that one thing be?

Ms. TURNER-LEE. Yes. That is such an interesting question, because I think the tech diversity issue has been one that Congress, as well as civil society actors and others, have really grappled with. And as we see technology evolve in the way that it is to a point where it is confusing, I would suggest that we have a lot more to do as these become much more ubiquitous and widespread.

On your question, I think what Congress can do first to quell algorithmic bias is to create guardrails. I think it has been mentioned that we need to ensure the tech companies know that they have to be in compliance with antidiscrimination laws. I think we start there. We create guardrails for best practices in design and development.

With regards to creating more diversity at the table, these are companies that are not necessarily regulated or in any way required to report diversity, in terms of who they serve and who is

sitting there. But I think we should reward best practices where we are seeing demonstrations of companies wanting to bring more actors to the table.

What does that mean? Years ago, when we had the ENERGY STAR standard imposed on appliances, most of us who go into a big box store know this appliance is going to save us money and it is going to be safe.

I think we should push in the algorithmic economy a gold standard: What is the Energy Star rating for what consumers understand of how their data is being used? And how will companies pushing the bar, raising the expectation that they are going to be in compliance, not only with those nondiscrimination laws, but they are going to be good stewards of our information and they are going to have environments where diversity is encouraged?

Ms. GARCIA OF TEXAS. Is there anything that we can do in terms of the criteria that they are using? Because I know one of the examples you gave on gender bias was just the word “woman” being on their resume somewhere caused to trigger the gender bias.

What can we do with regard to the criteria being used? For example, if you looked at my resume, I graduated from a Historically Black College, and I would hope that there is no assumption that I am African American, but a computer could do that, right?

Ms. TURNER-LEE. That is right.

Ms. GARCIA OF TEXAS. But I also go to a women’s college, so, obviously, that is going to peg me in that. But then they look at me, and I don’t look like I am Latina.

Ms. TURNER-LEE. That is right.

Ms. GARCIA OF TEXAS. I am going to have one confused computer.

Ms. TURNER-LEE. That is right. And you are going to have a double or triple jeopardy, right?

Ms. GARCIA OF TEXAS. But is there any way that we can do anything about what gets in the computer?

Ms. TURNER-LEE. Yes, as a policymaker myself at Brookings, it is so challenging to figure out how do we get companies to sort of adhere to a standard without overregulating them? And that is why I think those guardrails are particularly important.

But I also think it is important for us to continue this discussion on what does disparate impact mean when collective groups of people are denied loans or denied credit or denied some form of equitable opportunity in this country simply because the computer was wrong. Who is liable for that? Is it the developer?

I actually agree with what was said earlier. I don’t think developers necessarily walk around in a hoodie saying, “Today, I am going to discriminate against people.” I think it is the nature of what is in the black box that is not understood, which is why explainability models matter.

People need to understand what is going into this ocean. And for the layperson, I will give you this example that I use. It is like swimming in the ocean. At the top, you can see my legs and my hands, but when you go down, you begin to not see my body because the water becomes really cloudy.

I am okay if I actually search for camping gear for my son on one site and it shows up on another site. I am not okay if I am profiled because I am an African-American woman or a woman who

went to a Historically Black College, et cetera. Those are things that I can't see how you even got there to understand that from just my hand sticking out.

And so, we have to figure out what are those guardrails that will protect people, where are there pressure points to institute some other consumer protection, what is the role of privacy in terms of the data that is collected on people?

And I would suggest to you, where in the process can I recurate my identity and let them know that, "Hey, I am not this person that you keep thinking I am just because I buy camping gear. It is not me going out; it is my son."

Ms. GARCIA OF TEXAS. It is a good point. Thank you.

Chairman FOSTER. Thank you.

Ms. GARCIA OF TEXAS. I yield back. Thank you, Mr. Chairman.

Chairman FOSTER. This is a wonderful discussion that could go on forever.

The gentleman from Virginia, Mr. Riggelman, is recognized for 5 minutes.

Mr. RIGGELMAN. Thank you, Mr. Chairman, and thank you to Ranking Member Hill, and thank you to all of the witnesses for being here.

I would like to start by saying I am proud to be a member of the inaugural Artificial Intelligence Task Force. And I was going to send my avatar today, but it kept going in circles and bumping into walls, so I said, I am going to come here myself. That was a bad, bad joke.

But, anyway, my background experience with data analytics has taught me a lot, especially about the evolution I personally witnessed since 2002. And to get to my questions, I just want to talk really quickly about what I have done. My experience might be a little bit different than everybody up here.

I have been trying to aggregate big data and analyze big data for predictive analysis to go after actually network centers of gravity and critical touchpoints for a long time in the nonkinetic space on the military side.

And back in 2002, I want to tell you guys, that the big thing about the military—we have this incredible saying, that we try to solve today's problems with yesterday's technology tomorrow.

I think what I saw in 2002, there was never a statement of AI or machine learning. We were using these just really kludgy relational databases, trying to build arbitrary translators to try to make sure the nodes and attributes actually made sense for unproductized data, productized data, but mostly data that just didn't make a lot of sense to us in 2002.

What we have seen in the last 5 years, and I know this is crazy because sometimes the DOD is a little bit behind, but it is our work with places like Johns Hopkins University's Federally Funded Research and Development Centers (FFRDCs), working with the physics labs. And now you see a lot of not only private-public partnerships, but you see a lot of commercial and government partnerships in big data.

And what we have seen going forward is, that 5 years ago we might have been using relational databases, but now we are using graph databases and dynamic translators we could have never fore-

seen in the future. We had about 40 people working with us trying to find every touchpoint and every critical node in a network. So, I went from dropping bombs to actually dropping nonkinetic bombs, right, in specific types of networks, is pretty much what we did.

And it is just amazing to me, listening to all of you, that my background is so different, just based on trying to work with data, and the fact that machine learning and artificial intelligence, even up until 2010, 2011, in the military space, and big data with my companies, we really didn't talk about it much. We just really didn't. But now we can.

And what we see now is that now we are getting unproductized data. We are getting disparate data, multiple datasets. I am getting natural language processing. We are getting tons of unstructured data. We are able to go into dynamic translators we can put into graph databases, and now we are actually coding to what people are thinking when they are looking at a specific problem set. We are coding to an analyst's brain serially in parallel. Now, we have machine-learning templates.

And here is what happened after all that incredible stuff: It failed miserably the first time, because we were missing so much data.

The thing that I am going to ask, because I have my own reasons about this, and I will ask Mr. McWaters first, when you look at AI and ML, when you are looking at ML templates, machine-learning templates, when you are looking at what artificial intelligence is, the difference between templating and the difference between rules, where do you think the split is? And I want to ask some of you, where do you think the split is because definitions of machine learning and AI?

I know I have my own, but I would love to hear from you, because sometimes I even get sort of wrapped around the axle in trying to figure out where that split is and where we can actually look at some of the safeguards to make sure that we make the right jump from ML to AI.

Mr. MCWATERS. There is an old joke that artificial intelligence is whatever a computer can't do yet.

Popularly, our definitions of this have tended to move over time. Twenty years ago, you might have said that a computer would be intelligent if it could beat a grandmaster at chess. Today, we sort of think of that as being a relatively trivial case of intelligence. We think of it as being programmatic.

So, I think our definition of artificial intelligence tends to move over time. And, as Dr. Buchanan said, I don't think there is a clear articulation of exactly which techniques—ML, deep learning, and others—are specifically rested under the umbrella of that definition.

Mr. RIGGLEMAN. Dr. Merrill?

Mr. MERRILL. I think we can spend a lot of time trying to get our heads around the different definitions. When I started in the field, which is a long time ago now, AI was generally thought to be machines that tried to actually reason, that tried to start with an initial point and take steps to get to an end point, whereas ML was viewed more as just rote math, just like throw a computation at the problem.

Mr. RIGGLEMAN. Right.

Mr. MERRILL. You can still sort of throw that distinction out, but it just turns out to be a little bit unhelpful at the end, because AI failed when I started and it is roughly still failing, because it is just a really hard problem. People turn out to be really, really complicated beings.

And stuff which we said could never get done until AI worked is now relatively trivial in ML. To wit, your car's brakes are better than you are. And that is a case of ML that we said could never be done. You could never compute friction, but it turns out you can.

Ultimately, I think the most important class is maybe not whether it is AI or ML, but rather what are the characteristics of the problem you are trying to solve? AI-based techniques are trivial to explain. ML techniques are quite a bit harder to explain, but quite a bit more powerful. And so I guess I would encourage us to think less about the technique and more about the category of problem.

Mr. RIGGLEMAN. Thank you.

And that is why I am so excited about this. Thank you, Mr. Chairman. Because I think we have a chance to really solve some problems here, and I am happy to be here. Thank you, sir.

Chairman FOSTER. Thank you.

The gentleman from Georgia, Mr. Loudermilk, is recognized for 5 minutes.

Mr. LOUDERMILK. Thank you, Mr. Chairman.

I appreciate the panel being here. It is a very intriguing discussion we are having here today, especially as I spent 30 years in the information technology industry, as my good colleague, Mr. Riggleman, also spent time in the intelligence community in the Air Force in the earlier days where we were using analytics of massive amounts of data. And what is happening in that arena today is light years beyond anything that we were able to do with rooms full of main processing systems, mainframes back in the time.

And I am really interested in this field today, in what we can do with our artificial intelligence. I think it is also as important to understand our limitations of what we can't do and draw our boundaries around that, but yet on the periphery of that boundary having the sandboxes to where we can test and we can implement what we may be able to do in the future once we stabilize that.

One of the things I am interested in is what can we do today with artificial intelligence and fraud detection and prevention, because that is something that is really important in the industry, especially as we move more in the fintech arena.

My line goes back to the chip card industry. Since I have been in Congress, when I first started here, my debit card and my credit card had a chip, but I could only use it when I traveled overseas.

Once we implemented that ability here, the fraud went down by 76 percent. But criminals being criminals, all they do is shift their focus, and that focus has gone over into the digital payments arena, which is where we have a lot of challenges today.

And, Dr. Buchanan, I appreciate your discussion that you brought up in your testimony about how one of the payment card networks is using AI to help financial institutions reduce their fraud by \$25 billion annually. Can you tell us more detail about



how payment processors— financial institutions, insurance, retail, and others are using AI to combat the digital payment fraud?

Ms. BUCHANAN. When we are thinking about AI's automating simple and complex decisions—actually, that is my 10-word definition, so I think I have redeemed myself, Congressman.

One area that I can address to you is that 50 percent of phishing detections are now finance-related. And so what I detail in my report are some very encouraging examples around the world where financial services companies have tried to reduce phishing attacks.

There is a really good example in my report, IBK, a phishing voice detection app, and it is really a coordinated effort between regulators in South Korea and the financial services industry.

Basically what this app looks at is—and phishing in South Korea accounts for millions of dollars a year—a phone call is made, and it looks at picking particular keywords in the phone call. And if it meets a particular threshold, then an alert signal is sent that this is a potential voice phishing scam, and a significant financial transaction is halted.

In Estonia, Monese is using artificial intelligence in this arena as well, particularly when they are trying to on-board customers in the first place. So, they are looking at matching documents with video selfies in order to detect fraudulent IDs and fight identity theft.

Mr. LOUDERMILK. I traveled to Estonia last year, and what they are doing in the fintech industry is really a model for a lot of other nations. It is surprising, especially being an Eastern Bloc country, the suppression that they had during communism, to be able to come out to where they are now.

Regarding the things you just explained to us, payments.com showed that less than half of financial institutions use AI for fraud prevention. Why are we not seeing more use in the industry for fraud prevention?

Ms. BUCHANAN. That is an interesting question, Congressman. I think really it is because detecting fraud in the first place, we think about fraud as really being a latent variable. I mean, it is not necessarily directly observable, and so it is more challenging to machine-learn algorithms.

Actually, in some sense, you have a little bit of a self-defeating goal here. You could have the case of falsely declining transactions as fraudulent, okay. That actually costs the industry a lot in lost customer loyalty each year.

And apart from this erosion of customer loyalty and loss of retail losses, the machine-learning algorithms to detect fraud, as I said, they are more latent, in the sense that it is easier to track someone's shopping history directly. You see what they purchase. You see what they buy. But fraud is just another layer. It is not as directly observable. And I think that presents a complexity to the process.

Mr. LOUDERMILK. Thank you.

Chairman FOSTER. Given the time constraints on our occupancy of this hearing room, it looks like we will have time for only 5 minutes of questioning by the ranking member and the Chair. So, I would now like to recognize the distinguished ranking member for 5 additional minutes of questions.

Mr. HILL. I thank the chairman.

I thank, again, the panel for being here today. I appreciate your contributions to this important beginning of the task force work for this Congress.

Mr. McWaters, I wanted to start with you and just talk about some of the ways today that you are seeing AI being used in the financial services industry.

So, if you would talk about two or three of the biggest ways you are seeing artificial intelligence being used by the financial industry in customer acquisition, extension of credit, regulatory compliance costs? Name two or three or four specific elements in each of the main areas, if you would.

Mr. MCWATERS. I think we are seeing four key ways in which this is being deployed in financial services.

The first is driving increased efficiency, being able to do the same thing faster and with less manual input. And that can be a benefit both to the organization, obviously, in their bottom line, but also to the consumer, who is able to get an answer to their question or to their request more quickly.

Second, we are seeing an improvement in outcomes. Dr. Merrill made reference to this in terms of being able to originate more loans, accept more applications without a significant increase in defaults.

Third, we are seeing entities build out entirely new businesses. By virtue of some data flow that exists, is propagating through already, you may be able to create new value propositions. So, a payment network might be able to create a business of macroeconomic forecasting based on the data that flows through their network and monetize that separately.

And then finally, advice. Americans struggle to access the financial advice that they need to make good financial choices in the moment to plan for retirement. That advice traditionally has needed to be delivered by expert individuals and can be very expensive.

We are at the very beginning, I believe, of the opportunity to provide high-quality advice to individuals in real time that will help to address that issue. It is nascent today, but the opportunity is quite significant.

Mr. HILL. On that point, I believe in making sure that we have an economy that offers choices to consumers from the whole spectrum of the most machine-led robo-adviser to the most sophisticated one-on-one consultation. I don't think that government policy should bias towards that, and we have had some debates over the last 4 years where I think government policy actually directed people away from advice to machine-driven robo-advisers.

If I go through a sharp downturn in my portfolio and it has been dependent on a robo-adviser, who am I holding responsible for that? Who can I go talk to about that?

Mr. MCWATERS. I think that is an open question.

Mr. HILL. I don't like open questions. That is why we are here today. We need to make sure that those consumers know the risks of that. And that may be the trend of the moment or the trend of the time or it may be, in the short run, more affordable, but those are the kinds of things I think we have to talk about here in this, in our work.

Mr. MCWATERS. I would also note that I think that you will see in this space that even amongst some of the sort of highest echelons of private banking, what we now see is an appetite by those consumers to have a mix of both automated and in-person mediated items.

The other thing that I would note in response to your earlier question about consolidation in the marketplace is that these technologies can also provide an interesting opportunity for small and mid-sized financial institutions to rapidly catch up to large entities.

Mr. HILL. I do share your optimism there. All through the technology cycle, going back from a mainframe to a business size computer to the cloud, small broker-dealer competitors and small financial services competitors have had access to scaled-up technology through a vendor platform that in some ways helps them do a better job of being in full compliance of risk.

Data privacy, if each of you would just quickly answer, do you support the use of APIs when it comes to protecting customer service, customer data interfaces between aggregators or individual companies?

Dr. Turner-Lee, do you want to start?

Ms. TURNER-LEE. Yes, I do.

Mr. HILL. Dr. Buchanan?

Ms. BUCHANAN. Yes, I do.

Mr. HILL. Dr. Merrill?

Mr. MERRILL. Yes, I do.

Mr. HILL. Mr. McWaters?

Mr. MCWATERS. Yes, I do.

Mr. HILL. Good. Thank you. I yield back.

Chairman FOSTER. Thank you. And I guess as a follow-up on the API question, what do you think the state of the art is for authenticating yourself for access to those APIs?

Because one of the scariest things that I see about artificial intelligence is just the very impressive high-quality tools being used for phishing. Things, for example, where they will listen to your voicemail response, use that to synthesize your voice, and fake a phone call to one of your friends in your contact list saying, "Hey, Joe, I just sent you an email with an attachment, can you have a look at the attachment and call me back?" And everyone clicks on that attachment. And that is not even mentioning the video that is now available.

I think one very valuable thing the government can do is to at least provide citizens who are interested in having a high-quality way of digitally authenticating themselves online very much in the way Estonia has been leading the way.

And my closing question, I guess to each of you is, we have about 1 minute for each, if you look forward at the competitive environment, you see all of the giant banks trying to—they all have 10-year plans to turn themselves into tech firms. All of the tech firms are getting into banking as rapidly as you can imagine.

And so looking forward a decade, what do you think about the competitive landscape? Will there be any difference between giant financial institutions and tech firms, as we know them now?

Just march down the line.

Dr. Turner?

Ms. TURNER-LEE. I think we are going to go in this era of converged services, and it is going to be very challenging for regulators and Congress to discern what guardrails apply to whom. And right now, we have strong sectoral policies that affect the financial services sector, and we have loosely regulated policies that may apply to tech companies.

I think going forward we are going to have to figure out, particularly on behalf of consumers, where do those protections lie and where do we again place pressure for regulatory frameworks that allow for innovation while at the same time putting some stresses around the fact that we cannot have permissionless forgiveness in areas that have huge consequence for consumers.

And so, I completely agree with you. I think at some point, the lines are going to be so blurred we are not even going to know.

But keep in mind it has been consumers who are driving that demand for these services. So, I agree with you as well, we have to do—

Chairman FOSTER. And in Congress it is, obviously, a big issue, because I think there are seven committees that claim they are doing some part of IT, information technology, which means, of course, no one is doing it.

So, Dr. Buchanan, any thoughts on this?

Ms. BUCHANAN. The landscape I see moving forward, Chairman Foster, is more mergers and partnerships between banks, financial institutions, and big tech companies.

I do agree with Dr. Turner-Lee about drawing this line about how data is used. And I am very concerned, moving forward, that I want to make sure we don't give up privacy at the expense of convenience.

Chairman FOSTER. Thank you.

Dr. Merrill? And also, if you could comment on the role of the startup in this, where they may or may not have access to these giant datasets that seem to be essential for success in AI?

Mr. MERRILL. I guess I will be a little bit of an outlier here amongst my distinguished colleagues.

I think there is essentially no chance that in a decade we will see mergers and material consolidation between technology companies and big banks, because the cultural differences will be so great that the mergers will blow up.

I was responsible for a variety of our financial products when I was still at Google, all of which were carefully regulated really, because we were a bit weird about that. And it was clear that that was the wrong place to do those, those products, not because anyone had the wrong intent, but just because it just didn't fit.

I think ultimately, startups are at material risk, and I think that is very dangerous for the U.S. economy. We are at risk because it is hard to get data. We are at risk because a brief sideswipe by a large company, let alone the government, will crush any of us.

And I think over the last 20 years, for good or for ill, we have seen a lot of the development in this economy coming from startups. So, my biggest worry is that.

Chairman FOSTER. Mr. McWaters?

Mr. MCWATERS. I would argue that we need to think outside the bank, if you will, that we think about financial services in a heavily

verticalized and siloed fashion. We need to think about it in a more modular way.

And so when I look forward to the 10-year landscape, I would predict a world in which customer experiences for financial services increasingly trend towards the best of what big tech can offer, whether that is offered by a traditional financial entity or a technology entity, but that the products that the consumer accesses, the loans, the insurance, they need to fundamentally remain regulated.

And the data that is used to inform the entire experience needs to become more secure, the customer needs to have more control, and we need to really enfranchise the customer within a regulated framework.

Chairman FOSTER. Thank you.

And I would like to thank all of our witnesses for their testimony today.

The Chair notes that some Members may have additional questions for this panel, which they may wish to submit in writing. Without objection, the hearing record will remain open for 5 legislative days for Members to submit written questions to these witnesses and to place their responses in the record. Also, without objection, Members will have 5 legislative days to submit extraneous materials to the Chair for inclusion in the record.

This hearing is hereby adjourned.

[Whereupon, at 11:36 a.m., the hearing was adjourned.]



# **A P P E N D I X**

June 26, 2019

**Testimony of Dr. Bonnie Buchanan**

Head of Department of Finance and Accounting,  
Full Professor of Finance, Surrey Business School.  
The University of Surrey  
Guildford, Surrey, GU2 7XH UK

before the

The Task Force on Artificial Intelligence of the House Financial Services Committee

**Hearing Entitled “Ending Perspectives on Artificial Intelligence: Where We Are and the Next  
Frontier in Financial Services”**

Wednesday June 26, 2019  
Rayburn House Office Building, Room 2128



**Biography**

Professor Bonnie Buchanan is the Head of the Department of Finance and Accounting at the Surrey Business School, University of Surrey, UK. In 2018-2019, Professor Buchanan served as the Fulbright Distinguished Chair in Business and Economics at the Hanken School of Economics, Finland (her project was titled Fintech in the Nordic Countries). She has also served as the Bosanko Professor of International Economics and Finance at Seattle University as well as the George Albers Professor. She has taught university courses on Financial Institutions and Markets, International Finance and the History of Financial Crises. Dr Buchanan is the Editor in Chief of the Journal of Risk Finance.

Professor Buchanan conducts research in Fintech, Artificial Intelligence in Finance, Securitization and International Finance. She is the author of the book, *Securitization and the Global Economy*.

Professor Buchanan holds a Ph.D. in finance from the Terry College of Business, University of Georgia, a M.AppSc. in statistics from RMIT and a BSc. (Hons) in mathematics from the University of New South Wales.

Distinguished members of the Subcommittee, thank you for the opportunity to appear before you and provide testimony today to help inform discussion about artificial intelligence in the financial services sector. I am Dr Bonnie Buchanan, Professor of Finance at the Surrey Business School, University of Surrey. In this testimony I will provide some background on AI, its applications in finance as well as challenges and opportunities facing the financial services industry. I hope we can all work together to address the challenges and opportunities that artificial intelligence provides to the financial services industry.

#### **Overview**

Artificial Intelligence (AI) is rapidly transforming the global economy and the way we think about our financial future. Global revenues in the AI market grew from \$126 billion in 2015 to \$482 billion last year and are forecast to increase to \$3.061 trillion by 2024<sup>1</sup>. In 2017, 84.2% of cards and payments in the banking sector used AI techniques, mainly in online payment and credit card usage<sup>2</sup>. In 2017 AI was ranked as the key trend in financial services and Fintech<sup>3</sup>. AI, cloud computing and big data have created an affordable infrastructure to spur innovation in the financial services sector. There are two explanations for AI's impressive growth in a relatively short period of time. First, exponential advances in computing power have contributed to declining processing and data storage costs. Second, data availability is now more widespread.

Outside of the IT sector the financial services industry is experiencing the fastest growth and is in turn the biggest spender on AI services<sup>4</sup>. Until recently hedge funds and high frequency trading (HFT)<sup>5</sup> firms were the main users of AI in finance, but usage has now spread to other areas including banking, insurance, wealth management, personal financial planning and regulation. More specifically, AI financial applications include: algorithmic trading, portfolio composition and optimization, model validation, back testing, robo-advising, virtual customer assistants, market impact analysis, regulatory compliance and stress testing models.

AI is both disrupting and refining existing financial services. AI is not straightforward to define as there is no generally agreed upon definition. However, AI can be broadly thought of as a group of related technologies including machine learning and deep learning. Machine learning (ML) is concerned with general pattern recognition and universal approximations of relationships in data in cases where no a priori analytical solution exists. ML is best suited for situations that require extracting patterns from noisy data or sensory perception – or what is termed, a “data-up approach”. ML is primarily derived from sources such as experience,

practice, training and reasoning. A typical ML application is based on a problem, a data source, a model, an optimization algorithm and validation and testing.

ML uses algorithms to automatically optimize through experience with limited or no human intervention (or in other words, supervised versus unsupervised learning). An example of supervised ML in a banking context entails teaching an algorithm to learn from past regulatory breaches and to predict new breaches such as insider trading or cartel detection. In unsupervised learning, ML can help issue alerts such as low balance warnings. It can also be applied to bank overdraft charges to help assess what is happening to individual customers and what might be the causes of their current situation. Clustering algorithms help accomplish this objective. Regulators can use clustering algorithms to better understand trades and categorize business models of banks in advance of regulatory examinations. Topic models help us understand the behavioral drivers of different market participants and includes text mining and natural language processing (NLP). NLP links human language with computing. For example, the SEC has used topic models to detect accounting fraud.

The term “robo-advisor” was virtually unheard-of five years ago but is now commonplace in the financial services jargon. Chatbots and robo-advisors powered by NLP and ML algorithms have become powerful tools which provide a personalized and conversational experience to users in the financial services sector. For example, in September 2017 Allstate Insurance deployed Amelia, an AI powered chat bot, to assist employees. Amelia is trained in 40 insurance related topics and uses deep learning and NLP as well as data analytics to understand the intent of the user’s text and offer precise answers. To date, Amelia has helped call center representatives with more than 3 million customer conversations<sup>6</sup>. In another example, Lemonade is a platform providing property and casualty insurance to home owners and renters. Lemonade uses ML and chatbots for its customers. On average, it takes 90 seconds to get insured and 3 minutes to get paid for a claim.

Deep learning (DL) algorithms automate routine tasks, mitigate risk, help prevent fraud and assist in generating new insights. DL uses neural networks (NN) which are based on mimicking the way multiple layers of the brains’ neurons work (hence the term ‘deep’). For example, the Deutsche Bundesbank’s risk management area is already using NN to assess financial market soundness. NN have also been widely used in predicting financial distress and bankruptcy likelihood. NN, clustering and decision trees are AI techniques that assist financial institutions study customers’ buying behaviour, comparing it against other indicators to create a more complete picture of a transaction. Two primary advantages of DL are: (1) it is more resilient than machine learning to overfitting and (2) DL can address non-

linear events such as market volatility, which in standard quantitative models must usually be adjusted manually. However, one of the challenges with DL is model opacity, or the “black box” nature of its predictions. It is called “black box” because of the user’s limited ability to fully understand how the DL processes derive their predictions.

### **AI and Accountability**

AI is continuing to become more sophisticated and complex. But as we saw with the last financial crisis, financial markets are already very complex. This rapid growth and complexity in both AI and the financial system presents major new challenges regarding regulation and policy making, risk management as well as ethical, economic and social hurdles.

Like other Fintech product areas, AI should enhance financial inclusion. There are approximately 1.7 billion adults (or about 31% of adults) who are “unbanked”<sup>7</sup>. In these countries, cash economies are being supplemented by mobile access to digital funds. The increased application of AI technology to financial markets is likely to reduce barriers to entry for many individuals and business models that might not have previously had access to financial markets.

However, ML algorithms can potentially introduce bias and discrimination. Deep learning techniques provide predictions, but they do not provide insight into how the variables are being used to reach these predictions. This is especially important for trying to prevent discrimination in lending models. Hiring and credit scoring algorithms can exacerbate inequities due to biased data. Applications such as facial recognition can be inaccurate and biased. This can be demonstrated in the P2P lending industry. P2P business platform models depend on proprietary and complex algorithms. The interest rates applied are often based on credit e-scores (and sometimes other optional information provided). In the US, P2P platforms have come to represent an important market for debt consolidation. There is already a literature on P2P lending that investigates possible bias and discrimination in the industry. Duarte et al (2012) find that borrowers who appear more trustworthy (they have provided a photo on the platform) have a higher probability of getting funded. Online friendships of P2P borrowers can act as a signal of credit quality (Lin et al., 2013). They find that friendships increase the probability of successful funding and lower interest rates on funded loans. Unverifiable information affects lending decisions above and beyond the influence of verifiable and objective information with P2P loans (Herzenstein et al, 2011). Finally, Pope and Snyder (2011) find evidence of discrimination based on race, age and

weight. They find the market favors those listings that signal military involvement, being female or a desire to pay down credit card debt.

To combat potential bias in the mortgage lending market, Zest Finance applies AI models and big data. Through its ZAML Fair tool, Zest Finance reduces the impact of discriminatory credit data by excluding signals that tend to result in bias.

AI is also being applied to debt collection agencies. Consider the Chinese P2P lending market which has experienced platform failures in the last few years. After mid-2017 many P2P lenders shut down due to new lending controls and additional required licenses. Ziyitong launched an AI platform to help recover an estimated Rmb150 billion in delinquent loans<sup>8</sup>. The AI platform helps recover delinquent loans for approximately 600 debt collection agencies and over 200 lenders (including the Postal Savings Bank of China and Alibaba). A dialogue robot utilizes information about borrowers and their friends' network, and then uses the information to determine the phrasing with the highest likelihood of compelling the borrower to repay the loan. The dialogue robot will also call the borrower's friends, encouraging repayment of the loan. Ziyitong claims its recovery rate is 41 percent for large clients and loans that are delinquent up to one week, a rate that is twice that of traditional debt collection methods.

As financial services become increasingly automated, it remains unclear as to whether all borrowers will benefit from AI. If poor inputs are provided, then the biased outputs will be produced by the algorithms. In other words, bias in, bias out. This will have huge repercussions for low-income and minority consumers. Existing inequality could be exacerbated by ML algorithms that single out borrowers who are already disadvantaged as poor credit risks. In this scenario, borrowers might seek out alternative financial providers such as payday lenders and end up paying much higher interest rates than a traditional lender. Cathy O'Neil (2017) characterizes "weapons of math destruction" as being important, secret and destructive. This could be said of biased and discriminatory algorithms in financial services: they affect large numbers of people, are entirely opaque and destroy lives.

Resolving issues such as discrimination and bias requires being grounded in ethics and understanding what causes the bias in the algorithm in the first place. When it comes to AI in the financial services industry and a fairer future, policymakers need to be concerned about explainability and accountability of AI models. To overcome discriminatory bias, there needs to be robust oversight to ensure that AI applications in the financial services industry remains accountable to all members of society. An April 2018 UK House of Lords report<sup>9</sup> suggests that the AI sector's full potential would only be realized if potential risks such as

algorithmic bias and the opaqueness of "black box" systems that we see in DL techniques can be mollified.

#### **Fraud and Cybersecurity Issues**

The use of AI in the financial sector can assist in identifying fraud and cybersecurity crimes. A 2019 World Economic Forum report ranks the inappropriate use of customer data as one of the top two risks facing the global financial system. In banking, IT governance, fraud and cybersecurity are now equally important as capital and liquidity requirements. Online financial crimes have become more sophisticated and cost countries significant economic losses each year. Cybercrime costs the global economy over US\$400 billion annually with credit card fraud accounting for a large portion of this cost.<sup>10</sup> Due to massive fines imposed upon banks for failing to stop illegal financing, many banks have turned to AI techniques to improve their operations. For example, Feedzai uses real time ML to identify fraudulent transactions by recognizing behavioral patterns that could indicate fraudulent payment activity.

Another example of financial fraud is the voice phishing scam that recommends low interest loans to financially strained citizens. For example, last year in South Korea voice phishing scams entailed losses equivalent to \$391 million. As a regulatory response, the National Information Society Agency, the Financial Supervisory Service and the Industrial Bank of Korea co-developed "IBK Phishing Stop"<sup>11</sup> which is an AI-based voice phishing detection app. The algorithm is trained to detect stipulated keywords, special phrases and speech patterns. If at least 80 percent of the phone call is perceived to be fraudulent then an alert is sent before any significant financial transactions are made.

Another example of successful fraud detection involves Estonian company Transferwise, which moves nearly \$4 billion across borders every month. Multiple ML models instantly score each transfer, and Transferwise also uses ML to detect fraudulent behaviour and money laundering attempts. Another Estonian company, Veriff, uses ML applications for automation and fraud detection. Veriff achieves this in a cost-efficient manner by scanning 3200 different document types that are issued around the world. Human intervention only occurs when a more thorough examination of the transaction is needed.

Other challenges exist in this area, such as transactions wrongly declined due to suspected fraud, known as the "*false positive*". This works against the issuer because a false-positive declined transaction can result in erosion of customer loyalty and retail losses. False positives account for \$118 billion in retail losses and nearly 39 percent of declined

cardholders report that they abandoned their card after being falsely declined (Buchanan, 2019). ML methods can substantially reduce false declines and improve credit card approvals<sup>12</sup>.

#### **AI and the future of work in the financial sector**

In the financial services industry, AI has the potential to disrupt jobs across many levels. In 2017 Opimas LLC estimated that AI would result in approximately 230,000 job cuts in financial firms worldwide by 2025, with the hardest hit area being asset management (with an estimated 90,000 job cuts)<sup>13</sup>. In 2016, the GIS-Liquid Strategies group was managing \$13 billion with only 12 people. A major issue confronting the financial industry is how to balance the rapid deployment of AI, ML and DL against developing the best talent pool and skillsets. If algorithms struggle to distinguish the signal from the noise, then one really needs a person to step in and recalibrate ML models. Human talent and skills will become even more critical to sustaining competitive advantage in the financial services industry.

Robotic Process Automation (RPA) is being widely utilized in the banking industry. RPA enhances productivity, reduces transaction costs, eliminates manual errors and redeploys staff to higher skilled roles. One example of RPA is used by the UK Serious Fraud Office (SFO). In a typical year the SFO processes over 100 million documents in fraud and corruption cases. One notable case is the Rolls Royce bribery case, which resulted in the largest ever fine imposed in the UK for criminal conduct<sup>14</sup>. The SFO used the RAVN robotic system, which ended up costing £50K, and saved UK taxpayers hundreds of thousands of pounds. RAVN is referred to as a Legal Professional Privilege (LPP) robot and sifts documents into “privileged” versus “non-privileged” piles, indexes and compiles summaries. In the Rolls Royce case, RAVN processed 30 million documents at a rate of up to 600,000 per day (compared with a team of lawyers that would have processed 3,000 per day). Law clerks were deployed to other areas of the case.

As AI becomes more pervasive in the financial services industry there will need to be a shift towards appropriately educating workers. Graduates with tech and finance skills are in high demand. But as we move forward, and AI models become more ubiquitous in the finance, students will need to integrate other skills such as philosophy, economics, psychology, anthropology and sociology.

**Risks and Regulatory aspects**

AI is viewed in the financial services sector as a technique that has the potential to deliver huge analytical power, but many risks still need to be addressed. Many AI techniques remain untested in a financial crisis scenario. There have been several instances in which the algorithms implemented by financial firms appeared to act in ways quite unforeseen by their developers, leading to errors. In 2012 Knight Capital lost \$440 million in 45 minutes after deploying unverified trading software. The “Flash Crash” on May 6, 2010 was noteworthy for another reason. Proctor and Gamble swung in price between a penny and \$100,000, but the problem wasn’t caused by bugs or computer malfunctions that verification could have avoided. It was caused by expectations being violated: automatic trading programs from many companies found themselves operating in an unexpected situation where their assumptions were not valid (i.e., they were operating in “out-of-the-box” situations). In 2013, during a 17-minute computer glitch, Goldman Sachs flooded the US market with orders to purchase 800,000 contracts linked to equities and ETFs. During the same week, Chinese brokerage firm Everbright Securities, suffered a malfunction which resulted in it purchasing nearly \$4 billion worth of shares on the Shanghai market<sup>15</sup>. After the Brexit referendum in June 2016, Betterment LLC (a robo advisor that relied heavily on algorithmic trading) suspended trading in response to market volatility to spare its clients higher transactions costs. MIT economist, Andrew Lo has called for developing more robust AI technology capable of adapting to human foibles so that users can employ these tools safely, effectively and effortlessly.

AI and ML developments are moving fast to such an extent where it is almost outstripping the current legal and regulatory framework. The European Union and UK have adopted a more government-led approach to developing AI principles. In 2018, the UK’s introduction of Open Banking gave consumers the ability to compare product offerings and exchange data between providers in a secure way. Other countries such as Singapore, Canada and Iran are also considering adopting some form of open banking regulation.

In 2018, the General Data Protection Regulation (GDPR) also came into force. Under GDPR, EU citizens have the right to receive an explanation for decisions based solely on automatic processing. Furthermore, GDPR stipulates that companies must first obtain consent from an EU citizen before using consumer data. If the EU citizen data is stored on servers located outside of the EU region, GDPR rules apply. Failure to comply to GDPR can result in substantial fines (either up to \$22 million or 4% of a company’s revenues).



As of 2018, the European MIFID II<sup>16</sup> requires firms that apply AI and ML algorithmic models to have a robust development plan in place. Firms should also ensure that potential risks are included at every stage of the plan. In February 2018 the Financial Conduct Authority and Prudential Regulatory Authority released consultation papers on algorithmic trading which lists key areas of supervisory focus in relation to MIFID II.

Last month, 42 countries came together to support a global governance framework for AI. Singapore's AI governance structure is based on a "human-centric" approach, which emphasizes explainability, transparency and fairness to establish public trust in AI<sup>17</sup>.

#### **Other emerging trends that relate to AI.**

Banks are spending massive amounts of money on AI. For example, JP Morgan has invested in COiN, an AI technology that reviews documents and extracts data in far less time than a human. UBS has used AI to trade volatility and JP Morgan uses AI to execute equity trades. JP Morgan, Wells Fargo, Bank of America and Citigroup have increased their IT budgets to pursue AI innovation.

There is also a vibrant merging of financial services and tech companies that specialize in AI. For example, S&P acquired Kensho in 2017 for \$550 million in the biggest AI acquisition to date. Kensho was founded in 2013 with the intention of replacing bond and equity analysts. Its Warren algorithm<sup>18</sup> can process 65 million question combinations by scanning over 90,000 events such as economic reports, drug approvals, monetary policy changes and political events and its impact on financial assets. Google has purchased DeepMind Technologies and Intel has acquired Nervana Systems.

#### **Conclusion**

AI is becoming more ubiquitous in the financial services industry. This will present more legal, ethical, economic and social challenges. AI will also continue to bring new complexities to the global financial ecosystem. As more and more data become available and computing power increases, AI programs will become more complex. In response, AI in financial services needs to be technically robust, secure, protect privacy, be ethically sound and regulation compliant. Ultimately, AI in financial services needs to promote and maintain financial inclusion.

## References

Buchanan, Bonnie. (2019). Artificial intelligence in finance. Alan Turing Institute Research Paper. Available at: <http://doi.org/10.5281/zenodo.2612537>

Buchanan, B and C. Cao (2018) Quo Vadis? Fintech in China Versus the West. SWIFT Institute Working Paper. Available at: <https://swiftinstitute.org/research/quo-vadis-a-comparison-of-the-fintech-revolution-in-china-and-the-west/>

Citi (2018) Bank of the Future: the ABCs of Digital Disruption in Finance. Available at: <https://www.citivelocity.com/citigps/bank-future/>

Duarte, J., Siegel, S., & Young, L. (2012). Trust and credit: The role of appearance in peer-to-peer lending. *The Review of Financial Studies*, 25(8), 2455-2484.

Future Today Institute (2017) Tech Trends annual report. Available at: <https://futuretodayinstitute.com/2017-tech-trends/>.

Herzenstein, M., Sonenshein, S., & Dholakia, U. M. (2011). Tell me a good story and I may lend you money: The role of narratives in peer-to-peer lending decisions. *Journal of Marketing Research*, 48(SPL), S138-S149.

Lin, M., Prabhala, N. R., & Viswanathan, S. (2013). Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending. *Management Science*, 59(1), 17-35.

O'Neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Pope, D. G., & Sydnor, J. R. (2011). What's in a Picture? Evidence of Discrimination from Prosper. com. *Journal of Human Resources*, 46(1), 53-92.

World Economic Forum (2019) Global Risks Report. Available at: <https://www.weforum.org/agenda/2019/01/these-are-the-biggest-risks-facing-our-world-in-2019/>

<sup>1</sup> Source: Statista and Transparency Market Research.

<sup>2</sup> Source: Statista, 2018

<sup>3</sup> Future Today Institute (2017).

<sup>4</sup> Citi (2018).

<sup>5</sup> HFT is the most recognizable form of AT and use high-speed communications and algorithms in financial market transactions. In 2011, HFT firms accounted for 45-50% of US equities trading.

<sup>6</sup> Allstate's Digital Colleague Amelia Answers Questions for Call Center representatives. Sara Castellanos, WSJ. March 30, 2018.

<sup>7</sup> Worldbank (2017)

<sup>8</sup> China's debt collectors focus in on \$200 billion P2P debt pile", Don Weinland, FTimes. June 5, 2018.

<sup>9</sup> Britain urged to take ethical advantage in artificial intelligence. John Thornhill. FTimes. 04/15/2018.

<sup>10</sup> [https://csis-prod.s3.amazonaws.com/s3fs-public/legacy\\_files/files/attachments/140609\\_rp\\_economic\\_impact\\_cybercrime\\_report.pdf](https://csis-prod.s3.amazonaws.com/s3fs-public/legacy_files/files/attachments/140609_rp_economic_impact_cybercrime_report.pdf)

<sup>11</sup> Korea introduces AI app that auto-detects 'voice phishing' scams

<sup>12</sup> Citi (2018).

---

<sup>13</sup> Robots in Finance Bring New Risks to Stability, Regulators Warn. Silla Brush. Bloomberg Magazine. Nov 1, 2017. Available at: <https://www.bloomberg.com/news/articles/2017-11-01/robots-in-finance-bring-new-risks-to-stability-regulators-warn>.

<sup>14</sup> Serious Fraud Office CEO Ben Denison reveals how AI is transforming legal work, Thomas Macaulay, CIO, January 3, 2018.

<sup>15</sup> "17-minute trading glitch put Goldman's reputation on the line", Arash Massoudi and Tracy Alloway, FTimes. August 22, 2013.

<sup>16</sup> MiFID II or Markets in Financial Instruments Directive came into effect early 2018 and is designed to offer greater protection for investors and inject more transparency into all asset classes: from equities to fixed income, exchange traded funds and foreign exchange.

<sup>17</sup> How governments are beginning to regulate AI, Madhumita Murgia and Siddarth Shrikanth, Financial Times, May 30, 2019.

<sup>18</sup> Named after Warren Buffett.

Written Testimony before

**The Task Force on Artificial Intelligence of the  
House Financial Services Committee**

Hearing entitled:

**Ending Perspectives on Artificial Intelligence: Where We Are and the Next  
Frontier in Financial Services**

**R. Jesse McWaters, Financial Innovation Lead, World Economic Forum**

June 26th, 2019

Chairwoman Waters, Ranking Member McHenry, and distinguished members of the Task Force. I am honoured by the invitation to appear before you today at this important hearing on the implications of the growing use of Artificial Intelligence in financial services.

By way of background, the World Economic Forum is a Swiss not-for-profit International Organization for public-private cooperation. It is an independent and impartial organization with a mandate to serve as a neutral platform to support political, business and other societal leaders in shaping of global, regional, and industry agendas. Currently, significant organizational focus is devoted to understanding, and responding to, the 'Fourth Industrial Revolution' – a series of transformative technological breakthroughs in a range of fields (including artificial intelligence) that are disrupting traditional business models and straining traditional approaches to policy-making.

My work within the World Economic Forum is primarily focused on exploring the role that new technologies, including artificial intelligence, are playing in the rapid transformation of the financial system. This work includes: investigating the impact of new technologies on the operating models of financial institutions across a range of sub-sectors and geographies, analysing the shifting competitive dynamics of the industry, and identifying the challenges to effective governance of the financial system that may emerge as a result of these changes.

The testimony that follows is drawn from two documents. The first is a World Economic Forum report titled 'The New Physics of Financial Services – How Artificial Intelligence is Transforming the Financial Ecosystem' which was made publicly available on August 15<sup>th</sup>, 2018. The second is an, as yet untitled, World Economic Forum report currently under development that explores a range of

governance questions emerging from AI's application in the financial sector. This second report is tentatively scheduled for release in September of 2019. The content of both reports has been compiled through a mix of secondary research and in-depth interviews with subject matter experts including representatives from the financial sector, emerging fintechs, large technology companies and regulatory/supervisory authorities. Additional insights were gathered through a series of workshops, conducted around the world, that convened a sub-set of interviewees for structured discussions on the future of financial services. (N.B. All interviews and workshops were conducted under 'Chatham House Rule' where the general findings of the discussions may be publicly disseminated, but the identity and affiliation of individuals may not be shared).

The following testimony is organized around seven key points that summarize for this taskforce what I believe to be the most salient insights from these reports. It is important to understand that these reports, and the points drawn from them, seek to provide useful context and forward-looking insights about potential evolutionary paths for the financial ecosystem. However, they do not seek to diagnose specific regulatory and policy issues related to the technological transformation of the financial sector – nor do they seek to prescribe specific regulatory or policy approaches.

**1. A fundamental lack of understanding of what AI is and how it works poses a serious impediment to effective governance**

The computer systems that we refer to as 'artificial intelligence' possess capabilities that differ innately from those of human beings. Despite this, there is a strong bias to think of AI and its capabilities in human terms. This tendency can lead observers to deploy numerous unhelpful cognitive biases in their analysis of AI, resulting in both the overestimation and underestimation of AI capabilities. For example, we might assume that if a computer can be programmed to perform a task that most humans would find complex – such as analysing the many possible outcomes of a Chess game – that performing the motions necessary to fold laundry must be trivially easy, when in fact the opposite is the case. This inherent foreignness of AI makes it difficult to understand – and to some degree, predict – how AI systems will behave, making fear of these systems a natural response.

To make matters worse, despite being a frequent topic of discussion and debate, the very term 'artificial intelligence' lacks a consistent and broadly accepted definition even among technical experts in the field. This lack of precision can lead to significant confusion and disagreement when it

comes to evaluating the governance requirements and potential impacts of this technology. For the purposes of this testimony, I will employ a very expansive definition of AI: any analytical technique that uses a degree of self-learning for adaptive and predictive purposes. Moreover, rather than focusing on the underlying technical approaches used, this testimony will focus primarily on the ways in which the technology is being used to create value for financial institutions and their clients.

**2. AI's near-term impact on finance is best understood through the capabilities it enables.**

We identified four primary ways in which the suite of technologies commonly called AI are being used to drive value within the financial sector. The first is to improve the speed and efficiency of existing financial processes. Such techniques enable financial institutions to reduce operational costs, while at the same time often improving the experience of their customers. Examples of this category of AI application include the use of machine vision to automate the processing of minor 'scratch and dent' insurance claims and the use of natural language processing to more rapidly onboard complex corporate banking clients with less manual intervention.

The second category of value being created through the deployment of AI are improvements to the performance of existing activities. This is often achieved by combining new data points with advanced analytical techniques to identify new risks and opportunities that might not have been perceived using traditional methods. For example, certain hedge funds may use machine readable news and machine vision analysis of satellite imagery to identify new investment opportunities. Many new consumer lenders have sought to use non-traditional data inputs, such as bill payments or social graphs, to expand their lending operations to 'thin file' clients who might not have had enough traditional credit history to receive a loan.

The third category of value being enabled by AI is the deployment of entirely new value propositions. These are new sources of revenue, typically based on the analysis of data already being collected by the organization. For example, some payment networks may be able to offer macro-economic forecasting services based on the analysis of high-level metadata flowing through their networks, while a growing number of insurers are exploring opportunities to leverage AI's real-time analytic capabilities to help them stop incidents before they happen.

The fourth way in which financial institutions are using AI to create value is the deployment of highly personalized products and advice specifically calibrated to meet customer's needs. While such

offerings have always been possible, AI allows them to be deployed at near zero marginal cost. Many such offerings remain relatively nascent, but a growing number of personal financial advisors look to provide mass-market clients with real-time advice that helps them to better understand their financial situation and spending patterns.

**3. As the adoption of these AI becomes ubiquitous access to data will become a core strategic priority of financial institutions, driving a shift in the competitive landscape of the industry**

The capabilities discussed above represent a remarkable opportunity for financial institutions to differentiate themselves in terms of their efficiency, performance, and customer engagement; establishing an extremely strong rationale for investment in AI-enabled capabilities. Our research suggests that as the number of financial institutions focused on deploying these capabilities grows the implications may extend well beyond the performance of individual institutions – driving a foundational shift in the basis on which financial institutions compete.

To understand the reason for this, we must consider the ‘ingredients’ required to deploy AI-enabled capabilities. At their core, the majority of AI models are a combination of two parts: an ‘algorithm’ that takes certain inputs, analyses them, and provides specific outputs, and a dataset that is used to ‘train’ the algorithm to perform the required task. While the ‘algorithm’ can appear to be complex to the layman, a wide array of cutting-edge algorithms are available for free from open-source communities. This means that the other part – the dataset – is the critical component of any institution’s effort to build unique, high-quality AI models.

As a result, financial institutions’ appetite for data are likely to grow at an accelerating pace to feed the development of their AI systems; this will occur across three dimensions:

- Depth: data that is granular and specific, to develop nuanced understandings and deep insights that create real value for customers
- Breadth: data that spans across a wide set of use cases, to develop robust models that can adapt to a variety of situations
- Exclusivity: data that other institutions cannot easily replicate or procure, to provide value that cannot be replicated by others and create sustained competitive differentiation

This growing importance of data is likely to further underline existing discussion on the importance of adequate data protection that have surfaced in the wake of several high-profile data scandals in

recent years. Numerous contributors to our research stressed that the growing demand for data – particularly where that data is of personal or sensitive nature – will demand the development of improved systems for both safeguarding personal data and enabling customers to share their data in a secure easy to understand fashion.

**4. The early adopters of AI may be well-positioned to establish and entrench their market dominance in specific activities, leading to a more interconnected financial system**

Many contributors to our research stressed that the competitive dynamics of markets where businesses seek to differentiate themselves based on AI capabilities differ significantly from the dynamics in more traditional markets. A key reason for this that AI systems tend to create a self-reinforcing loop, wherein the improvement of a product attracts more users who then provide the system with additional data, leading to further improvements in the product. This loop – sometimes referred to as the AI flywheel, tends to create a ‘winner take all’ dynamic where certain products secure a degree performance and quality that other competitors cannot compete with, making it difficult for them to access the data needed to make improvements.

It is likely that early movers in the deployment of AI capabilities who successfully establish flywheel effects will be able to establish difficult-to-challenge offerings around specific service offerings. In some cases, these will be end-customer facing offerings, while in others they may be back- and middle- office capabilities. Success in the later would create an impetus for institutions to consider if it is rational to continue performing those activities in house, or if it would be more reasonable to outsource those activities to the third party whose AI has established itself as best-in-class at that activity. The provider of such services might be a competing financial institution, a large technology firm, or even a highly specialized fintech. As adoption of such services became more common, the landscape of financial services would be fundamentally transformed as complex interconnections between more specialized financial institutions (and non-financial entities such as cloud service providers) became typical.

**5. The shifting landscape of financial services resulting from increased use of AI will create new uncertainties and has the potential to drive governance gaps**

As financial institutions seek to incorporate AI-enabled capabilities into their operating models and brace themselves for longer-term changes to their competitive environment, it is unsurprising that



members of both the public and private sector are concerned about ensuring good governance and stability. Over the course of the last eight months we have conducted workshops and interviews with experts from leading financial institutions, emerging fintechs, and regulatory authorities around the world. Three issues that were commonly raised by stakeholder from a variety of organizations were the following:

Systemic risk: As AI becomes more ubiquitous and as financial institutions become more interconnected (relying on one-another for core capabilities as proposed in point four) new vectors for the propagation of systemic risk could emerge that existing prudential models are not equipped to identify or defend against.

Bias: In seeking an 'information edge', financial institutions are increasingly using new types of data (e.g., social media data, telecom data) to inform their AI systems. The use of new forms of data could potentially allow for bias to be propagated through the financial ecosystem and lead to unfair outcomes for specific populations.

Explainability: Understanding AI models is difficult – sophisticated systems can involve orders of magnitude more intermediary steps than traditional systems. This makes it almost impossible to follow how the provided inputs led to the outputs of an AI model, and often even the developers who built a model cannot fully explain how it works. This creates significant complexity in adequately governing AI systems both from an internal business perspective and a regulatory perspective.

**6. While AI may demand changes to existing regulatory and governance practices it does not require a fundamental re-thinking of regulatory principles and objectives**

The risks and potential governance gaps illustrated in point five are clearly serious: ensuring the continued stability of the financial system is critical to enabling growth in the broader economy and safeguarding the fairness of the financial system is a cornerstone of the industry's social licence. At the same time, it is important to highlight that few, if any, of the risks raised by the increased use of AI in financial services demand a fundamental rethinking of the objectives of financial regulation or the high-level principles that the financial system should be held to. Instead, establishing effective governance requires a granular investigation of how best to achieve these objectives in the context

of new tools and operating models. For example, when we consider the case of the three risks identified in point five, it is critical to keep in mind that...

Systemic risk: New supervisory capabilities supported by AI can augment regulator's ability to monitor and proactively respond to emerging systemic risks against a changing financial landscape. This potentially allows the increasing complexity of the financial ecosystem to be balanced out by the increasing sophistication of oversight and governance mechanisms.

Bias: Extensive work has been and continues to be done on limiting the propagation of bias in financial decision-making systems. Many well-established techniques exist to safeguard against the introduction of bias via human factors, decision-making systems, and the input data that informs the decision-making system. Additional techniques can be used to identify unfair outcomes across various 'fairness metrics'; proactive and reactive techniques to managing bias can both still be used in AI contexts.

Explainability: 'Explaining' a model can take on many forms and providing full interpretability of a model may not be necessary in all cases, depending on the need that an explanation seeks to fill. Other techniques, such as 'guard-rails', context generation engines, and the ability to interrogate limited portions of a model may be sufficient to deliver the level of context, transparency and control deemed necessary to have informed trust in a model for a particular use case.

**7. The use of AI in financial services introduces both opportunities and risks – effectively balancing them is the responsibility all stakeholders of the financial system**

As discussed in point one, popular discourse around AI is often tinged by fear and frustrated by the highly complex nature of this topic. Many of the concerns around AI's deployment in financial services – including but by no means limited to systemic risk, bias, and explainability – require close consideration.

However, it is critical to recall that when strong governance practices can be put in place and risks can be suitably managed, that AI has the potential to deliver significant benefits to end-users of financial products and services – particularly individuals and small businesses. AI can help provide faster and more efficient processing of customer requests, increase accessibility by opening models

up to new data points, and improve the quality of financial advice by allowing institutions to more deeply understand their customers.

The core challenge for this committee, the regulatory apparatus of the US financial system, and the financial institutions that it regulates will be to effectively navigate the balancing of these risks and opportunities based on the specific context of individual use cases of the technology. AI is not inherently good nor evil; it is a toolkit that has many potential benefits, but like any toolkit, needs to be used responsibly.

**Douglas Merrill**  
**CEO ZestFinance**  
**Testimony to the House Committee on Financial Services AI Task Force**  
**June 26, 2019**

Chairman Foster, Ranking Member Hill, and members of the task force, thank you for the opportunity to appear before you to discuss the use of artificial intelligence in financial services.

My name is Douglas Merrill. I'm the CEO of ZestFinance, which I founded ten years ago with the mission to make fair and transparent credit available to everyone. Lenders use our software to increase loan approval rates, lower defaults, and make their lending fairer. Before ZestFinance, I was Chief Information Officer at Google. I have a Ph.D. in Artificial Intelligence from Princeton University.

The use of artificial intelligence in the financial industry is growing in areas like credit decisioning, marketing, and fraud detection. Today I will discuss a type of AI — machine learning (a.k.a ML) — that discovers relationships between many variables in a dataset to make better predictions. Because ML-powered credit scores substantially outperform traditional credit scores, companies will increasingly use machine learning to make more accurate decisions. For example, customers using our ML underwriting tools to predict creditworthiness have seen a 10% approval rate increase for credit card applications, a 15% approval rate increase for auto loans, and a 51% increase in approval rates for personal loans — each with no increase in defaults.

Overall, this is good news and it should be encouraged. Machine learning increases access to credit especially for low-income and minority borrowers. Regulators understand these benefits and, in our experience, want to facilitate, not hinder, the use of ML.

At the same time, ML can raise serious risks for institutions and consumers. ML models are opaque and inherently biased. Thus, lenders put themselves, consumers, and the safety and soundness of our financial system at risk if they do not appropriately validate and monitor ML models.

Getting this mix right—enjoying ML's benefits while employing responsible safeguards—is very difficult. Specifically, ML models have a “black box” problem; lenders know only that an ML algorithm made a decision, not why it made a decision.

Without understanding why a model made a decision, bad outcomes will occur. For example, a used-car lender we work with had two seemingly benign signals in their model. One signal was that higher mileage cars tend to yield higher risk loans. Another was that borrowers from a particular state were slightly less risky than those from other states. Neither of these signals raises redlining or other compliance concerns. However, our ML tools noted that, taken together, these signals predicted a borrower to be African-American and more likely to be denied. Without visibility into how seemingly fair signals interact in a model to hide bias, lenders will make decisions which tend to adversely affect minority borrowers.

There are purported to be a variety of methods for understanding how ML models make decisions. Most don't actually work. As explained in our White Paper and recent essay on a technique called SHAP, both of which I've submitted for the record, many explainability techniques are inconsistent, inaccurate, computationally expensive, or fail to spot discriminatory outcomes. At ZestFinance, we've developed explainability methods that render ML models truly transparent. As a result, we can assess disparities in outcomes and create less-discriminatory models. This means we can identify approval rate gaps in protected classes such as race, national origin and gender and then minimize or eliminate those gaps. In this way, ZestFinance's tools decrease disparate impacts across protected groups and ensure that the use of machine learning-based underwriting mitigates, rather than exacerbates, bias in lending.

Congress could regulate the entirety of ML in finance to avoid bad outcomes, but it need not do so. Regulators have the authority necessary to balance the risks and benefits of ML underwriting. In 2011, the Federal Reserve, OCC, and FDIC published guidance on effective model risk management.<sup>1</sup> ML was not commonly in use in 2011, so the guidance does not directly address best practices in ML model development, validation and monitoring. We recently produced a short FAQ, which we've also submitted for the record, that suggests updates to bring the guidance into the ML era. Congress should encourage regulators to set high standards for ML model development, validation and monitoring.

We stand upon the brink of a new age of credit. An age that is fairer and more inclusive, enabled by new technology — machine learning. However, “brink” can also imply the edge of a cliff; without rigorous standards for understanding why models work, ML will surely drive us over the edge. Every day that we wait to responsibly implement ML keeps tens of millions of Americans out of the credit market or poorly treated by it. Thank you for your time and attention.

---

<sup>1</sup> <https://www.occ.treas.gov/news-issuances/bulletins/2011/bulletin-2011-12a.pdf>



**Douglas Merrill**  
CEO & Founder  
ZestFinance

(Former Chief Information Officer of Google)

Douglas Merrill is the CEO and founder of ZestFinance, a Los Angeles-based financial services technology company that uses machine learning and data science to predict credit risk and make more accurate underwriting decisions. Zest helps lenders deploy fully explainable machine learning models and to make their lending fairer and more inclusive. Backed by some of Silicon Valley's most prominent venture capitalists, Zest's partners include Discover Financial Services, Ford Motor Co., Synchrony Financial, one of Turkey's largest banks, and Baidu.

In 2009, Douglas started ZestFinance with a hypothesis: Google-like algorithms could be applied to make consumer credit more transparent, available to more people, and significantly less expensive. ZestFinance's team of data scientists and mathematicians are united by a unique mission: to make fair and transparent credit available to everyone.

Prior to founding Zest, Douglas was the Chief Information Officer of Google for six years. Douglas led an organization of 15,000 staff, oversaw all aspects of internal engineering and technology, and drove multiple strategic efforts, including Google's IPO auction in 2004.

Douglas also served as Senior Vice President of Infrastructure and HR Strategy at Charles Schwab. In academia, Douglas was an Information Scientist at the RAND Corporation, where he conducted highly classified research for several branches of the U.S. armed services.

Douglas holds a Ph.D. in artificial intelligence from Princeton and is the author of *Getting Organized in the Google Era: How to Get Stuff Out of Your Head, Find It When You Need It, and Get It Done Right*. His academic publications include articles in *The Journal of the Learning Sciences*, *Cognition and Instruction*, *Reliable Distributed Systems*, and a paper in the book series *Lecture Notes in Computer Science*.

# Beyond The Black-Box: A Better Framework For Explainable AI

By Evan Kriminger, Mark Eberstein, Sean Kamkar, Jose Valentin, Douglas Merrill  
ZestFinance

November 2018



Copyright ©2018 ZestFinance, Inc. All Rights Reserved. Confidential and proprietary. No part of this document may be disclosed in any manner to any third party without ZestFinance's prior written consent.

## 1 Table Of Contents

- [1 Table Of Contents](#)
- [2 Introduction](#)
- [3 What is explainability?](#)
  - [3.1 Conveying explanations](#)
  - [3.2 Interpreting a model](#)
- [4 Importance of explainability](#)
  - [4.1 Stakeholders](#)
  - [4.2 Explainability in credit](#)
- [5 Explainability techniques](#)
  - [5.1 Intrinsic explainability](#)
  - [5.2 The challenge of explainability](#)
  - [5.3 Univariate perturbations](#)
  - [5.4 Visualization](#)
  - [5.5 Proxy models](#)
  - [5.6 Gradient methods](#)
- [6 Evaluating explanations](#)
  - [6.1 Comparison metrics](#)
  - [6.2 Properties of a good explanation](#)
  - [6.3 User testing](#)
- [7 Two experiments show limits to current explainability techniques](#)
  - [7.1 Experiment One: Implementation issues of LIME](#)
  - [7.2 Experiment Two: Explaining performance on real credit data](#)
    - [Variance](#)
    - [Sensitivity to hyperparameters](#)
    - [Smoothness](#)
    - [Precision](#)
    - [Accuracy](#)
- [8 Conclusion](#)
- [9 Explainability At Zest](#)
- [10 References](#)
  - [10.1 Internal References & Code](#)



---

## 2 Introduction

Machine learning (ML) is a subset of artificial intelligence that focuses on the design of systems that can learn from and make decisions and predictions based on large information sets. It has become the standard for producing powerful data models that automate decision-making, often in high-stakes use cases. Its effectiveness has been proven in diverse fields such as natural language processing, robotics, recommendation engines, finance, and healthcare. The research community continues to substantiate the superior predictive power of these new algorithms over traditional methods such as logistic regression. Unlike status quo methods, ML models accommodate non-linearities, multivariate interactions, and generalize well to new datasets all within a single model -- improving accuracy and reducing complexity and risk.

Despite the clear benefits of machine learning, the use of logistic regression models continue to be the norm, especially in risk- and prediction-related business such as credit and underwriting. There are several reasons for this. One is that financial institutions (FIs) do not have the in-house expertise required to build, train, and deploy advanced ML models. More user-friendly ML modeling tools will help close this knowledge gap. The more imposing obstacle to adoption is regulatory and business risk. The Federal Reserve, the Office of the Comptroller of the Currency, and the Federal Deposit Insurance Corporation have all issued guidance dictating clear and documented model risk management: how and why a model that an FI has put into production arrives at the results. Explainable machine learning models should be the standard for FIs not only to meet regulatory requirements but also to illustrate their decision-making process to clients and business stakeholders.

In this paper, we define explainability in terms of the problems it solves, the principles on which it is based, and the way in which it conveys information about the model. We present an overview of methods in use in the market, along with techniques for evaluating the quality of explanations each technique delivers. Popular explainability methods have systematic shortcomings: they are often computationally expensive, restricted to certain classes of models, and they suffer from failure modes, which could lead to catastrophic outcomes such as race-based discrimination. Solving these issues is the focus of a considerable research effort, with the goal of efficiently explaining expressive models such that the explanations provide an accurate picture of the model's behavior in a human-interpretable form. Enabling the safe application of modern machine learning techniques is the key to revolutionizing high-stakes business problems like credit underwriting.

## 3 What is explainability?

To explain a model is to relate the model's decisions to the input data on which its decisions are based. This is a notably vague definition compared to the performance goal of machine learning, which is to make highly accurate predictions in a range of high-stakes applications,

Copyright ©2018 ZestFinance, Inc. All Rights Reserved. Confidential and proprietary. No part of this document may be disclosed in any manner to any third party without ZestFinance's prior written consent.

such as consumer finance and healthcare. While a useful optimization criterion for modeling might be classification error, the success of explaining a model is a more qualitative outcome. What does it mean for a model to be properly explained? To approach the difficult task of explaining machine learning models, the problem setting must be well-established. Which type of model is to be explained? What form does the information provided by the explanation take? What are the desired outcomes for interpreting this model? These questions dictate the principles of a satisfactory explainability approach.

### 3.1 Conveying explanations

The typical end-product of an explainability tool reveals the contributions or influence of each input feature towards the model prediction. In the context of machine learning, features are individual input columns. This feature-level explanation is considered local if it applies to a single input sample or global if it describes feature contributions over all samples. A different approach is to use example data points. Exemplar-level approaches explain a prediction in terms of the training input which led to that prediction. Thus, explanations can exist at the level of features, individual samples, or the model as a whole.

For example, in an autonomous vehicle, explaining the decision to turn involves generating a heat map over the input image revealing which features in the image led to that decision. An English to Spanish translation model can be explained by highlighting which English words led to each Spanish word in the translation. For credit underwriting, the decision to reject a loan applicant is explained by highlighting the fields in the loan application which led to the rejection decision. This may also include the features with the most important contributions, both positive and negative, to the applicant's score.

The interpretation of model behavior at a particular input is often less meaningful than explaining behavior at the input relative to a point of reference. This referential form of explainability is more intuitive for certain application domains. In credit underwriting, returning the reasons for a rejection are more informative in the context of an accepted applicant. The customer and regulators are less interested in the explanation of the model's exact probability output than the explanation of why the applicant was rejected compared to an accepted reference applicant.

### 3.2 Interpreting a model

Explanations reveal how a model uses its inputs to make predictions. With feature-level explanations, this is a breakdown of each feature's contribution to the output. However, it is only in the case of linear models that the feature contributions are exactly the linear model coefficients for each feature. For nonlinear models, the explanation does not convey information as transparently. It is therefore important to assure that the explanation of a complex model can be trusted to depict model behavior accurately.

---

There are a few defining properties of a reliable explanation, which aid in the development and evaluation of explainability techniques:

**Consistency.** A consistent explainer should not rely on meticulously tuned parameters and should provide reasonable results for a wide range of parameters. It should not be prone to large random fluctuations between repeated runs of the program. Logically equivalent models should yield the same explanations. Similar inputs should receive similar explanations.

**Accuracy.** An explainer should be fully representative of the true dynamics and behavior of the model. The problem of interpreting complex machine learning often leads to some simplifying assumptions. For example, the proxy model explanation technique assumes that a simple, interpretable model can serve as a proxy to explain a more complicated model, as long as their input-output behavior is similar enough. The proxy model assumption is reasonable only if a high-level interpretation of the model is needed, such as the feature importance across all inputs. For explaining individuals, the proxy is going to produce very different results from the target model. Other assumptions such as monotonicity and independence of the input variables cannot be guaranteed and result in explanations which do not reflect reality.

One tangible target for the accuracy of an explanation is the sensitivity of a model, which refers to how a model's output is affected by a small perturbation to its input. Variables which are declared to be important by an explanation should have a significant impact on the model's output when perturbed.

**Interpretability.** Explanation values are not normalized quantities such as with probabilities, and amplitudes of explanations may vary greatly between methods. If the desired end product of an explanation is the relative feature importance, then this is not an issue, but providing values in interpretable units is preferable. Some explainers provide an attribution to the features which sums to the model's output. Thus, the explanation can be directly interpreted as each feature's contribution to the output. Explanations are harder to interpret in isolation, so some techniques are referential. This means that they explain the prediction for a particular input in reference to the model's treatment to a baseline input. For many problems a referential explanation is natural. For example, in credit underwriting, the decision to deny an applicant credit may be explained relative to a reference-approved applicant.

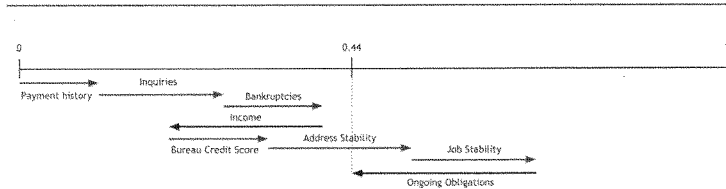


Figure 1: A credit model's prediction of 0.44 decomposed into contributions from the input features.

**Practicality.** Explainability techniques can be quite computationally expensive. Explainers must scale well with the complexity of the model and the size of the dataset if they are to offer real-time explanations. An important consideration is the types of models which a technique is capable of explaining. Black-box methods can explain any model, while others focus on trees or neural networks. Ensembles of models offer robustness and improved performance, yet interpreting heterogeneous submodels poses a difficult task for explainability.

#### 4 Importance of explainability

The predictive ability of a machine learning model is only one component of a complete explanation. Predictions optimize a particular objective function, such as mean squared error, but this is an incomplete picture of the overall success of the model. A complete explanation needs to address numerous regulatory concerns that the standard model validation process takes for granted, as it was developed for more easily interpretable logistic regression-based models. In order to put machine learning models into production for everyday use, their explainability methods have to meet the following regulatory model validation requirements [Doshi-Velez and Kim, 2017]:

- **Fairness:** It is necessary to identify unlawful or inappropriate model biases. For example, some features may be a proxy for race or gender. A model should not be overly reliant on these features, and not treat individuals from protected classes differently than those from an unprotected class with comparable inputs.
- **Safety:** Identify failure modes which may be extrapolated from the features the model uses in making its decision. Such failure modes may not be obvious in summary statistics derived from validation sets. The model may be getting good results from obviously flawed logic or poor results due to unobvious changes in input data. Consider the stopping decisions an autonomous vehicle makes for pedestrians. It is important to interpret the model's decisions to ensure that the pedestrian is the driving force of the stop, and not just the presence of signs or crosswalks. Otherwise, the model may not respond to pedestrians in unprotected crossing situations.

- **Objective misalignment:** Although the model's objective function may be a proxy for a business objective, the model must be analyzed to ensure it is adequately addressing its intended purpose. For example, optimizing a diet to lower cholesterol may not produce a healthy diet if the dieter chooses nonfat, but high sugar foods.
- **Security:** A model that is not well-understood could be exploitable by attackers. The model can be manipulated if easily modifiable features can change the outcome. Confirming that the model does not rely on such "gameable" features helps ensure its security.
- **Model health monitoring:** In production, the model's performance could degrade considerably if the characteristics of the population diverge from the training set. It may be difficult to acquire ground truth target values for more recent examples to detect such a change. Additionally, in more critical applications, monitoring may need to be real-time. Using explainability to monitor model health ensures that the functionality of the model is the same as it was during training. This is a more direct test for the model's health than detecting shifts in the input data alone, with no regard to the model.

#### 4.1 Stakeholders

Explainability impacts many parties, and understanding how these impacts play out is an important step in designing machine learning products.

- **Regulators** have laid out model risk management criteria. While regulatory guidance was based on old techniques, that guidance still applies to ML models in production. Unfortunately, you cannot meet regulatory requirements for ML models without explainability. In the European Union, for instance, the General Data Protection Regulation (GDPR), as of 2018, grants a "right to explanation" for users subject to automated or AI-based decision-making. In the U.S., state regulators are also moving in the direction of demanding greater explainability, and federal regulators are actively reviewing the use of alternative data and advanced mathematical techniques in automated decision-making.
- **Consumers** also benefit from receiving explanations of decisions that can have a profound impact on their lives. Automated decisions can be troubling, and clear explanations of how an automated process arrived at a decision can provide confidence that they were treated fairly, and offer a path of correct behavior towards a different outcome. At ZestFinance, our goal is to expand the availability of fair and transparent credit, which requires providing consumers with interpretable decisions.
- **Management** must assess the business impacts and the risks associated with deploying machine learning models. Explaining models puts algorithmic decisions in a context that is consistent with the business logic of human decision makers. Explainability tools may themselves be an important product offering alongside machine learning models. Business leadership has more options for what to include in a complete machine learning product.

- **Data scientists** must consider explainability in their modeling decisions. This is an emerging field, with an ever-evolving set of techniques and empirical results. Research and development efforts may need to focus more on explainability than the actual modeling itself. Developments in explainability enable data scientists to build more powerful models and provide a valuable tool for debugging and validation.

#### 4.2 Explainability in credit

In consumer credit underwriting, explainability is just as important as the model itself. The need for explainability is explicit for credit underwriting systems, as specified by the Equal Credit Opportunity Act (ECOA) Regulation B and Fair Credit Reporting Act (FCRA). These regulations require that lenders supply *adverse action* notices, which inform the applicant of the reasons for the denial of credit. The law establishes the basis on which applicants cannot be denied credit. For example, discrimination is prohibited on the basis of race, sex, age, national origin, or marital status, i.e., the *disparate impact* of particular classes.

The legal requirement of interpretability has led the industry to be dominated by simple, inexpressive models such as logistic regression. Expanding credit to more good borrowers (without added risk) can only happen with the wider adoption of machine learning. This can only happen by applying novel explainability techniques. In the following sections, we describe the current explainability methods and their performance ability, followed by a comparison to Zest's explainability method and performance.

### 5 Explainability techniques

Methods for explainability can be categorized based on the type of model they explain and the criteria which these explanations seek to satisfy. The explanation can also convey information in different forms, as described earlier.

Black-box explainability techniques derive explanations solely from input-output behavior, without considering the model internals. Black-box techniques have the benefit of versatility and apply to any class of model. The tradeoff for this versatility is the difficulty in accurately characterizing model behavior without utilizing any specific information from the model. Black-box techniques are also often computationally expensive, which roughly stems from the need to enumerate many test cases in an attempt to fully explain the decision space. White-box techniques exploit the structure of the model in interpreting its decisions. While white-box techniques have more information available to produce explanations, this comes at the expense of losing the flexibility of black-box methods.

### 5.1 Intrinsic explainability

Two model types are considered to be inherently explainable: linear models and decision trees. For this reason, they have been workhorses in regulated industries. A linear model makes predictions of the form

$$f(x) = a_1x_1 + a_2x_2 + \dots + a_nx_n + \varepsilon$$

Where  $x_i$  is the  $i$ th feature on the input,  $a_i$  is the corresponding coefficient (weight), and  $\varepsilon$  is the bias term. Explaining a linear model is simple because the contributions of each feature are by definition additive. Each feature contributes its value weighted by the coefficient associated with it, i.e.  $a_ix_i$ . The coefficients of the linear model serve as measures of feature importance, and these do not change across all inputs. Thus, local and global explainability are the same for linear models. For classification problems, with a discrete number of classes, modelers use a logistic regression model that takes the form:

$$f(x) = \sigma(a_1x_1 + a_2x_2 + \dots + a_nx_n + \varepsilon)$$

In logistic regression, the linear model is transformed by the nonlinear logistic function, which scales the output to behave like a class probability. Since this function is monotonic, logistic regression models can be explained by their underlying linear model.

Decision trees arrive at classification decisions by following decision paths determined by querying individual features. Each node in a decision tree is associated with a test that a certain feature is above a given threshold, with the result of this test determining the next node. Leaf nodes in the decision tree represent decisions.

Predictions made by a decision tree are the result of very explicitly stated conditions on a handful of features. Explaining a decision tree is as simple as returning the decision path. This is a local explanation, but a global summary of model behavior can be derived from measures of the frequency with which each feature is used in the decisions.

Other simple classifiers may be considered explainable. The k-nearest neighbor's classifier assigns inputs to the class of the nearest input in the training set, and thus inherently offers exemplar-level explanations. The naive Bayes classifier uses the independence assumption to represent the class posterior probability as the product of likelihoods for each feature. The feature likelihoods represent the relative importance of each feature towards the model's prediction.

## 5.2 The challenge of explainability

While simple models offer out-of-the-box explainability, unlocking more powerful models requires advanced explainability methods. The most expressive models with the greatest potential performance gains provide no innate interpretability. The classifiers discussed above do not scale to high dimensional data such as images, making them unusable for many real-world problems. Ensembled models offer robustness, but explaining them carries the added difficulty of ensuring that the submodel explanations are compatible. The relative ordering of feature-level explanations may be the only meaningful information that can be extracted, which prohibits the comparison of explanations between models.

## 5.3 Univariate perturbations

Explainability of a black box model requires understanding how the model responds to its inputs. Consider the task of local explainability. A natural approach is to perturb a given input and observe the effect on the output. If the model is highly sensitive to the perturbation, then the features involved were important to the prediction. Fully characterizing the model may require testing any possible perturbation, which is computationally intractable and cannot return a concise explanation of the model function. The most common procedure is to observe the effect that a single feature has on the model output, and use this a measure of that feature's importance.

One type of perturbation is to remove a feature entirely, which is the basis for Leave-one-covariate-out (LOCO) [Lei et al. 2018] feature importance. The impact of removing feature  $x_i$  is  $f(x) - f(x_{-i})$ , where  $x$  is the original input and  $x_{-i}$  is the input without feature  $i$ . Another approach is to add noise to a feature rather than removing it entirely. Permutation feature importance (PMI) [Breiman 2001] is a global explainability method in which, for each feature, the values of that feature are shuffled across all input samples. The intuition is that if a feature is not being incorporated into the model's decision, then replacing it with arbitrary values will not affect performance.

Univariate perturbation approaches are simple to implement and model-independent but suffer from a few important drawbacks. The process of perturbing each feature individually is computationally demanding. The model must be evaluated for each feature, and for each perturbation that is required to get a sufficient estimate of the model impact.

Not only are univariate perturbation approaches computationally intensive, they often also yield wildly inaccurate results. Measuring only univariate variable effects does not explain model behavior that depends on variable interactions. For example, in credit underwriting, the impact that an applicant's income has on their score depends on the loan amount. For a small loan, greatly increasing an applicant's income is probably not going to have a large impact on their



score. A permutation of income, even a large increase, might appear irrelevant to credit. This is likely to be wrong.

One must take care when perturbing inputs that the perturbed input still makes sense. Consider a variable such as a car's down payment. Permutation feature importance could, by substituting a down payment for a Porsche on a cheaper Ford Fusion, end up evaluating the model on applications with down payments that are greater than the loan amount itself. Not only do these "made up" data points fall well outside of the space of data that the model was trained on, they violate the fundamental logic of the problem domain.

#### 5.4 Visualization

Univariate or bivariate feature importance can be demonstrated graphically. Partial Dependence Plots (PDPs) [Friedman 2001] compute the model output for fixed values of features under study, averaged over all input samples. This produces a curve (1D) or heatmap (2D) of how a feature's values generally influence the output. PDPs lose information by averaging over all inputs, which again ties into the problem of variable interaction. Independent Conditional Expectation (ICE) [Goldstein et al. 2015] plots attempt to alleviate this issue by plotting multiple response curves, which represents a split of the input samples by conditioning on certain variables. While useful, these tools provide graphical insight into the effect of a feature rather than true model explanations. In addition, such plots require human interpretation for each feature, which is hard for large numbers of features.

#### 5.5 Proxy models

A complicated model can sometimes be sufficiently approximated by a simpler explainable model. If this proxy model can be shown to behave closely to the original model, then its explanation may be similar as well. Student-teacher [Bucilua et al. 2006] learning is one way to generate simple proxy models. The teacher is trained on the given task and used to generate predictions on the whole dataset. The student model is a less complex model that is trained on the same input data but with the teacher's predictions as its target. The intuition here is that the predictions of the teacher provide a more easily learnable target for the simple model with less expressive power. While this paradigm makes sense for models of the same type, such as neural networks, differing only in architecture, it is unlikely that an inherently explainable model such as logistic regression or a decision tree would be able to accurately represent a powerful teacher. This can be clearly seen by trying to fit a linear proxy model to a parabola  $y = x^2$ . Even if two models make the same predictions on a set of examples, this is not a strong argument that they use the data in the same way.

Local Interpretable Model-Agnostic Explanations (LIME) [Ribeiro et al. 2016] notes that simple, explainable models can locally approximate complex models. Explaining a model, therefore, involves building an explainable model at the point of interest. In LIME, data is sampled in the vicinity around the test point and used to train a proxy model (logistic regression or decision

tree) with the samples weighted inversely proportional to their distance from the test point. LIME provides black-box explainability and has empirically shown promising results, but suffers from some practical drawbacks. The process of sampling and training a proxy model for each input is computationally expensive. Sampling itself may be difficult in high-dimensional datasets where it is difficult to choose appropriate metrics and parameters to define a local neighborhood.

### 5.6 Gradient methods

Another approach to explainability is to capture the sensitivity of a model's output to its inputs. For differentiable models, such as neural networks, this amounts to taking the gradient of the output with respect to the input [Simonyan 2013]. The simple case of a linear model reveals the effectiveness of this approach. Given the model,

$$f(x) = a_1x_1 + a_2x_2 + \dots + a_nx_n + \varepsilon,$$

take the gradient with respect to a feature  $x_i$ ,

$$\partial f / \partial x_i = a_i,$$

to reveal that the sensitivity of the model to  $x_i$  is simply the coefficient  $a_i$ , which weights that feature.

While the gradient applies to arbitrary differentiable models and corresponds to an intuitive definition of explainability, it suffers from a few drawbacks. Backpropagation is used to pass gradients from the output of the network to the inputs. Due to the use of nonlinear activation functions which rectify and clip the signals, neural networks have many "flat" regions in which the value of the gradient is zero. A zero gradient suggests that the factor does not matter, which could be correct, but complicated nonlinear models do actually extract information from flat regions. For this reason, plain gradients are not a perfect solution for propagating contribution through a network. Recent research [Kindermans et al. 2017] addresses this issue with techniques to pass gradients through flat regions. Many of these techniques are specific to certain architectures, activation functions, and application domains.

## 6 Evaluating explanations

The quality of an explanation is difficult to evaluate relative to the well-defined goal of model accuracy in supervised machine learning. Evaluating explanations is easier for visual machine learning tasks, such as object recognition, as the quality of the explanation can be compared to a human's understanding of the important aspects of a scene. Other problem domains lack an obvious intuitive explanation and a human oracle may not be available for every input of interest. Automated evaluation of explanation quality is important to advance the field of explainability. For business applications, metrics can ensure the reliability of the explanation.

Copyright ©2018 ZestFinance, Inc. All Rights Reserved. Confidential and proprietary. No part of this document may be disclosed in any manner to any third party without ZestFinance's prior written consent.

---

Aside from the lack of a ground truth, explaining a model is an ambiguous problem. If an explanation seems suspicious, there are two possibilities:

1. The model is behaving erratically, and the explainer is accurately describing this behavior.
2. The explainer is not properly describing model behavior.

Deciding which of these two possibilities is truth is hard even for tasks where useful features are inherently known. This is problematic. A model explainability technique should be able to accurately describe model behavior. What is needed are means of assessing the accuracy and utility of an explainability approach for a given task.

The process of evaluating an explainer is therefore not as simple as ensuring that its explanations put more weight on generally important features. In the literature, explainability techniques are often proposed along with fundamental properties that they satisfy. Confidence in the explainer is derived by showing, either theoretically or experimentally, that it satisfies the properties of a good explainer. Another approach to evaluating explainability techniques is based on the true goal of model interpretability, which is the information it conveys to the human user. A variety of experiments can be devised to test how explanations assist a user in the desired task.

### 6.1 Comparison metrics

The ability to compare two explanations is useful for evaluation purposes. Metrics for explainability are vital in research and development as well, providing a means of measuring improvement or regression from the state of the art. This is similar to distance metrics which compare two input samples. For two explanations to be similar, they must rank the features similarly in importance. Ranks may be compared with Spearman's rank correlation coefficient [Pirrie 2004], a nonparametric estimator of correlation in the orderings of two variables. Two explanations which produce the same ordering of feature importance receive a Spearman's coefficient of +1. The rank of the least important features is not very meaningful because many features contribute trivially and their relative ordering is noise.

Spearman's coefficient weights the entire ordering equally, which may result in an explainability metric that does not align with human intuition. In top- $k$  intersection, the top- $k$  features for each explanation are computed and the number of features in common is used as a measure of similarity. This is more in line with the human perception of which features are driving a decision. The parameter  $k$  can be set by determining the number of features after which the contribution values fall to a trivial level.

## 6.2 Properties of a good explanation

The following properties, while not an exhaustive list, provide an extensive framework for comparing and evaluating explainability methods. These were developed through research and development at ZestFinance and curated from the literature on explainability.

- **Sensitivity to hyperparameters**

In supervised learning, the selection of hyperparameters can be conducted methodically through a carefully designed cross-validation process. Explainability methods also have hyperparameters, but lack of ground truth values preclude the same cross-validation process. Many practitioners resort to simply observing the top features from the training set in aggregate. With this difficulty in validation, it is important that explainers not be overly sensitive to their hyperparameter values. The property can be tested by perturbing hyperparameters and observing the change in the explanations.

- **Variance**

Explainers that are not deterministic will produce different explanations each time they are run, even on the same input sample. Small variations in the amplitudes of the feature attributions are permissible, but if the rank ordering of features changes between runs this creates serious issues in the reliability of the explainer.

- **Smoothness**

If two input samples are extremely similar and are scored identically by the model, then one would expect the explanations for each of these inputs to be similar as well. Without such smoothness, explanations appear random and unreliable. To evaluate the smoothness of an explainer, take the nearest neighbor samples with the same predictions, and measure how close their explanations are. It is expected that if two samples are close in input space, then their explanations are highly correlated.

- **Precision**

Complex machine learning models make their decisions based on complicated relationships between their variables. It is important that an explainer has the power to follow the subtle behavior of a model. Some explainability methods seek to simplify the problem with proxy models, but if the proxy model is too simple, the details of local behavior will be missing in the explanations. One way to test for the precision of an explainer is to look at the diversity of features reported in the top-k. An explainer with no locality, such as a global linear proxy model, will always return the same features.

- **Accuracy**

If a given variable or feature is important to a model's output, the explanation should report it as important. Establishing the accuracy of feature impact is a principle that guides many of the commonly used explainability methods such as permutation

Copyright ©2018 ZestFinance, Inc. All Rights Reserved. Confidential and proprietary. No part of this document may be disclosed in any manner to any third party without ZestFinance's prior written consent.

importance and partial dependence plots. The ability to change the prediction of a model by changing the most important features is, therefore, a very intuitive measure of the explainer's accuracy. It would be disconcerting to a denied credit applicant, for example, if the most important factors in the decision could not, in fact, change their score.

- **Treatment of correlated variables**

It is difficult to assign credit to correlated features even in linear models. Perturbation-based explainability techniques may miss the impact of these features entirely by perturbing them in isolation. The desired attribution amongst correlated variables is not clear and depends heavily on what the model is doing, however, the behavior of an explainer should be understood and consistent. A related problem is with one-hot encodings of categorical variables. If one of the categorical values is a particularly important predictor, then it is preferable to attribute the indicator of this value. Because only one of the one-hot indicators may be 1 for a given input, an explainability method may decide to weight the less important indicator, since its value being 1 implies that the truly important indicator is 0.

- **Robustness to outliers**

Explainability methods may grow unstable near the edges of the data space. This is a concern for methods like LIME which are based on samples from the distribution of the data. Ensuring consistency for outliers is important since explanations ensure that the model is using reasonable features for inputs in regions where the model did not have much training data.

- **Computational cost**

For production systems, explanations may need to be real-time or near real-time. Thus, an explainer should be able to efficiently explain both single inputs and batches of inputs. Explainability techniques vary greatly in their computational demands. Brute force computation of some techniques is completely intractable, while a gradient, for example, can be evaluated at the same time as a model prediction.

### 6.3 User testing

An explanation is only useful if people can use it to understand a model and act based on the model's insights. This utility should be verifiable by experiment. For example, modelers should be able to use the explanations to predict which of two given models will generalize better, or use more robust logic in generating predictions from its inputs. A human user should be able to identify irregularities in a model and predict inputs which would be misclassified. If the explanations are accurately describing the model, the user will be able to make these predictions.

## 7 Two experiments show limits to current explainability techniques

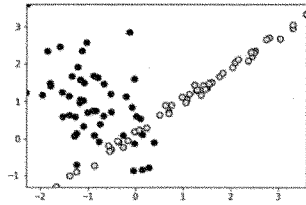
Explainability in machine learning is still very much an emerging field, lacking thorough testing and comparison procedures. In this paper, we perform two experiments to better understand the limits of the ability to explain supervised learning models using current techniques. The first experiment compares three popular explainers on a tree model trained on a toy classification dataset to visualize the mechanics of the algorithms, understand the effect of model hyperparameters, and diagnose potential failure modes. In the second experiment, we evaluate the explanations of a neural network model using a more robust Lending Club loan application dataset.

Experimental evaluation of explainability methods has overwhelmingly focused on vision problems, as saliency maps can be compared by visual inspection. There has been less experimentation on other domain-specific machine learning applications such as credit underwriting and natural language processing. The tabular data of problems like credit underwriting lacks the local structure between neighboring columns that pixels have in image data. Categorical variables behave very differently than continuous variables, which poses additional difficulties.

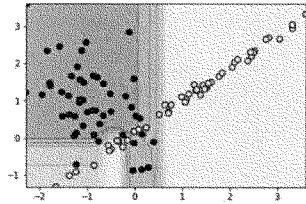
The results of the two experiments in this paper offer insight into the implementation of explainability methods and their performance for credit problems. We develop an extensive set of tests to evaluate the stability and validity of explainers. It is shown that popular explainers require extensive parameter tuning, and may produce unreliable results. Many explainers do not scale well computationally with the number of rows or columns of the dataset and are thus unsuitable for production. The ZAML explainer is evaluated and is shown to satisfy all the desired properties of a reliable explainability technique.

### 7.1 Experiment One: Implementation issues of LIME

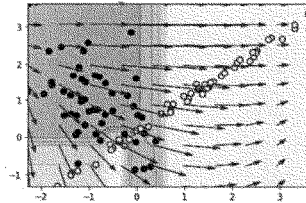
To demonstrate the performance issues of LIME, we begin with a simple 2D classification problem. This toy dataset is shown in the figure below. The purple dots represent class 0 and the yellow dots class 1. The goal of the model is to label each input as belonging to one of the two classes.



An XGBoost classifier model is trained on this data. In the plot below, a heat map of the model output reveals the decision surface of the model.



LIME can be used to understand the importance each feature plays in the model's decision in different regions of the data space. The LIME explanations are computed at evenly spaced grid points, and the resulting explanation is displayed as a red arrow. The horizontal component of the red arrow corresponds to the explanation of the x feature, and the vertical component is the explanation of the y feature.

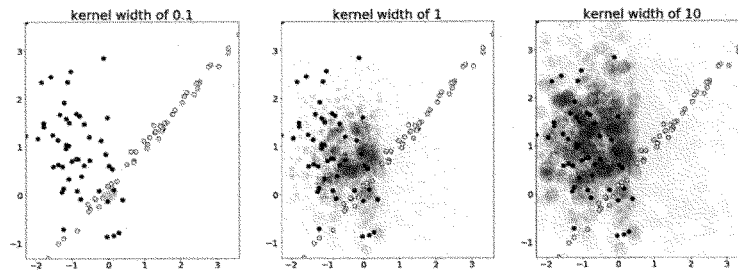


The nature of LIME's explanations depends heavily on its two main parameters. The first parameter is the kernel size, which determines the size of the local neighborhood around a point from which the linear proxy model is built. As the kernel size grows large, LIME approximates the model globally with a single linear model. The second parameter is the number of samples

Copyright ©2018 ZestFinance, Inc. All Rights Reserved. Confidential and proprietary. No part of this document may be disclosed in any manner to any third party without ZestFinance's prior written consent.

that are used in estimating the proxy model. Drawing more samples results in a more stable explanation, but increases the computation time.

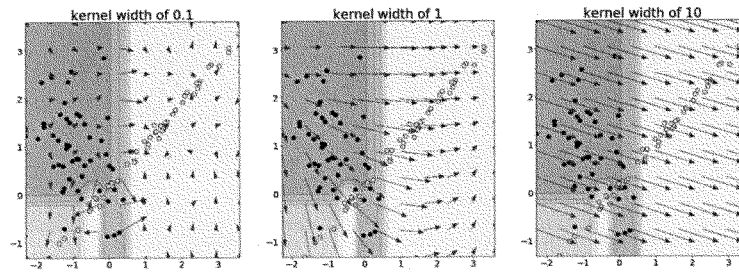
To examine the effect of the kernel size, the figures below capture the neighborhoods LIME would use to explain a prediction at (0,0) given different sizes. The sampled neighborhood points are plotted beneath the data point, colored according to the prediction of the XGBoost model. The size of each neighborhood point is proportional to its weight in the estimation of the linear proxy model. The color represents the score given to each point by the model. Open source implementations of LIME choose a relatively high kernel size, as this ensures that samples from all classes will be present in the neighborhood. However, as can be seen from the figure on the right, if the kernel size is too large, the locality of the proxy model is lost.



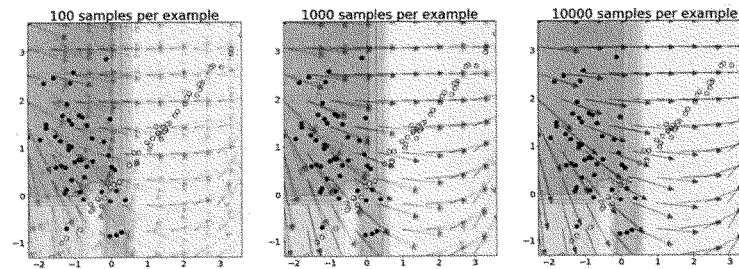
The effects of the kernel size on the resulting LIME explanations (the coefficients of the linear proxy models) are seen in the figure below. The arrow plot may be interpreted as the slope of the linear proxy model at that particular point, thus pointing in the direction of increasing model scores. For the smallest kernel size, the explanations point towards the closest prediction boundary. This provides a more precise description of model behavior, but in sparse regions of the data space, behavior becomes erratic. For samples where  $x > 1$ , the x-coordinate is entirely responsible for the prediction of class 1. The LIME neighborhood is so small that it does not contain a reasonably weighted sample from class 0 this far from the boundary. Thus, when LIME is truly local, the predictions become unreliable for many regions of the data space.

When the kernel size is 1, LIME provides a reasonable explanation of samples near the simplest part of the classification boundary ( $y > 1$ ). The larger neighborhood results in a smooth explanation space, but the finer details in the lower left corner are lost. For an extremely large kernel size of 10, LIME has lost all locality, and each proxy model is the same linear model fit on the unweighted dataset.





The effect of the number of samples that are drawn for the purpose of estimating a proxy model is illustrated in the figure below. With a kernel size of 1, LIME is repeatedly run 10 times, and the resulting explanation arrows are plotted to show the volatility under different numbers of neighborhood samples. With only 100 samples to estimate each proxy model, the explanations vary wildly, to the point where either feature could be considered more important, depending on the draw. As the number of samples increases, the LIME explanations approach determinism. There is a tradeoff between the volatility and computation time of LIME. The proxy model must be learned for each sample that is to be explained. In this simple 2-dimensional problem, 1,000 samples suffice for a reasonable explanation, but as the dimensionality of the data grows, this might become a concern.



One of the more difficult details in implementing LIME is drawing samples from the distribution of the data. Generative models [Goodfellow et al. 2016] address the problem of drawing accurate samples from the distribution of the data but are amongst the most difficult models to train in machine learning. The authors of LIME simply estimate the univariate means and standard deviations for each feature and draw each feature independently from these independent normal approximations. For some types of data, this may be a reasonable

approximation, but this will pose problems, especially with categorical data types, as will be shown in the next section.

From this experiment, it is clear that LIME is very sensitive to its parameters. Even in 2 dimensions, a very large number of generated samples is needed to get stable performance. This will worsen in higher dimensions. The kernel width parameter is easier to estimate for a visualizable dataset, but will also be a much more difficult problem for real datasets. For real data, the Gaussian assumption for the sampling distribution may be unreasonable.

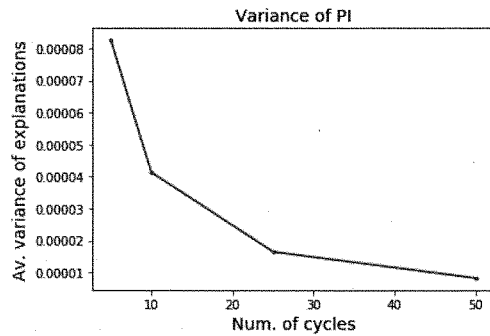
## 7.2 Experiment Two: Explaining performance on real credit data

In our second experiment, we went beyond the limitations presented for toy datasets by attempting to explain a model built on a large, publicly available 2007-2011 data set from the online loan marketplace The Lending Club. The data included 42,538 loan applications and their payment status. Records with a loan status of "Charged Off" or "Fully Paid" were selected, with these statuses serving as the classification target. Feature engineering was deliberately kept simple. Commonly used variables were selected by hand, categorical variables were one-hot encoded, and missing values were imputed with the mean, while adding a binary column to indicate if the value was missing. This resulted in 39,088 input records with 35 dimensions. The first 35,000 records were used for training and the remaining 4,088 were reserved for evaluating explainers. The data was standardized by subtracting the mean of each column and dividing by the standard deviation.

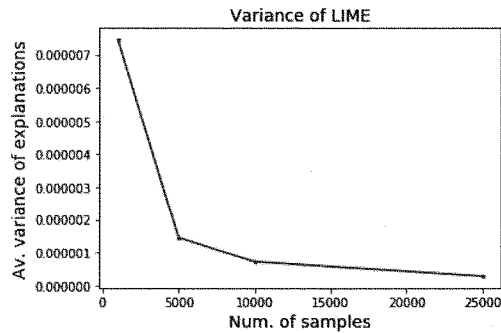
We built a neural network model with 4 dense layers with ReLU activations and dropout regularization. The model had a test set performance of 0.703 AUC. Three explainers were run on this model: permutation importance (PI), LIME, and the ZAML explainer (ZAML), which is based on vector calculus. We evaluate their performance here based on the properties described in the "Evaluating explanations" section of this paper.

### Variance

Permutation importance (PI) and LIME are stochastic algorithms, while the ZAML explainer is deterministic. While some randomness in an explanation is acceptable, an explainer that provides different feature rankings across subsequent runs is undesirable. The randomness in PI originates from the shuffling of the columns to determine model sensitivity to that feature. The number of times each column is shuffled is a parameter called *cycles*, and it controls the tradeoff between determinism of the explanation and computational cost. In the plot below, the explanations from PI are generated for the Lending Club test set, and the variance across 10 runs is computed for each feature level attribution. We plot the average variance over all inputs and columns as a function of the *cycles* parameter. By increasing the cycles, the variance of PI can be greatly reduced at the expense of increased computational cost. Each additional cycle requires  $D$  additional model evaluations, where  $D$  is the dimensionality of the feature space.



Randomness in LIME arises from the neighborhood sampling procedure. As with PI, there is a tradeoff between determinism and computational cost, which is controlled by the number of samples. Estimating a proxy model with more samples has increased cost, but results in a more stable explanation. In the plot below, the average variance of explanations over 10 runs is shown for LIME as a function of the number of samples. For a low number of samples, the variance is very large relative to permutation impact.

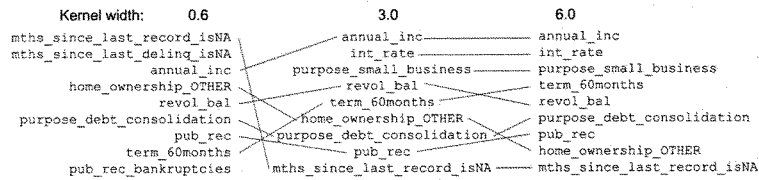


#### Sensitivity to hyperparameters

Aside from the parameters that affect randomness, explainers have additional hyperparameters that control the nature of the explanations. LIME has the kernel width, which dictates the effective neighborhood size around the point of interest. The kernel width greatly affects the

Copyright ©2018 ZestFinance, Inc. All Rights Reserved. Confidential and proprietary. No part of this document may be disclosed in any manner to any third party without ZestFinance's prior written consent.

resulting explanation. In the figure below, the top 9 most important features are shown for increasing kernel widths.



In addition to the kernel width, the sampling procedure itself is a design choice in LIME. Sampling a realistic local neighborhood around a point of interest is a very difficult task and common implementations of LIME model each feature as a univariate Gaussian random variable. This means that LIME explains the prediction of a model based on a neighborhood of samples which may be highly unrealistic.

The first row in the table below is an actual applicant in the Lending Club dataset, followed by 9 rows of application data generated by LIME as the neighborhood. The highly correlated variables of `loan_amnt` (loan amount) and `installment` are no longer correlated. Additionally, variables like `annual_inc` (annual income) may contain negative values. The model was not trained on data like these samples and they do not provide reliable insight about its behavior.

loan_amnt	int_rate	installment	annual_inc	dti	delinq_2yrs
2000.0	0.0662	61.41	50000.00	8.11	0.0
10338.0	0.1071	76.78	-30125.32	21.46	-1.0
17928.0	0.1565	152.43	92680.99	13.69	-0.0
5794.0	0.1322	493.70	79560.33	14.50	0.0
15996.0	0.0395	558.82	79275.55	14.03	0.0
11432.0	0.1101	-125.05	16307.60	18.46	-0.0
13813.0	0.1080	776.08	-36931.54	-1.40	1.0
2498.0	0.0356	426.58	20338.34	5.87	0.0
10847.0	0.1776	229.98	36622.50	20.42	-0.0
-7156.0	0.1369	489.08	27108.77	8.50	-0.0

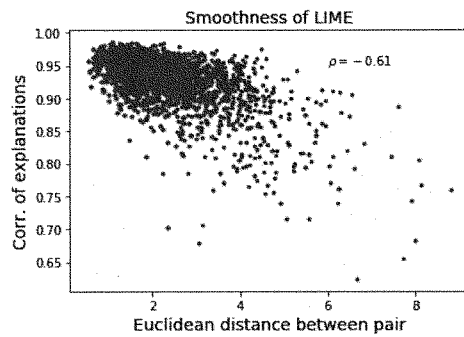
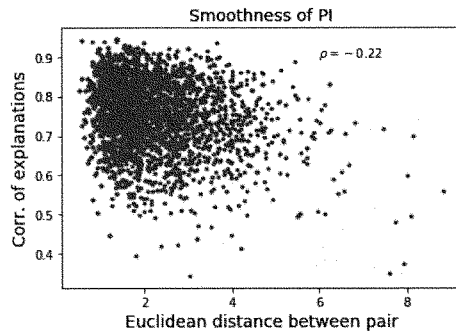
The ZAML explainer involves the computation of a path integral along the manifold of the data and is approximated with discrete steps. The number of steps determines the quality of the approximation. Taking the top 9 most important features for the number of steps set to 2, 10, and 20 reveals that the explanation is not overly sensitive to this parameter.

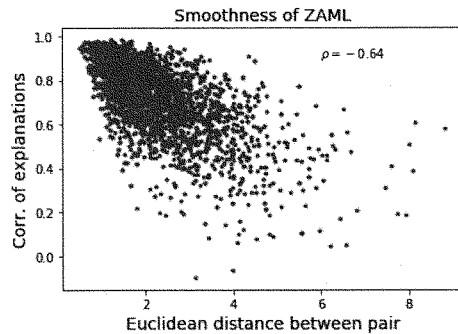
Copyright ©2018 ZestFinance, Inc. All Rights Reserved. Confidential and proprietary. No part of this document may be disclosed in any manner to any third party without ZestFinance's prior written consent.

Num. steps:	2	10	20
	int_rate	int_rate	int_rate
	annual_inc	annual_inc	annual_inc
	purpose_debt_consolidation	purpose_debt_consolidation	purpose_debt_consolidation
	purpose_credit_card	purpose_credit_card	purpose_credit_card
	term_60months	term_60months	term_60months
	term_36months	term_36months	term_36months
	home_ownership_RENT	mths_since_last_delinq_isNA	mths_since_last_delinq_isNA
	inq_last_6mths	home_ownership_RENT	home_ownership_RENT
	mths_since_last_delinq_isNA	inq_last_6mths	inq_last_6mths

### Smoothness

If two inputs to a model are very similar and receive the same scores from the model, then one would expect the explanations of these inputs to be similar as well. This property is known as smoothness. To measure smoothness, we find the nearest neighbor for each test applicant (if it received the same score from the model). For each pair of neighboring inputs, the distance between the two is calculated. Then the explanations of the two samples are compared with Spearman's rank correlation coefficient. For a consistent explainer, if the distance between a pair of inputs is very small, then the explanation will have a Spearman's coefficient near 1.



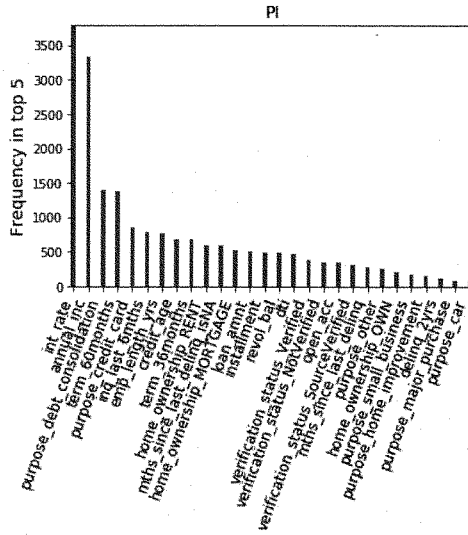


In the plots above, the ZAML explainer displays the strongest negative correlation between the distance of a sample pair and the similarity of their explanations, followed by LIME. For permutation importance, similar samples can result in quite different explanations. From a business perspective, smoothness in the space of explanations provides a sense of fairness in the model's decisions. Discrepancies in model outcomes for two similar scenarios creates dissonance in the human interpretability of the model's functionality.

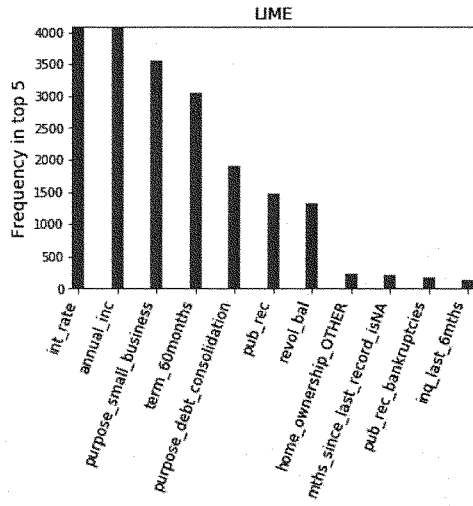
### Precision

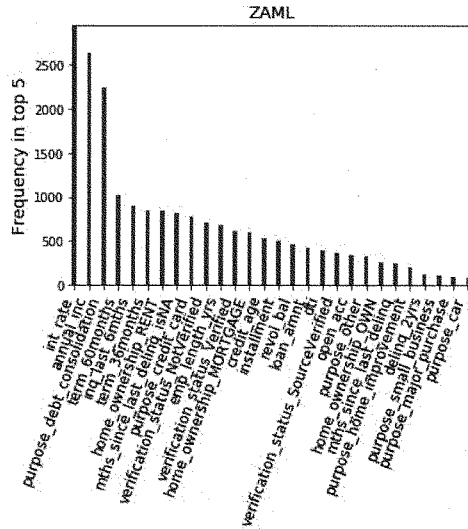
As we mentioned earlier, the precision of an explanation is based on how varied the explanations are across the data space. Modern machine learning models get their predictive power because they can model complex local behavior, extracting signals among different features in different regions of the data space. A global linear proxy model returns the same explanation for every input and is thus not a good candidate for explaining a nonlinear model like the one we built off the Learning Club data. To show the relative precision of the various techniques, we took the top 5 features for each sample in the testing set, and build a histogram of how often each feature appears in the top 5.

These histograms are shown below. ZAML and permutation impact show a few features which almost always are in the top 5. In fact, these methods have significant overlap in their explanations. The frequency of less common features decays gradually. LIME only selects 2 features with significant frequency: the interest rate and the flag indicating if a public record is available. The rest of the features receive very little attribution and essentially fluctuate as noise.









Why does LIME produce such unvaried explanations? The kernel width parameter must be set relatively high in order to keep a sufficient number of samples in the local neighborhood from which the linear proxy model is estimated. If the neighborhood size is too small, particularly with higher dimensional data, the linear model estimation will be ill-posed. Thus, a larger kernel width is favored, and the proxy model no longer accurately represents the neural network, but rather a highly smoothed version of it with no complex local behavior.

**Accuracy**

The accuracy of an explanation is especially important in the credit domain as the Fair Credit Reporting Act requires that all credit denials come with reasons that accurately describe the key factors that led to an applicant being denied. The adverse action notice provided to the customer must describe key factors a consumer can change in order to improve their likelihood of being approved. Accurate reasons are therefore required to comply with the law in the United States and many other jurisdictions.

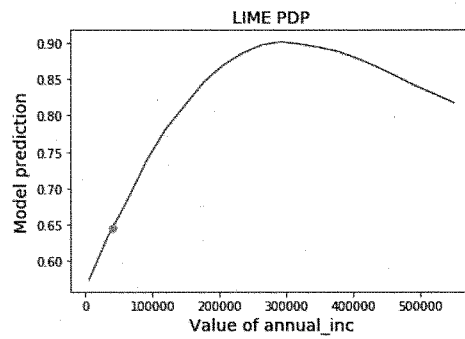
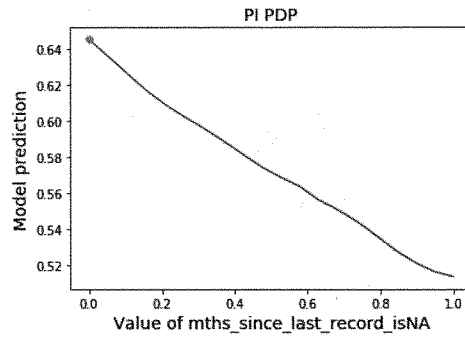
To evaluate the accuracy of an explainer, consider the top feature for each loan applicant. We replicated the input, but replaced the top feature with every possible value in its range. The model was evaluated with these synthetic inputs to determine the maximum possible change in

Copyright ©2018 ZestFinance, Inc. All Rights Reserved. Confidential and proprietary. No part of this document may be disclosed in any manner to any third party without ZestFinance's prior written consent.

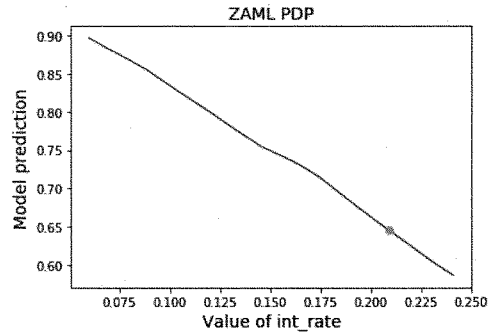
---

score from the original input. If this value was small, then the top feature was not really influential in the model's decision.

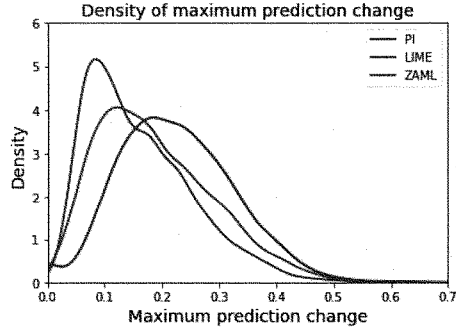
For a single input, this can be visually interpreted from a partial dependence plot (PDP). The x-axis represents the value of the feature of interest, and the y-axis represents the resulting model prediction. In the plots below, the impacts of the top features predicted by the 3 explainers are shown for a single input. In this case, all explainers produce a meaningful explanation.



Copyright ©2018 ZestFinance, Inc. All Rights Reserved. Confidential and proprietary. No part of this document may be disclosed in any manner to any third party without ZestFinance's prior written consent.



To compare the accuracies of the 3 explainers, we show the distributions of the maximum impacts of the top features for the test set. Permutation importance produces a top feature with the highest impact, which is a reasonable outcome considering this test is essentially what the permutation importance uses to produce its explanation. The ZAML explainer, while built from completely different principles, produces reasonably impactful top features. The top feature of LIME is not impactful. The LIME technique fails to identify the most impactful variables.



---

## 8 Conclusion

The techniques used today to explain linear models are not safe to use for machine learning, nor are many of the leading methods created to explain machine learning models. For explaining a neural network model, like the one we built using Lending Club data, LIME was not a reliable explainability technique. Sampling a local neighborhood becomes increasingly difficult as the number of features grows. Setting hyperparameters for LIME is unfortunately too much of an art given the sensitivity of the results to its parameters and the lack of ground truth. It is possible that for many problems, a single neighborhood size parameter will not work globally. The randomness in the algorithm can only be mitigated by drawing a large number of neighborhood samples, which increases computational costs.

Permutation importance and ZAML both present reasonable explanations, which is supported by the fact that their top features can be manipulated to greatly affect the model prediction. Both methods have significant overlap in the features they identify as important, which is additional evidence in the correctness of their results. The diversity of top 5 features shows that they both capture local behavior, rather than a single global explanation. ZAML provides a more consistent explanation in the sense that similar inputs receive similar explanations. It is also a deterministic algorithm, which is an important property, especially in credit, where fluctuations between two runs of the algorithm may be hard to defend.

Computational cost is an important property of explainers for systems that must run in near real-time. The cost of LIME is essentially fitting a linear model for each input that is to be explained, a task that quickly becomes intractable when a model has even just hundreds of variables. The cost of least-squares regression depends on the number of samples in the local neighborhood and the dimensionality of the data.

Permutation importance and the ZAML explanation can be compared more directly. For a dataset with  $D$  features, permutation importance requires  $D \cdot \text{cycles}$  model evaluations to explain each input. The ZAML explanation requires model evaluations equal to the `number_of_steps` parameter. Experimentally, we observed stable explanations for permutation importance with about 50 cycles, and for ZAML with 10 steps. In a dataset with 100 features for example, PI will take about 500 times longer than ZAML, and this gap only increases with the dimensionality of the data.

## 9 Explainability At Zest

Zest has developed over the last few years wholly new model explainability methods that provide accurate and repeatable explanations at a row, segment, and global level for ensembles of heterogeneous ML models. The explainability is achieved through methods that cover differentiable and non-differentiable models. We have solved for numerical precision issues and

Copyright ©2018 ZestFinance, Inc. All Rights Reserved. Confidential and proprietary. No part of this document may be disclosed in any manner to any third party without ZestFinance's prior written consent.

---

provided methods for combining explanations of diverse submodels. This capability, when paired with ZAML's analysis, monitoring, and automated documentation enables lenders to safely and quickly apply advanced ML models in credit underwriting and in other regulated applications where consistency, accuracy, and performance is paramount.

## 10 References

1. Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning." arXiv preprint arXiv:1702.08608 (2017).
2. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)
3. Breiman, Leo. "Random forests." *Machine learning* 45.1 (2001): 5-32.
4. Friedman, Jerome H. "Greedy function approximation: a gradient boosting machine." *Annals of statistics* (2001): 1189-1232.
5. Goldstein, Alex, et al. "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation." *Journal of Computational and Graphical Statistics* 24.1 (2015): 44-65.
6. Lei, Jing, et al. "Distribution-free predictive inference for regression." *Journal of the American Statistical Association* (2018): 1-18.
7. Buciluă, Cristian, Rich Caruana, and Alexandru Niculescu-Mizil. "Model compression." *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2006.
8. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should i trust you?: Explaining the predictions of any classifier." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2016.
9. Shapley, Lloyd S. "A value for n-person games." *Contributions to the Theory of Games* 2.28 (1953): 307-317.
10. Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Advances in Neural Information Processing Systems*. 2017.
11. Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." arXiv preprint arXiv:1312.6034 (2013).
12. Kindermans, Pieter-Jan, et al. "PatternNet and PatternLRP—improving the interpretability of neural networks." *stat* 1050 (2017): 16.
13. Pirie, W. "Spearman rank correlation coefficient." *Encyclopedia of statistical sciences* 12 (2004).
14. Goodfellow, Ian, et al. *Deep learning*. Vol. 1. Cambridge: MIT press, 2016.

### 10.1 Internal References & Code

<https://github.com/Katlean/Experimental/blob/master/users/mfe/experiments/simulations/Demo%20Explainer%20Issues.ipynb>

Copyright ©2018 ZestFinance, Inc. All Rights Reserved. Confidential and proprietary. No part of this document may be disclosed in any manner to any third party without ZestFinance's prior written consent.





## **Why Lenders Shouldn't 'Just Use SHAP' To Explain Machine Learning Credit Models**

Everyone wants to solve AI's black box problem: the dilemma of understanding how a machine learning (ML) computer model arrives at its decisions. The hard part is figuring out the influence of each of the hundreds or thousands of variables interacting in nearly infinite combinations to derive an outcome in an ML model.

In 2017 two computer scientists from the University of Washington published a technique for generating fast and practical explanations of a particular kind of ML called tree-based models (specifically, a variant called XGBoost). The algorithm's authors named their work SHAP, for [Shapley additive explanations](#), and it's been used hundreds of times for coding projects.

The Shapley name refers to American economist and Nobelist Lloyd Shapley, who in 1953 first published his formulas for assigning credit to "players" in a multi-dimensional game where no player acts alone. Shapley's seminal game theory work has influenced voting systems, college admissions, and scouting in professional sports. Shapley Values work well in machine learning, too. The catch is that they're expensive to compute. In a game or model with just 50 variables you're already looking at considering more options than there are stars in the universe.

That's where SHAP comes in. SHAP approximates Shapley values quickly by cleverly using the tree structure of XGBoost models, speeding up the explanation time enough to make it practical to assign credit to each variable. Some banks and lenders eager to use machine learning in credit underwriting or other models are asking themselves, "Why not just use SHAP to power my explanation requirements?"

Fair question. The answer? Because that would be irresponsible for a bunch of reasons. For credit and finance applications, bridging from off-the-shelf SHAP to a safe application takes a lot of care and work, even if you just want to explain XGBoost models. Credit risk models must be treated particularly carefully because they highly regulated and significantly impact consumers' lives. When a consumer is denied credit, the [Fair Credit Reporting Act of 1970](#) requires accurate and actionable reasons for the decision so that consumers can repair their credit and re-apply successfully.

SHAP is a practical solution for some use cases of ML, but in credit underwriting, it just doesn't hold water on its own. Here are a few reasons why we've faced serious challenges in our attempts to apply SHAP in credit risk -- and why we had to invent something new.

**Score space vs margin space - these details really matter**

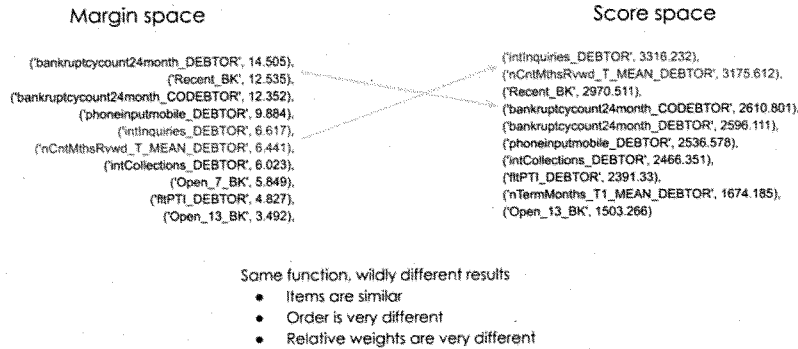
Lending businesses want to be able to set a target and approve, say, 20% of applicants. That means the business wants a function that outputs numbers between 0 and 100, where 0 is the worst and 100 is the best, and for which exactly 20% of all scores lie above 80, 30% of all scores lie above 70%, and so on. This well-defined output is said to be in "score space."

The score space is very different mathematically from the credit model's actual output, which is said to be in "margin space." Margin space numbers fall in a narrow range from 0 to 1. In general, the relationship between the model's actual output in margin space and the acceptance threshold in score space is extremely non-linear, and you have to transform the model's output to generate the number the lending business wants. Don't worry, you're not the only one that struggles to keep track: we do too, and while technical, the margin space/score space transition really matters.

The problem with SHAP is that, because of the way it computes its Shapley values, it really only works in margin space. If you compute the set of weighted key factors in margin space, you'll get a very different set of factors and weights than if you compute them in score space, which is where banks derive their top five explanations for rejecting a borrower. Even if you are using the same populations and are only looking at the transformed values, you will not get the same importance weights. Worse, you likely won't end up with the same factors in the top five.

The table below shows how this plays out for a real applicant for an auto loan. The reasons returned to the rejected borrower were dramatically different when translated from margin space to score space. If you skipped this important step, and just used SHAP out of the box, you would have thought the main reason for denial was the bankruptcy count. But the real top reason for denial, in score space, was the number of credit inquiries. A consumer relying on reasons generated by margin space attribution would be misled. Getting this wrong could have devastating consequences to consumers seeking to access financing for their first house or car, who rely on denial reasons to improve their ability to access credit. It could also cause a lender to run afoul of fair lending and fair credit rules.

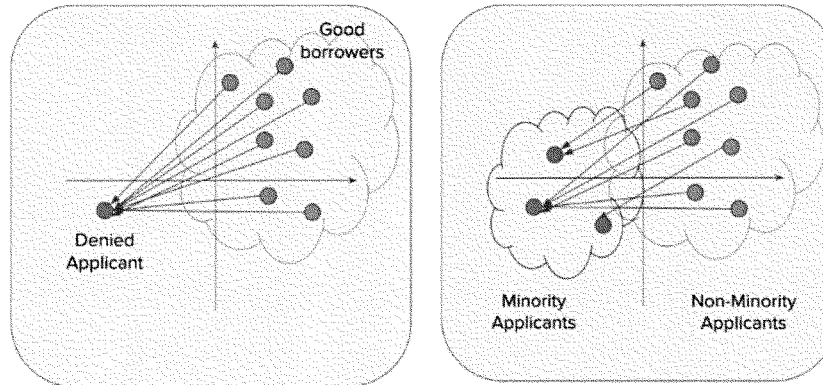
## Assigning credit in margin space can be deceptive



Why does this happen? Because SHAP derives its values by looking at all the results of taking a path down each tree in the model, and it assumes that the sum of the values along a set of paths down a tree gives you the score -- basically, you can compute the score with only the data in the trees. That's not true when you transform into score space; the transformation destroys that structure. SHAP can also have trouble recovering even a simple model's internal structure, as we'll explain in the last point.

### Explanation by reference

SHAP computes variable importance globally, which means it shows how the model behaves for every applicant (in margin space) with respect to the overall model itself. In credit risk modeling, it is often required to understand an applicant's score in terms of another applicant or applicant population, that is, with respect to a reference population. For example, when lenders compute the reasons an applicant was rejected (for [adverse action notices](#)), they want to explain the applicant's score in terms of the approved applicants. When they do [disparate impact analysis](#) lenders want to understand the drivers of approval rate disparity. This requires comparing the feature importance for the population of, say, white non-Hispanic male applicants to protected groups and performing a search for less discriminatory alternatives. These are illustrated in the diagrams below.



Left: Adverse action requires comparing the denied applicant to good borrowers.

Right: Fair lending analysis requires comparing minority applicants with non-minority applicants.

There are many details you need to get right in this process, including the appropriate application of sample weights, mapping to score space at the approval cut-off, sampling methods, and accompanying documentation. Out of the box, SHAP doesn't allow you to easily do this.

#### You want to use modeling methods other than XGBoost

Using SHAP is hard enough because it outputs values in margin space that you have to correctly map into score space. But it has other important limitations. Although SHAP provides fast explanations for gradient-boosted tree models, there are many other mechanisms for building scoring functions, including many alternative forms of tree models such as random forests and extremely random forests, not to mention other implementations of gradient boosting such as [LightGBM](#).

You may also want to use continuous modeling methods such as radial basis function networks, Gaussian mixture models, and, perhaps most commonly, deep neural networks. The current implementation of SHAP cannot explain any of the other types of tree models, and cannot explain any continuous model outside a small collection, and only by importing algorithms other than SHAP.

What's more, SHAP cannot explain ensembles of continuous and tree-based models, such as stacked or deeply stacked models that combine xgboost and deep neural networks. In our experience ([and the experience of others](#)), these types of ensemble models are more accurate and stable over time. That's why we built [ZAML](#) to explain a

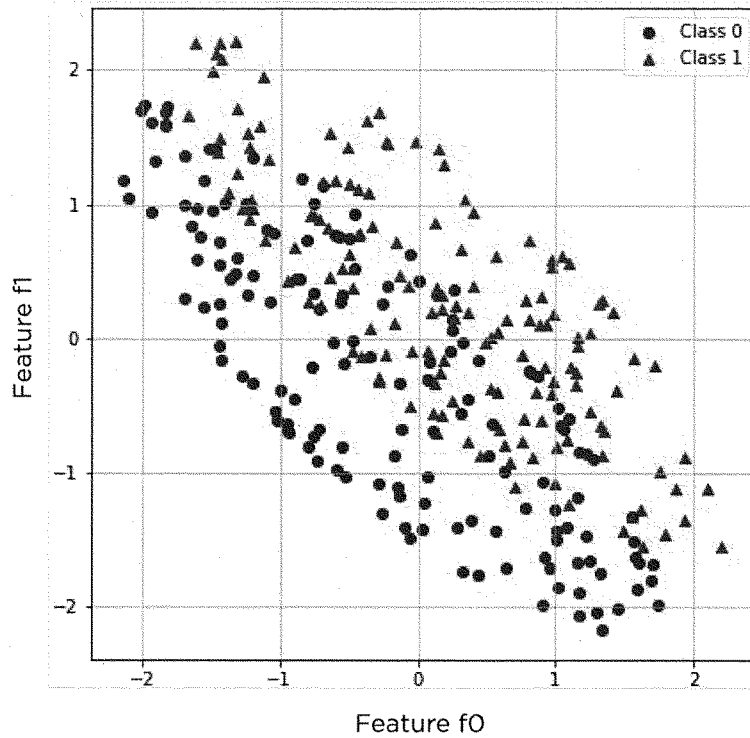
Copyright © 2019 ZestFinance, Inc. All Rights Reserved. Confidential and proprietary. No part of this document may be disclosed in any manner to any third party without ZestFinance's prior written consent.

much wider variety of model types, enabling you to use world-beating ensemble models to drive your lending business.

**Even on a simple XGBoost model, SHAP fails to uncover the underlying geometry**

Machine learning models are effectively geometric entities: they embody the idea that things near to one another will tend to be mapped to the same place and then produce systems which reflect that structure. A good example of this is the ovals dataset, a two-dimensional dataset consisting of a set of points drawn uniformly from two overlapping ovals with the same number of points drawn from each. The ovals from which the points are drawn are arranged roughly vertically in the chart below, and the model is trained to predict membership in one or the other oval given the coordinates of a point. For convenience, the oval with a greater y value is arbitrarily assigned the target value 1 and the oval with the less y value is assigned the value 0.

Actual Classification

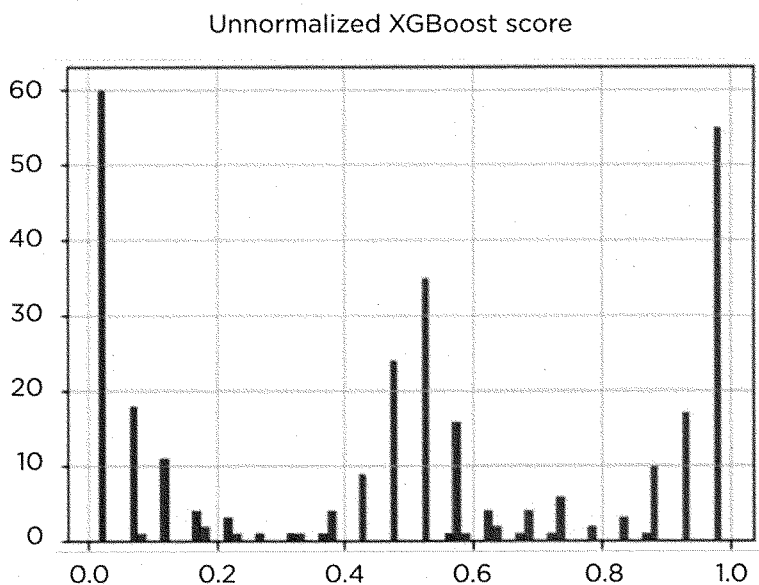


When viewed geometrically, this dataset is inseparable: points in the overlapping region are equally likely to have been drawn from either of the two ovals and so no classifier can predict membership for any such point.

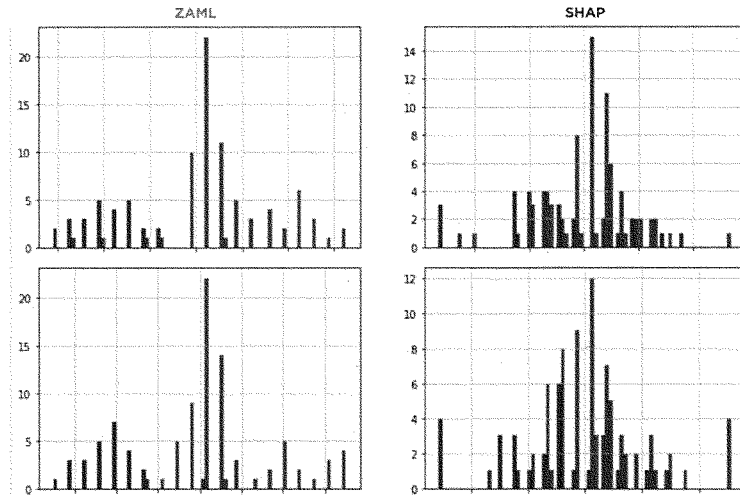
Intuitively, one would expect a classification function defined for the ovals dataset to correspond to three regions: a region of points belonging only to the upper oval, a region of points common to the two ovals, and a region of points belonging only to the lower oval. We trained an XGBoost model on a random sample of half of the ovals dataset, and looked at the model's predictions on the other half.

The chart below shows the model's predictions, and we can see the three regions we expected. The blue represents scores the model assigned to the bottom region, the

green the middle, and the red, the top. As you can see the model produces nicely separated outputs.



We should see that same separation when we look at the explainer outputs. One would intuitively expect that items in the upper region will have average attributions which are relatively large and positive, items in the common area to have attributions which are relatively close to zero, and items in the lower region to have average attributions which are relatively large and negative. If the explainer doesn't reflect this structure, it isn't really explaining the model, and probably shouldn't be trusted. To investigate that question, we compared SHAP attribution weights with the attribution weights generated by Zest's ZAML software in the charts below.



Let's walk through what we're seeing. The left column shows the feature importance for each model prediction, as assigned by ZAML. The right column shows the feature importance for each model prediction as generated by SHAP. The top row is the feature importance for f0, the x coordinate in our ovals dataset. The bottom row is the feature importance for f1, the y coordinate. The blue, green and red colors correspond to the bottom, middle and top regions, respectively.

As you can see, ZAML readily separates the top, middle, and bottom regions -- notice how the blue green and red bars are all nicely separated in the charts in the left column -- while SHAP, shown on the right, gets them all jumbled up. The results suggest that SHAP may not be the right tool to use off the shelf for the rigorous and regulated requirements of credit underwriting.

We did not expect these results when we first saw them, and frankly we thought they were wrong. After careful review by multiple teams inside and outside the company, however, we're confident they're not. Look for a scientific paper describing our algorithm, a mathematical proof of its correctness and uniqueness, and other empirical results to be published soon. In the meantime, if you care about getting your model explanations right, [feel free to reach out to us](#).

SHAP was a giant leap forward in model explainability. The use of a game-theoretic framework to explain models is powerful and creative. Nonetheless, as the above analyses show, you really need more than just the out-of-the-box SHAP to provide the kind of accurate explanations required for real-world credit decisioning applications.

Copyright © 2019 ZestFinance, Inc. All Rights Reserved. Confidential and proprietary. No part of this document may be disclosed in any manner to any third party without ZestFinance's prior written consent.



Even on a simple XGBoost model, SHAP can provide inaccurate explanations, and care must be taken to map into score space correctly and to mitigate numerical precision issues, when computing explanations by reference. Before diving head first into ML explainability with SHAP, it is important to understand its limitations and determine whether or how you will address those limitations in your ML application. Credit decisions make lasting impacts on people's lives and getting the explanations right matters.



## **FAQs ZestFinance has received on the Application of the Federal Agencies' Banking Supervisory Guidance on Model Risk Management to Machine Learning Models**

**NOTE TO READERS:** ZestFinance, Inc. helps lenders transition from conventional underwriting methods to machine learning-based underwriting. In the course of our work, we often receive questions from executives at financial institutions about the applicability of the federal banking agencies' Supervisory Guidance on Model Risk Management (the Guidance) to machine learning models.<sup>1</sup> The Guidance clearly applies to machine learning-based underwriting models. Machine learning models, however, differ from traditional models in ways that raise unique issues regarding their evaluation, testing, and documentation under the Guidance. The FAQs below reflect the questions Zest receives most frequently on this issue. The accompanying answers set forth Zest's current views on best practices for the responsible adoption of machine learning models consistent with the Guidance, as well as the goals of ensuring safety and soundness in the financial system, increasing access to credit, and minimizing fair lending and other compliance risk.

### **SUMMARY**

The financial services industry is increasing its adoption of machine learning (ML) for a range of applications. ML models are powerful at predicting outcomes because they can consider more data than traditional models and apply sophisticated mathematical techniques to evaluate multiple variables and the relationships between them, and continually refine and improve their underlying algorithms to enhance performance and predictive power on an ongoing basis. ML technologies have the potential to bring unbanked and underbanked consumers into the financial system, enhance access to responsible credit, and contribute positively to the overall safety and soundness of the financial system. Increased predictive power, however, comes with increased

---

<sup>1</sup> The Guidance was issued by the Board of Governors of the Federal Reserve System (FRB) and the Office of the Comptroller of the Currency (OCC) in 2011, and adopted by the Federal Deposit Insurance Corporation (FDIC), with technical conforming changes, in 2017.

complexity. The use of these advanced new modeling techniques in financial services raises important issues of risk management.

The Guidance establishes a framework for effective model risk management that focuses on appropriate development, documentation, validation, and governance standards for models used by financial institutions. The guidance applies to all modern modeling techniques, including machine learning models. However, the Guidance was adopted in 2011, largely before ML was common, and thus does not address ML models. While the principles articulated in the Guidance remain sound and appropriate for all models, certain particulars and examples in the Guidance do not reflect the way ML models function. Different validation approaches, built largely upon current approaches, are more effective at meeting the Guidance's goals when using ML models.

These FAQs cover the application of the Guidance to the use of ML models by financial institutions and describes methods for complying with key aspects articulated in the Guidance given the unique risks posed by ML techniques. The questions below do not address all aspects of the Guidance; instead, they are the questions Zest is most frequently asked by executives at financial institutions considering the use of machine learning. The accompanying answers represent Zest's current thinking on how financial institutions may use ML models responsibly and consistent with the Guidance.

### **SECTION III: OVERVIEW OF MODEL RISK MANAGEMENT**

*For new machine learning models, can lenders use techniques for model risk management different from those outlined in the Guidance that they determine to be more appropriate for such models?*

- Yes. The Guidance is not prescriptive, but illustrative. The selection of model risk management techniques should be based, in part, on the type, complexity, and functional attributes of the model. ML models operate differently than traditional models; thus, it is appropriate to consider alternative approaches to model risk management.

*Is it acceptable to use machine learning models in high stakes financial services decision-making?*

- Yes. Nothing in the Guidance precludes the use of machine learning models. The Guidance applies to a financial institution's use of any "model," which it defines as a "quantitative method, system, or approach that applies statistical, economic, financial, or mathematical theories, techniques, and assumptions to process

input data into quantitative estimates." ML models fit squarely within this definition.

- Fairness, anti-discrimination, and safety and soundness goals tend to support the use of more predictive models, including machine learning models. As many as fifty million Americans have incomplete or inaccurate credit bureau data. Millions of these consumers are denied access to credit by lenders using conventional, static credit scoring techniques because those models often inaccurately predict default risk. Machine learning models are resilient to incomplete data, able to consider more variables, and capable of creating models that more accurately assess credit risk. Consequently, ML's enhanced predictive power has the potential to safely expand access to credit while reducing losses and systemic risk.
- However, ML-based credit risk models must be validated, documented, and monitored using methods appropriate to the modeling approach selected in order to comply with the principles articulated in the Guidance. As discussed below, conventional validation approaches are not sufficient to evaluate ML models. ML model developers and institutions should take care to conform their practices to the principles in the Guidance regarding Model Development Implementation and Use, and Model Evaluation and Verification standards using techniques robust enough to assess and explain the performance of ML models.

#### **SECTION IV: MODEL DEVELOPMENT, IMPLEMENTATION, AND USE**

*Can you use as many variables as desired in a model?*

- Yes. The Guidance does not address, or limit, the number of variables that may be used in a model, and nothing in the Guidance suggests that fewer variables necessarily decreases risk. ML models can consider many more variables than traditional methods, which is a key reason why ML models often provide greater predictive power, and deliver superior results, compared to traditional models.
- The same data review and documentation practices outlined in the Guidance still apply to ML models even though ML models consider many more variables than traditional models. As the Guidance indicates, "there should be rigorous assessment of data quality and relevance, and appropriate documentation. Developers should be able to demonstrate that such data and information are suitable for the model."

*Can model developers analyze vastly more variables and still comply with the Guidance?*

- Yes. As the Guidance states: “Developers should be able to demonstrate that such data and information are suitable for the model and that they are consistent with the theory behind the approach and with the chosen methodology. If data proxies are used, they should be carefully identified, justified, and documented.”
- ML models consider hundreds or even thousands of variables, so it may be impractical to manually review all of them. An automated variable review may be the most effective way to support comprehensive analysis and documentation of the data and the model. Automated variable review methods should identify and document data issues that could raise questions about the predictive power, fairness, and safety and soundness of a model. Notably, variables should be reviewed for unexpected and/or inconsistent distributions, mappings, and other data degradation issues that can lead to model misbehavior. In connection with reviewing data variables, ML models will detect patterns and relationships among variables that no human would detect. This continuously evolving multivariate analysis is what makes any assessment of the data during the development phase problematic. The Guidance calls for documentation of these review methods and descriptions of the assumptions and theoretical basis for their use.

#### **SECTION V: MODEL VALIDATION**

*What methods are permissible for assessing the soundness of an ML model?*

- The Guidance does not prescribe any specific method for validating any model, including a machine learning model. Nonetheless, the Guidance sets out a core framework for effective model validation: evaluation of conceptual soundness, ongoing monitoring, and outcomes analysis.
- Regarding soundness, certain conventional evaluation methods described in the Guidance would, if applied to ML models, be ineffective and would likely produce misleading results. For example, one of the testing methods identified by the Guidance is sensitivity analysis. Common implementations of sensitivity analysis include exploring all combinations of inputs and permuting these inputs one-by-one (univariate permutation) in order to understand the influence of each variable (or a combination thereof) on model scores. Exploring all combinations of inputs (exhaustive search) is computationally infeasible for most ML models. Univariate permutation (permuting inputs one-by-one), while more computationally tractable, yields incorrect results for ML models that capture and evaluate multivariate interactions.
- Effective ML model evaluation techniques should be efficient and tractable, and designed to test the how ML models actually work. Such techniques should also

assess the impact of multivariate interactions because ML models evaluate such interactions. Appropriate methods of evaluating ML models include techniques derived from game theory, multivariate calculus, and probabilistic simulation.

*How do the Guidance's monitoring standards apply to ML models?*

- The Guidance calls for ongoing model monitoring: "Such monitoring confirms that the model is appropriately implemented and is being used and is performing as intended..." The Guidance further states: "Many of the tests employed as part of model development should be included in ongoing monitoring and be conducted on a regular basis to incorporate additional information as it becomes available."
- A thorough approach for monitoring ML models should include:
  - Input distribution monitoring: Recent model input data may be compared with model training data to determine whether incoming credit applications are significantly different from model training data. The more that live data differs from training data, the less accurate the model is likely to be. This data comparison is typically done by looking at variable distributions and ensuring recent data is drawn from a similar distribution as occurred in the model training data. For ML models, multivariate input variable distributions should be monitored to identify input data where combinations of values that were unlikely to appear together during model development are now occurring in production. Systems for monitoring model inputs should trigger alerts to monitors or validators when they spot anomalies or shifts that exceed pre-defined safe bounds.
  - Missing input data monitoring: Comprehensive model monitoring should include monitoring for missing input data. Model input data comes from a variety of sources, some of which is retrieved over networks from third parties. Data sources could become unavailable in production. A complete model monitoring program should monitor and trigger alerts to monitors and validators when the rate of missing data, and its impact on model outputs and downstream business outcomes, exceed pre-defined thresholds.
  - Output distribution monitoring: Model outputs should be monitored by comparing distributions of model scores over time. Monitoring systems should compute statistics that establish the degree to which the score distribution has shifted from the scores generated by the model in prior periods such as those contained in training and validation data sets.

- Execution failure monitoring: Error and warning alerts generated during model execution can indicate flaws in model code that may affect model outputs. Such alerts should, therefore, be closely monitored, the causes of such alerts should be investigated and identified, and appropriate remediation should be implemented where necessary.
- Latency monitoring: Model response times should be monitored to ensure model execution code and infrastructure meet the latency requirements of applications and workflows that rely on model outputs. Models that perform slowly or with unreliable execution time may cause intermittent timing issues, which can result in the generation of inaccurate scores. Establishing clear latency objectives and pre-defined alert thresholds should be part of a comprehensive model monitoring management program.
- Economic performance monitoring: A complete ML model monitoring solution should include business dashboards that enable analysts to configure or pre-define alert triggers on key performance indicators such as default rate, approval rate, and volumes. Substantial changes in these indicators can signal operational issues with model execution and, at a minimum, should be investigated and understood in order to manage risk.
- Reason code stability: Reason codes explain the key drivers of a model's score. Reason code distributions should be monitored because material changes to the distributions can indicate a change in the character of the applicant population or even in the decision-making logic of the ML model.
- Fair lending analysis: Machine learning models can develop unintended biases for a variety of reasons. Relatedly, like any model, ML models can result in disparities between protected classes. To ensure that all applicants are treated fairly and in a non-discriminatory manner, it is important to monitor loan approvals, declines, and default rates across protected classes. Historically, this monitoring has been done far after the fact. Because of the possibility of bias and the advanced predictive fit of ML models, monitoring of these models should occur in real time.

*Should model monitoring include automation?*

- Yes. The Guidance states: "monitoring should continue periodically over time, with a frequency appropriate to the nature of the model." Given the complexity of ML models, automated model monitoring, which can run concurrently with model operations, is essential to meet the expectations set by the Guidance, especially when combined with multivariate input monitoring and alerts. Changes to input

and output distributions should be monitored in real time to identify problems promptly and reflected in periodic reports.

*How should model outcomes be analyzed?*

- As the Guidance recommends, model outcomes should be thoroughly understood prior to adoption and deployment of any new model, including ML models. Because machine learning models can consider many more data points than traditional models, traditional tools such as manual review of partial dependence plots can be cumbersome or inaccurate. Such tools can also miss crucial aspects of ML model behavior, such as the influence of variable interactions. In addition to understanding fully how a model arrives at a score, it is important to understand the swap sets generated by switching to a new model: that is, which applicants will now be approved (swap-ins) and which will now be denied (swap-outs). While the quantity of applicants is important, so is the quality of applicants. Outcomes validation methods should include an examination of the distribution of values for all model attributes of swap-ins and swap-outs, as well as a comparison with populations already accepted and with known credit performance.

## **SECTION VI: GOVERNANCE, POLICIES, AND CONTROLS**

*How do the Guidance documentation requirements apply to ML models?*

- As the Guidance states, “documentation of model development and validation should be sufficiently detailed so that parties unfamiliar with a model can understand how the model operates, its limitations, and its key assumptions.”
- Meeting the requirement for thorough documentation of advanced modeling techniques can be challenging for model developers because ML models can process many more variables than traditional models, ML algorithms often have many tunable parameters, ML “ensembles” can join both many variables and many tunable parameters, and all of these must be thoroughly documented so the model can be reproduced.
- These issues largely do not apply to logistic regression-based underwriting models, which are easier to understand and explain but less predictive.
- In the case of ML models, documenting how a model operates, its limitations, and its key assumptions requires using explainability techniques that accurately reveal how the model reached its decisions and why.



- Entities should ensure that they use explainability methods that accurately explain how a model operates. Most commonly used explainability methods are unable to provide accurate explanations. For example, some methods (e.g., LOCO, LIME, PI) look only at model inputs and outputs, as opposed to the internal structure of a model. Probing the model only externally in this way is an imperfect process leading to potential mistakes and inaccuracies. Similarly, methods that analyze refitted and/or proxy models (e.g., LOCO and LIME), as opposed to the actual final model, result in limited accuracy. Explainability methods that use “drop one” or permutation impact methods (e.g., LOCO and PI) rely on univariate analysis, which fails to properly capture feature interactions and correlation effects. Finally, methods that rely on subjective judgement (e.g., LIME) create explanations that are both difficult to reproduce and overly reliant on the initial judgement. These errors in explanation cause model accuracy to suffer. Even slight inaccuracies in explanations can lead to models that discriminate against protected classes, are unstable, and/or produce high default rates. Models that rely on mathematical analyses of the underlying model itself, including high-order interactions, and do not need subjective judgement are appropriate explainability methods.

*Should model documentation include automation?*

- Yes. Although the Guidance is silent on whether model documentation may be generated automatically, automated model documentation is the most practical solution for ML models. ML model development is complex, and operationalizing and monitoring ML models is even harder. It is not feasible for a human, unaided, to keep track of all that was done to ensure proper model development, testing and validation. There are tools to automate model documentation for review by model developers, compliance teams, and other stakeholders in the model risk governance process. Given the number of variables in ML models, automated documentation is likely to provide a higher degree of accuracy and completeness than manual documentation. In general, participants in model risk management should not rely upon manually generated documentation for ML models.

*Are there other best practices for ML model risk management?*

- Yes. The Guidance makes clear that the quality of a bank’s model development, testing, and validation process turns in large part on “the extent and clarity of documentation.” Therefore, model documentation should be clear, comprehensive, and complete so that others can quickly and accurately revise or reproduce the model and verification steps. Documentation should explain the business rationale for adopting a model and enable validation of its regulatory compliance.

- Records of model development decisions and data artifacts should be kept together so that a model may be more easily adjusted, recalibrated, or redeveloped when conditions change. Such artifacts include development data, data transformation code, modeling notebooks, source code and development files, the final model code, model verification testing code, and documentation.
- Model documentation should be clear, comprehensive, and complete so that others can quickly and accurately revise or reproduce the model and verification steps. Documentation should explain the business rationale for adopting a model and enable validation of its regulatory compliance.

**TESTIMONY OF NICOL TURNER LEE**  
**Fellow, Center for Technology Innovation, Brookings Institution**

**Before the**  
**Task Force on Artificial Intelligence**  
**United States House Committee on Financial Services**

**Hearing on “Perspectives on Artificial Intelligence: Where We Are and the Next Frontier in Financial Services”**

**June 26, 2019**

Chairwoman Waters, Ranking Member McHenry and Members of the Committee, thank you for the opportunity to testify. I am encouraged by the interest of this committee on artificial intelligence (AI) and the application of autonomous systems to the financial services sector. I am Nicol Turner Lee, Fellow in the Center for Technology Innovation at the Brookings Institution. With a history of over 100 years, Brookings is committed to evidenced-based, nonpartisan research in a range of focus areas. My particular research expertise encompasses data collection and analysis around regulatory and legislative policies that govern telecommunications and high-tech companies, along with the impacts of digital exclusion, artificial intelligence and machine-learning algorithms on vulnerable consumers. My forthcoming book, *The digitally invisible: How the internet is creating the new underclass*, also addresses this topic.

**Introduction**

Increasingly, the private and public sectors are turning to artificial intelligence (AI) systems and machine learning algorithms to automate simple and complex decision-making processes. The mass-scale digitization of data and the emerging technologies that use them are disrupting most economic sectors, including transportation, retail, advertising, financial services and energy, and other areas. AI is also having an impact on democracy and governance as computerized systems are being deployed to improve accuracy and drive objectivity in government functions.

It is the availability of massive data sets which has made it easy to derive new insights through computers. As a result, *machine learning algorithms*, which are a set of step-by-step instructions that computers follow to perform a task, have become more sophisticated and pervasive tools for automated decision-making.<sup>1</sup> While algorithms are used in many contexts from making recommendations about movies to credit products, I rely on a definition that I made recently in a newly released paper<sup>2</sup> which refers to them as computer models that make inferences from data about people, including their identities, their demographic attributes, their preferences, and their likely future behaviors, as well as the objects related to them.<sup>3</sup>

In machine learning, algorithms rely on multiple data sets, or training data, that specifies what the correct outputs are for some people or objects. From that training data, it then learns a model which can be applied to other people or objects and make predictions about what the correct outputs should be for them.<sup>4</sup> However, because machines can treat similarly-situated people and objects differently, research is starting to reveal some troubling examples in which the reality of algorithmic decision-making falls short of our expectations, or is simply wrong. For example, automated risk assessments used by U.S. judges to determine bail and sentencing limits can generate incorrect conclusions, resulting

---

<sup>1</sup> The concepts of AI, algorithms and machine learning are often conflated and used interchangeably. In this paper, we will follow generally understood definitions of these terms as set out in publications for the general reader. See, e.g., Stephen F. DeAngelus. "Artificial intelligence: How algorithms make systems smart," *Wired Magazine*, September 2014. Available at <https://www.wired.com/insights/2014/09/artificial-intelligence-algorithms-2/> (last accessed June 25, 2019). See also, Michael J. Garbade. "Clearing the Confusion: AI vs. Machine Learning vs. Deep Learning Differences," *Towards Data Science*, September 14, 2018. Available at <https://towardsdatascience//clearing-the-confusion-ai-vs-machine-learning-vs-deep-learning-differences-fce69b21d5eb> (last accessed April 12, 2019).

<sup>2</sup> Turner Lee, N., Resnick, P., and Barton, G. Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. Brookings. (2019). Available at <https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/> (last accessed June 25, 2019). Angwin, J. Tobin, A. Varner, M. (2017). Facebook (Still) Letting Housing Advertisers Exclude Users by Race. *ProPublica*. Available at: <https://goo.gl/Vk4irs> (last accessed June 25, 2019).

<sup>3</sup> Andrea Blass and Yuri Gurevich. Algorithms: A Quest for Absolute Definitions. *Bulletin of European Association for Theoretical Computer Science* 81, 2003. <https://www.microsoft.com/en-us/research/wp-content/uploads/2017/01/164.pdf> (last accessed June 25, 2019).

<sup>4</sup> Technically, this describes what is called "supervised machine learning."

in large cumulative effects on certain groups, like longer prison sentences or higher bails imposed on people of color. Or, credit decisions based on inferential data about applicants, such as their zip code, social media profiles or web browsing histories, can lead to higher rejection rates.

Referring back to my recent paper on this subject, my co-authors and I determine that an algorithmic decision generates “bias” when its outcomes are systematically less favorable to individuals within a particular group and where there is no relevant difference between groups that justifies such harms.<sup>5</sup> Bias in algorithms can come from unrepresentative or incomplete training data, or the reliance on flawed information that reflects historical inequalities. The bottom line is that if left unchecked, biased algorithms can lead to decisions which can have a collective, disparate impact on certain groups of people even without the programmer’s intention to discriminate.

In my testimony, I hope to further unpack the concept of algorithmic bias and outline why we need to proactively work to identify and mitigate online biases. I conclude this written testimony with a series of recommendations – whether driven by policymakers or the self-regulatory actions of industries – that can facilitate more ethical, fair and just algorithmic models. If not carefully identified and mitigated, algorithms – especially those associated with the sensitive use cases to be discussed in this hearing – have the potential to replicate and amplify stereotypes historically prescribed to people of color and other vulnerable populations.

#### **Racial and Ethnic Biases in the Online Economy**

I’d like to start with an *initial truth* about emerging technologies. Despite their facilitation of greater efficiencies and cognition due to the programming of machines, the online economy has not resolved

---

<sup>5</sup> Blog, “Understanding bias in algorithmic design,” Impact.Engineered, September 5, 2017. Available at <https://medium.com/impact-engineered/understanding-bias-in-algorithmic-design-db9847103b6e> (last accessed June 25, 2019). This definition is intended to include the concepts of disparate treatment and disparate impact, but the legal definitions were not designed with AI in mind. For example, the demonstration of disparate treatment does not describe the ways in which an algorithm can learn to treat similarly situated groups differently, as will be discussed later in the paper.

the issue of racial bias in its applications. In 2013, online search results for “black-sounding” names were more likely to link arrest records with profiles, even when false.<sup>6</sup> Two years later, Google apologized for an algorithm that automatically tagged and labeled two African Americans as “gorillas” after an innocuous online word search.<sup>7</sup> In 2017, a report by ProPublica exposed a controversial online function on Facebook that allowed advertisers to exclude members of its “ethnic affinity” groups, primarily people of color, from targeted marketing for certain ads.<sup>8</sup> Those ads were specifically focused on housing, employment, and the extension of credit.

In their controversies, Google explained their biases as problems associated with the algorithm or the inappropriate meta-tagging of images. Facebook immediately ended their practices and forbade advertisers from engaging in discriminatory practices on their site. In both cases, certain online users were wrongly characterized based upon their race.

We live in a society where online data are collected in real-time from users through a series of interactions with web sites, social media communities, e-commerce vehicles, and general online inquiries for information of interest. These small portions of data become compiled, mined, and eventually regenerated for commercial or public use. Big data serves a variety of purposes, from helping to advance breakthroughs in science, health care, energy, and transportation to enhancing government efficiencies by aggregating citizen input.

Big data can also exclude people. In a report published by the Federal Trade Commission (FTC), when big data analytics are misapplied, online users can be tracked or profiled based on their online

---

<sup>6</sup> BBC. (2013). Google searches expose racial bias, says study of names. *BBC*. Available at: <https://goo.gl/P8oodF> (last accessed June 25, 2019).

<sup>7</sup> Kasperkevic, J. (2015). Google says sorry for racist auto-tag in photo app. *The Guardian*. Available at: <https://goo.gl/ZEJYng> (last accessed June 25, 2019).

<sup>8</sup> Angwin, J. Tobin, A. Varner, M. (2017). Facebook (Still) Letting Housing Advertisers Exclude Users by Race. *ProPublica*. Available at: <https://goo.gl/Vk4irs> (last accessed June 25, 2019).

activities and behaviors.<sup>9</sup> Consequently, online users can be denied credit based on their web browsing history, or aggregated, predictive analytics can wrongly determine an individual's suitability for future employment or an educational opportunity. Online proxies, including one's zip code, can also be used by marketers to extrapolate an individual's socioeconomic status based on neighborhood, resulting in incorrect assumptions about one's lifestyle or preferences.<sup>10</sup> In these and other examples, big data, when misapplied, can lead to the disparate treatment of individuals and groups, especially those that comprise protected classes by race, gender, age, ability, religion, and sexual orientation.

In these cases, the algorithm - when applied to these vulnerable populations - may repeat historical discrimination, or generate new forms of bias, whether explicit, implicit, or unconscious. In the instances of *explicit bias*, algorithms may not start out being discriminatory or have prejudicial intent. Instead, the algorithm can adapt to the societal biases that exist within communities of online users, leading to stereotypes and unfair profiling. Latanya Sweeney, Harvard researcher and former chief technology officer at the Federal Trade Commission (FTC), found the micro-targeting of higher-interest credit cards and other financial products when the computer inferred that the subjects were African-Americans, despite having similar backgrounds to whites.<sup>11</sup> During a public presentation at a FTC hearing on big data, Sweeney demonstrated how a web site, which marketed the centennial celebration of an all-black fraternity, received continuous ad suggestions for purchasing "arrest records" or accepting high-interest credit card offerings.<sup>12</sup>

---

<sup>9</sup> Ramirez, E. Brill, J. Ohlhausen, K. McSweeney, T. (2016). Big Data: A Tool for Inclusion or Exclusion. *FTC*. Available at: <https://goo.gl/wUxwU1> (last accessed June 25, 2019).

<sup>10</sup> Noyes, K. (2015). Will big data help end discrimination—or make it worse? *Fortune*. Available at: <https://goo.gl/VnPM1j> (last accessed June 25, 2019).

<sup>11</sup> Sweeney, Latanya and Jinyan Zang. "How appropriate might big data analytics decisions be when placing ads?" Powerpoint presentation presented at the Big Data: A tool for inclusion or exclusion, Federal Trade Commission conference, Washington, DC. September 15, 2014. Available at [https://www.ftc.gov/systems/files/documents/public\\_events/313371/bigdata-slides-sweeneyzang-9\\_15\\_14.pdf](https://www.ftc.gov/systems/files/documents/public_events/313371/bigdata-slides-sweeneyzang-9_15_14.pdf) (last accessed June 25, 2019).

<sup>12</sup> "FTC Hearing #7: The Competition and Consumer Protection Issues of Algorithms, Artificial Intelligence, and Predictive Analytics," § Federal Trade Commission (2018),

When the values and beliefs of the programmer factors into the design of the algorithm, there is a risk of *implicit or unconscious biases*. Here, implicit bias can extend into the complex calculations of machine learning and artificial intelligence concealed within the design of the algorithmic procedure. Online retailer Amazon, whose global workforce is 60 percent male and where men hold 74 percent of the company's managerial positions, recently discontinued use of a recruiting algorithm after discovering gender bias.<sup>13</sup> The data that engineers used to create the algorithm were derived from the resumes submitted to Amazon over a 10-year period, which were predominantly from white males. The algorithm was taught to recognize word patterns in the resumes, rather than relevant skill sets, and these data were benchmarked against the company's predominantly male engineering department to determine an applicant's fit. As a result, the AI software penalized any resume that contained the word "women's" in the text and downgraded the resumes of women who attended women's colleges, resulting in gender bias.<sup>14</sup>

MIT researcher Joy Buolamwini found that the algorithms powering three commercially available facial recognition software systems were failing to recognize darker-skinned complexions.<sup>15</sup> Generally, most facial recognition training data sets are estimated to be more than 75 percent male and more than 80 percent white. When the person in the photo was a white man, the software was accurate 99 percent of the time at identifying the person as male. According to Buolamwini's research, the product error rates for the three products were less than one percent overall, but increased to more than 20

---

<sup>13</sup> Hamilton, Isobel Asher. "Why It's Totally Unsurprising That Amazon's Recruitment AI Was Biased against Women." Business Insider, October 13, 2018. Available at <https://www.businessinsider.com/amazon-ai-biased-against-women-no-surprise-sandra-wachter-2018-10> (last accessed June 25, 2019).

<sup>14</sup> Vincent, James. "Amazon Reportedly Scraps Internal AI Recruiting Tool That Was Biased against Women." The Verge, October 10, 2018. Available at <https://www.theverge.com/2018/10/10/17958784/ai-recruiting-tool-bias-amazon-report> (last accessed April 20, 2019). Although Amazon scrubbed the data of the particular references that appeared to discriminate against female candidates, there was no guarantee that the algorithm could not find other ways to sort and rank male candidates higher so it was scrapped by the company.

<sup>15</sup> Hardesty, Larry. "Study Finds Gender and Skin-Type Bias in Commercial Artificial-Intelligence Systems." MIT News, February 11, 2018. Available at <http://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212> (last accessed June 25, 2019). These companies were selected because they provided gender classification features in their software and the code was publicly available for testing.



percent in one product and 34 percent in the other two in the identification of darker-skinned women as female.<sup>16</sup> In response to Buolamwini's facial-analysis findings, both IBM and Microsoft committed to improving the accuracy of their recognition software for darker-skinned faces. Not surprising, Buolamwini is an African-American female researcher, suggesting that implicit and unconscious biases can often go undetected in high-tech industries where diverse populations are clearly underrepresented.

Generally, these examples of explicit, implicit and unconscious biases, complicated by historical realities, unmask the fact that algorithms are not necessarily devoid of societal biases, prejudices, stereotypes, and even incorrect assumptions.

#### **Causes of algorithmic biases**

Before delving into the specific use cases impacting the financial services sector, it is imperative to understand the root causes of online biases. *First*, historical human biases are shaped by pervasive and often deeply embedded prejudices against certain groups, which can lead to their reproduction and amplification in computer models. In the Amazon recruitment algorithm, men were the benchmark for professional "fit," resulting in female applicants and their attributes being downgraded. Unfortunately, historical realities often find their way into the algorithm's development and execution, and they are exacerbated by the lack of diversity which exists within the computer and data science fields.<sup>17</sup>

*Second*, online biases can also be reinforced and perpetuated without the user's knowledge. For example, African-Americans who are primarily the target for high-interest credit card options might find themselves clicking on this type of ad without realizing that they will continue to receive such predatory online suggestions. In this and other cases, the algorithm may never accumulate counter-factual ad

---

<sup>16</sup> Ibid.

<sup>17</sup> Turner Lee, Nicol. "Inclusion in Tech: How Diversity Benefits All Americans," § Subcommittee on Consumer Protection and Commerce, United States House Committee on Energy and Commerce (2019). Also available on Brookings web site, <https://www.brookings.edu/testimonies/inclusion-in-tech-how-diversity-benefits-all-americans/> (last accessed June 25, 2019).

suggestions (e.g., lower-interest credit options) that the consumer could be eligible for and prefer. Thus, it is important for algorithm designers and operators to watch for such potential negative feedback loops that cause an algorithm to become increasingly biased over time.

*Third*, Insufficient training data is another cause of algorithmic bias, particularly if the data used to train the algorithm are more representative of some groups of people than others. In this case, the predictions from the model may be systematically worse for unrepresented or under-representative groups. For example, in Buolamwini's facial-analysis experiments, the poor recognition of darker-skinned faces was largely due to their statistical under-representation in the training data. Conversely, algorithms with too much data, or an over-representation, can skew the decision toward a particular result. Researchers at Georgetown Law School found that an estimated 117 million American adults are in facial recognition networks used by law enforcement, and that African-Americans were more likely to be singled out primarily because of their *over-representation* in mug-shot databases.<sup>18</sup> Consequently, African-American faces had more opportunities to be falsely matched, which produced a biased effect.

#### **Managing bias detection and mitigating out biases**

In our recent paper, the co-authors and I argue that detection approaches should begin with careful handling of the sensitive information of users, including data that identify a person's membership in a federally protected group (e.g., race, gender). Moreover, developers and other entities that are tasked in the design and deployment of algorithms must address and guard against the systemic bias waged on protected classes, especially when it leased to collective *disparate impacts*. Some of these outcomes may have a basis for legally cognizable harms, such as the denial of credit, online racial profiling, or mass surveillance.<sup>19</sup> While other cases may be justification for action simply due to the outputs of the

---

<sup>18</sup> Sydell, Laura. "It Ain't Me, Babe: Researchers Find Flaws In Police Facial Recognition Technology." NPR.org, October 25, 2016. Available at <https://www.npr.org/sections/alltechconsidered/2016/10/25/499176469/it-aint-me-babe-researchers-find-flaws-in-police-facial-recognition> (last accessed June 25, 2019).

<sup>19</sup> Guerin, Lisa. "Disparate Impact Discrimination." www.nolo.com. Available at <https://www.nolo.com/legal-encyclopedia/disparate-impact-discrimination.htm> (last accessed June 25, 2019). See also, *Jewel v. NSA* where the

algorithm, which may produce *unequal outcomes* or unequal error rates for different groups, despite not having an intent to discriminate, e.g., the mislabeling African-Americans as primates.

The argument could also be made that algorithms cannot be blind to sensitive attributes, despite all of the efforts of the developers.<sup>20</sup> Critics have pointed out that an algorithm may classify information based on online proxies for the sensitive attributes, yielding a bias against a group even without making decisions directly based on one's membership in that group. Barocas and Selbst define online proxies as "factors used in the scoring process of an algorithm which are mere stand-ins for protected groups, such as zip code as proxies for race, or height and weight as proxies for gender."<sup>21</sup> They argue that proxies often linked to algorithms can produce both errors and discriminatory outcomes, such as instances where a zip code is used to determine digital lending decisions or one's race triggers a disparate outcome.<sup>22</sup> Similarly, a job-matching algorithm may not receive the gender field as an input, but it may produce different match scores for two resumes that differ only in the substitution of the name "Mary" for "Mark" because the algorithm is trained to make these distinctions over time.

Going forward, operators of algorithms must be more transparent in their handling of sensitive information, especially if the potential proxy could itself be a legal classificatory harm.<sup>23</sup>

#### **Addressing algorithmic bias in the financial services sector**

For years, research has made clear that there are historical and contemporary inequalities that exist in the financial services industries. In the area of banking services for historically disadvantaged populations, the 2017 Federal Deposit Insurance Corporation's National Survey of Unbanked and

---

Electronic Frontier Foundation argues that massive (or dragnet) surveillance is illegal. Information about case available at <https://www EFF.org/cases/Iewel> (last accessed June 25, 2019).

<sup>20</sup> This is often called an anti-classification criterion that the algorithm cannot classify based on membership in the protected or sensitive classes.

<sup>21</sup> Zarsky, Tal. "Understanding Discrimination in the Scored Society." SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, January 15, 2015. <https://papers.ssrn.com/abstract=2550248>.

<sup>22</sup> Larson, Jeff, Surya Mattu, and Julia Angwin. "Unintended Consequences of Geographic Targeting." Technology Science, September 1, 2015. Available at <https://techscience.org/a/2015090103/> (last accessed June 25, 2019).

<sup>23</sup> Solon Barocas and Andrew D. Selbst, "Big Data's Disparate Impact," SSRN Scholarly Paper (Rochester, NY: Social Science Research Network, 2016. Available at <https://papers.ssrn.com/abstract=2477899>.

Underbanked Households reported that 17 percent of African-Americans and 14 percent of Hispanics were completely unbanked, compared to three percent of whites.<sup>24</sup> A further 30 percent of African-Americans and 29 percent of Hispanics were underbanked, compared to 14 percent of whites.

When coupled with other demographics, the disparities appear more glaring. Fifteen percent of unmarried female-headed family households are unbanked, as are 22 percent of American households without a high school diploma, and 24 percent of households where Spanish is the predominant language. As this evidence suggests, these populations are poorly served by the banking system as it currently operates.

African-Americans and non-White Hispanics are also poorly represented in homeownership. For example, Philadelphia has perhaps one of the most glaring displays of redlining, a practice which persuades and dissuades individuals toward change. Despite being part of 44 percent of the state's population, African-Americans received 10 times fewer mortgage loans than their white counterparts.

Despite a strengthening economy, record low unemployment and higher wages for whites, African-American homeownership has decreased every year since 2004 while all other groups have made gains. In 2017, 19.3 percent of African American applicants were denied home loans, while only 7.9 percent of white applicants were rejected. Brookings fellow Andre Perry found that "owner-occupied homes in black neighborhoods are undervalued by \$48,000 per home on average, amounting to \$156 billion in cumulative losses."<sup>25</sup> In other words, for every \$100 in white family wealth, black families hold just \$5.04. This type of physical redlining is now manifesting in the form of applications discrimination, or

---

<sup>24</sup> "FDIC National Survey of Unbanked and Underbanked Households." Federal Deposit Insurance Corporation. (2017). Available at [economicinclusion.gov/downloads/2017\\_FDIC\\_Unbanked\\_HH\\_Survey\\_Appendix.pdf](https://economicinclusion.gov/downloads/2017_FDIC_Unbanked_HH_Survey_Appendix.pdf). (last accessed June 25, 2019).

<sup>25</sup> Perry, A., Rothwell, J., and Harshbarger, D. The devaluation of black assets in black neighborhoods: the case of residential property. Brookings. (2018). Available at <https://www.brookings.edu/research/devaluation-of-assets-in-black-neighborhoods/> (last accessed June 25, 2019).

what Frank Pasquale has coined as “weblining,” where whole communities are classified by their credit characteristics and associated risks.

In his paper on credit denial in the age of AI, Brookings scholar Aaron Klein argues that AI and machine learning algorithms can begin to find “empirical relationships between new factors and consumer behaviors.”<sup>26</sup> In fact, he asserts that one’s social media profile, the type of computer one is using, what a person is wearing and where they buy their clothes could potentially factor into a credit model, denying loans to individuals whose choices and preferences suggest their inability to re-pay a loan. These new deployments of credit-algorithms are challenging legally cognizable harms and make it more difficult for consumers to discern the reasons for deniability.

Brookings scholar Henry-Nickie presents a similar argument in pointing to how the over-reliance on AI-driven financial services can create “wicked problems” when bank and fintech algorithms choose which consumers to serve.<sup>27</sup> In particular, the range of problems created by less thoughtful AI implementation can encompass: product steering, discriminatory pricing, unfair credit rationing, exclusionary filtering, and digital redlining.<sup>28</sup>

The historical and contemporary realities of certain populations when it comes to wealth- and asset-building suggest that more work needs to be done to avert a potential “double” and “triple” jeopardy of potential exclusion from the burgeoning online economy, particularly when algorithmic decision-making models are baked with assumptions about certain groups.

#### **Recommendations**

---

<sup>26</sup> Klein, A. Credit denial in the age of AI. Brookings. (2019). Available at <https://www.brookings.edu/research/credit-denial-in-the-age-of-ai/> (last accessed June 25, 2019).

<sup>27</sup> Henry-Nickie, M. How artificial intelligence affects financial consumers. Brookings. (2019). Available at <https://www.brookings.edu/research/how-artificial-intelligence-affects-financial-consumers/> (last accessed June 25, 2019).

<sup>28</sup> Ibid.

While this committee is embarking on both an educational pathway and serious legislative dialogue on the application of AI to financial services, I outline in my final section of this written testimony a set of high-level recommendations for consideration among Members of the committee and Congress as a whole.

**1. Congress must modernize civil rights laws and other consumer protections to safeguard protected classes from online discrimination.**

To develop trust from policymakers, computer programmers, businesses, and other operators of algorithms must abide by U.S. laws and statutes that currently forbid discrimination in public spaces. Historically, nondiscrimination laws and statutes unambiguously define the thresholds and parameters for the disparate treatment of protected classes. The 1964 Civil Rights Act “forbade discrimination on the basis of sex as well as race in hiring, promoting, and firing.” The 1968 Fair Housing Act prohibits discrimination in the sale, rental, and financing of dwellings, and in other housing-related transactions to federally protected classes. Enacted in 1974, the Equal Credit Opportunity Act stops any creditor from discriminating against any applicant from any type of credit transaction based on protected characteristics. While these laws do not necessarily mitigate and resolve other implicit or unconscious biases that can be baked into algorithms, companies and other operators should guard against violating these statutory guardrails in the design of algorithms, as well as mitigating their implicit concern to prevent past discrimination from continuing.

To quell algorithmic bias, Congress should start by clarifying how these nondiscrimination laws apply to the types of grievances recently found in the digital space, since most of these laws were written before the advent of the internet.<sup>29</sup> Such legislative action can provide clearer guardrails that are triggered when algorithms are contributing to legally recognizable harms. Moreover, when creators and

---

<sup>29</sup> Tobin, Ariana. “HUD sues Facebook over housing discrimination and says the company’s algorithms have made the problem worse.” ProPublica (March 28, 2019). Available at <https://www.propublica.org/article/hud-sues-facebook-housing-discrimination-advertising-algorithms> (last accessed June 25, 2019).

operators of algorithms understand that these may be more or less non-negotiable factors, the technical design will be more thoughtful in moving away from models that may trigger and exacerbate explicit discrimination, such as design frames that exclude rather than include certain inputs or are not checked for bias.<sup>30</sup>

Henry-Nickie also points to the importance of maintaining consumer financial protections in the age of AI, which implore regulators to engage in targeted, strategic and analytical exploration of emerging technologies in the sector. Further, she argues that “the deluge of data generated by connected devices and machine learning applications creates a prime opportunity to collect and mine publicly available data to inform critical regulation burden analyses,” for which she references the Small Business Regulatory Enforcement Fairness Act and the Paperwork Reduction Act.<sup>31</sup>

In the end, it is important for Congress to determine what role, if any, they want to play in prescribing some level of accountability to companies developing and disseminating algorithms going forward. It may be the case that without accountability or further conversation between policymakers, technologists and civil society, this conversation will be for naught.

**2. Companies that design and deploy algorithms must exercise some level of algorithmic accountability, which involves the creation of a bias impact statement, regular auditing and more human involvement in risk-adverse decisions, like credit and lending.**

As a self-regulatory practice, a *bias impact statement* can help probe and avert any potential biases that are baked into or are resultant from the algorithmic decision. As a best practice, operators of algorithms should brainstorm a core set of initial assumptions about the algorithm’s purpose prior to its development and execution. The bias impact statement should assess the algorithm’s purpose, process

<sup>30</sup> Elejalde-Ruiz, Alexia. “The end of the resume? Hiring is in the midst of technological revolution with algorithms, chatbots.” Chicago Tribune (July 19, 2018). Available at <http://www.chicagotribune.com/business/ct-biz-artificial-intelligence-hiring-20180719-story.html>.

<sup>31</sup> Henry-Nickie, M. How artificial intelligence affects financial consumers. Brookings. (2019). Available at <https://www.brookings.edu/research/how-artificial-intelligence-affects-financial-consumers/> (last accessed June 25, 2019).

and production, where appropriate. Operators of algorithms should also consider the role of diversity within their work teams, training data, and the level of cultural sensitivity within their decision-making processes. Employing diversity in the design of algorithms upfront will trigger and potentially avoid harmful discriminatory effects on certain protected groups, especially racial and ethnic minorities. While the immediate consequences of biases in these areas may be small, the sheer quantity of digital interactions and inferences can amount to a new form of systemic bias. Therefore, the operators of algorithms should not discount the possibility or prevalence of bias and should seek to have a diverse workforce developing the algorithm, integrate inclusive spaces within their products, or employ “diversity-in-design,” where deliberate and transparent actions will be taken to ensure that cultural biases and stereotypes are addressed upfront and appropriately. Adding inclusivity into the algorithm’s design can potentially vet the cultural inclusivity and sensitivity of the algorithms for various groups and help companies avoid what can be litigious and embarrassing algorithmic outcomes.

The bias impact statement should not be an exhaustive tool. As a self-regulatory tool, its goal should be to ward off disparate impacts resulting from the algorithm that border on unethical, unfair, and unjust decision-making. When the process of identifying and forecasting the purpose of the algorithm is achieved, a robust feedback loop will aid in the detection of bias, which leads to the next recommendation of promoting regular audits of algorithms and their decisions. Where appropriate, more humans should be involved in these processes to ensure that subjective criteria are not dominating the final outcome.

Congress should promote, and in some cases reward self-regulatory models where businesses identify, monitor and correct biases that negatively impact the online experiences of users. For example, Google’s decision to ban ads that promoted payday loans was an example of self-regulation. Or, Facebook’s updates to its ad policies to prevent race-based targeting, especially those that attempt to include or exclude demographic groups in housing, employment and credit, is another example of how



companies are correcting ill-advised practices. A potentially novel idea may be to reward best practices with some type of “gold seal of approval” when companies demonstrate a strict adherence to standards and practices which highlights their outperformance in creating more ethical algorithms.

**3. Congress should support the use of regulatory sandboxes and safe harbors to curb online biases.**

Regulatory sandboxes could be another policy strategy for the creation of temporary reprieves from regulation to allow the technology and rules surrounding its use to evolve together. These policies could apply to algorithmic bias and other areas where the technology in question has no analog covered by existing regulations. Rather than broaden the scope of existing regulations or create rules in anticipation of potential harms, a sandbox allows for innovation both in technology and its regulation. Even in a highly regulated industry, the creation of sandboxes where innovations can be tested alongside with lighter touch regulations can yield benefits.

For example, companies within the financial sector that are leveraging technology, or fintech, have shown how regulatory sandboxes can spur innovation in the development of new products and services.<sup>32</sup> These companies make extensive use of algorithms for everything from spotting fraud to deciding to extend credit. Some of these activities mirror those of regular banks, and those would still fall under existing rules, but new ways of approaching tasks would be allowed within the sandbox.<sup>33</sup> Because sandboxes give innovators greater leeway in developing new products and services, they will require active oversight until technology and regulations mature. The U.S. Treasury recently reported not only on the benefits that countries that have adopted fintech regulatory sandboxes have realized,

---

<sup>32</sup> Fintech regulatory sandboxes in [UK](#), [Singapore](#), and [states in the U.S.](#) are beginning to authorize them. They allow freedom to offer new financial products and use [new technologies such as blockchain](#).

<sup>33</sup> In March, the state of Arizona became [the first U.S. state to create a “regulatory sandbox” for fintech companies](#), allowing them to test financial products on customers with lighter regulations. The U.K. has run a similar initiative called [Project Innovate](#) since 2014. The application of a sandbox can allow both startup companies and incumbent banks to experiment with more innovative products without worrying about how to reconcile them with existing rules.

but recommended that the U.S. adopt fintech sandboxes to spur innovation.<sup>34</sup> Given the broad usefulness of algorithms to spur innovation in various regulated industries, participants in the roundtables considered the potential usefulness of extending regulatory sandboxes to other areas where algorithms can help to spur innovations.

Regulatory safe harbors could also be employed, where a regulator could specify which activities do not violate existing regulations.<sup>35</sup> This approach has the advantage of increasing regulatory certainty for algorithm developers and operators. For example, Section 230 of the Communications Decency Act removed liability from websites for the actions of their users, a provision widely credited with the growth of internet companies like Facebook and Google. The exemption later narrowed to exclude sex trafficking with the passage of the Stop Enabling Online Sex Trafficking Act and Fight Online Sex Trafficking Act. Applying a similar approach to algorithms could exempt their operators from liabilities in certain contexts while still upholding protections in others where harms are easier to identify. In line with the previous discussion on the use of certain protected attributes, safe harbors could be considered in instances where the collection of sensitive personal information is used for the specific purposes of anti-bias detection and mitigation.

**4. The tech sector must be more deliberate and systematic in the recruitment, hiring and retention of diverse talent to avert and address the mishaps generated by online discrimination, especially algorithmic bias.**

Less diverse workforces contribute to algorithmic bias, whether intentional or not. Recent diversity statistics report these companies employ less than two percent of African Americans in senior executive

<sup>34</sup> Mnuchin, Steven T., and Craig S. Phillips. "A Financial System That Creates Economic Opportunities - Nonbank Financials, Fintech, and Innovation." Washington, D.C.: U.S. Department of the Treasury, July 2018. Available at [https://home.treasury.gov/sites/default/files/2018-08/A-Financial-System-that-Creates-Economic-Opportunities---Nonbank-Financials-Fintech-and-Innovation\\_0.pdf](https://home.treasury.gov/sites/default/files/2018-08/A-Financial-System-that-Creates-Economic-Opportunities---Nonbank-Financials-Fintech-and-Innovation_0.pdf) (last accessed June 25, 2019).

<sup>35</sup> Another major tech-related Safe Harbor is the EU-US Privacy Shield after the previous Safe Harbor was declared invalid in the EU. Available at [https://en.wikipedia.org/wiki/EU%E2%80%93US\\_Privacy\\_Shield](https://en.wikipedia.org/wiki/EU%E2%80%93US_Privacy_Shield) (last accessed June 25, 2019).

positions, and three percent of Hispanics when compared to 83 percent of whites.<sup>36</sup> Asian-Americans comprise just 11 percent of executives in high tech companies.<sup>37</sup> In the occupations of computer programmers, software developers, database administrators, and even data scientists, African-Americans and Hispanics collectively are under six percent of the total workforce, while whites make up 68 percent.<sup>38</sup> Even when people of color are employed in high tech industries, the feelings of professional and social isolation also have been shown to marginalize these employees, potentially restricting their active workplace engagement, affecting their participation in the feedback loop, and contributing to higher rates of attrition.<sup>39</sup> At Google, employees have been subjected to anti-diversity memos,<sup>40</sup> and women have experienced documented backlash from male employees on hiring. This alienation within high-tech workforces neither encourages nor welcomes diverse input into work products. It also may distract from efforts to incorporate elements of “diversity in the design” of algorithms, where biases can be avoided at the onset. Technologists may not be necessarily trained to identify cues that are outside of their cultural context and can be fenced into work groups that share similar experiences, values and beliefs. This is what some researchers have dubbed *inattentional blindness*.

These largely unconscious bias errors strongly support why high-tech companies should be striving for more diverse workforces to identify and quell online discrimination. Companies that are disrupting societal norms through the sharing economy, social media and the internet of things must do better to

---

<sup>36</sup> Atwell, J. (2016). Lack of women and minorities in senior investment roles at venture capital firms. *Deloitte*. Available at: <https://goo.gl/iah1VZ> (accessed June 25, 2019).

<sup>37</sup> Ibid.

<sup>38</sup> EEOC. (2016). Diversity in High Tech. *EEOC*. Available at: <https://goo.gl/EwKBUJ> (accessed June 25, 2019).

<sup>39</sup> Scott, A. Kapor Klein, F. Onovakpuri, U. (2017). Tech Leavers Study. *Kapor Center*. Available at: <https://goo.gl/Zgf6dg> (accessed June 25, 2019).

<sup>40</sup> Conger, K. (2017). Here's the 10-page anti-diversity screed circulating internally at Google. *Gizmodo*. <https://goo.gl/UEYNhx>. Available at: <https://goo.gl/9ctiyF> (accessed June 25, 2019).

address the less than remarkable representation of people of color as creators, influencers and decision makers.

As in the case of HBCUs and HSIs, the tech sector should work to strengthen those relationships and programs, which target these students for future employment. Congress and federal agencies, including the U.S. Department of Education, need to also do more to ensure that minority-serving institutions are establishing premiere programs that include both technology access and cutting-edge career development in fields where the nation will soon face massive shortages. We need to take notes from the former Obama administration that pushed the U.S. toward a "race to the top," urging collaboration between the private and public sectors to realize the nation's global competitiveness and edge over our international counterparts.

#### **Conclusion**

The prevalence of AI and machine learning should trigger alarms when we fail to have collaborative, proactive and productive discussions on their applications design and use. While many innocuous decisions will be best served by algorithms, others, especially those emanating from the financial services industry, may need more thoughtful consideration on their intended and unintended consequences. For developers seeking to deploy these emerging technologies, the engagement in conversations about its responsible and ethical deployment are at the core of these conversations, and potentially result in reduced risk to consumers and more deliberation in the identification and mitigation of online biases.

I want to thank Members of this Committee for including me in this conversation and look forward to your questions.



---

June 2019

# INSURANCE MARKETS

## Benefits and Challenges Presented by Innovative Uses of Technology

## GAO Highlights

Highlights of GAO-19-423, a report to congressional requesters

### Why GAO Did This Study

The innovative use of technology by insurance companies (insurtech) is growing and offers the potential to improve customer experiences while also lowering insurer costs. Some stakeholders have raised questions about how certain uses of insurtech could create both risks for consumers and challenges for regulators, and whether some challenges might slow technological innovation in the insurance sector.

GAO was asked to provide information on insurtech activities in the property/casualty and life insurance sectors. This report (1) identifies new uses of technologies and potential benefits and challenges for insurers and their customers; and (2) discusses what stakeholders identified as key challenges that could affect the adoption of new technologies, and actions taken to address those challenges. GAO reviewed available literature; analyzed relevant laws and regulations; and conducted interviews with more than 35 stakeholders, including federal and state regulators, technology companies, insurers, and consumer groups (selected based on literature reviews and recommendations, and for relevance to the scope of GAO's review).

GAO is not making any recommendations in this report.

View GAO-19-423. For more information, contact Anna Maria Ortiz at (202) 512-8678 or [ortiza@gao.gov](mailto:ortiza@gao.gov).

June 2019

## INSURANCE MARKETS

### Benefits and Challenges Presented by Innovative Uses of Technology

#### What GAO Found

Insurtech companies (recently established companies bringing technology-enabled innovations to the insurance industry) as well as established insurers have begun to use technologies, including artificial intelligence (AI) and mobile applications, in an attempt to improve risk assessment and enhance customer experiences. For example:

- Consumers can purchase insurance products specifically tailored to their situation and needs, such as renters or auto insurance that can be turned on and off as needed using a mobile app.
- Some insurers have begun to use nontraditional data (such as from social media) to analyze policyholder risk, and use AI and complex algorithms to reduce costs by automating information gathering and risk assessment.

However, implementing these technologies can create potential challenges for insurers and risks for consumers, including the following:

- The use of AI to create underwriting models for determining premium rates can make it challenging for insurers to ensure that factors prohibited by regulation (such as race) are not used in models. Such models are often developed by data scientists who, unlike actuaries, may not fully understand insurance-specific requirements.
- Insurer collection and use of consumer data not provided by the consumer raise questions about data accuracy, privacy, and ownership.
- Some insurtechs sell coverage through nonadmitted insurers. As we have previously reported, nonadmitted insurers—unlike traditional insurers—are not required to be licensed in each state in which they sell insurance, and receive less regulatory oversight of their policies and rates. Also, if nonadmitted insurers became insolvent, state guaranty funds would not be available to help pay policyholder claims.

Stakeholders with whom GAO spoke identified challenges they said might affect adoption of innovative technologies. These include paper-based documentation requirements that do not accommodate online insurance transactions, and challenges for regulators in the evaluation of complex rating models. The National Association of Insurance Commissioners (NAIC) and state regulators have initiated a number of actions designed to address such concerns. For example:

- State insurance regulators, through an NAIC task force, have been examining regulatory areas that may pose obstacles for innovation, such as requirements for paper documentation or signatures.
- NAIC issued draft best practices for states to use when reviewing complex rating models.
- NAIC adopted a model law that creates a legal framework for states to use to require insurance companies to operate cybersecurity programs and protect consumer data.

Because many of these regulatory initiatives are still in development (or recently developed), the effect on innovation and consumer protection is unknown.

---

## Contents

---

Letter		1
	Background	3
	Emerging Use of Technologies Can Reduce Insurance Costs and Expand Product Choices but Creates Privacy and Other Challenges	9
	NAIC and State Regulators Initiated Actions to Address Challenges That Stakeholders Said Could Affect Adoption of Technologies	26
	Agency and Third Party Comments	33
Appendix I	Objectives, Scope, and Methodology	36
Appendix II	GAO Contact and Staff Acknowledgments	38
Figures		
	Figure 1: Examples of How Technology Can Automate the Automobile Claim Process	12
	Figure 2: Potential Benefits and Challenges Technologies Present for Insurers and Consumers	14

---

**Abbreviations**

AI	artificial intelligence
app	application
EU	European Union
insurtech	insurance technology
NAIC	National Association of Insurance Commissioners
Treasury	Department of the Treasury

This is a work of the U.S. government and is not subject to copyright protection in the United States. The published product may be reproduced and distributed in its entirety without further permission from GAO. However, because this work may contain copyrighted images or other material, permission from the copyright holder may be necessary if you wish to reproduce this material separately.





June 7, 2019

Congressional Requesters

The innovative use of technology by insurance companies (insurtech) is growing and offers the potential to reduce insurer costs while enhancing customer experiences. In recent years, both insurtech companies (recently established companies bringing technology-enabled innovations to the insurance industry) and established insurers have begun to use technologies, such as artificial intelligence (AI), to explore ways in which to improve operations and functions such as risk assessment, marketing, and product development. As consumers, and millennials in particular, have become well-versed in new technologies and taken a more hands-on approach to purchasing insurance, insurtechs have emerged to offer customized insurance products and streamlined customer experiences.<sup>1</sup>

At the same time, some stakeholders have expressed concerns that certain uses of technology could create risks for consumers, including potential misuse of data. Some stakeholders also have said the current insurance regulatory system slows technological innovation. As we noted in recent reports on data, analytics, and AI, the technologies have produced benefits such as reduced cost and increased accuracy in some areas of business, but also can pose privacy and civil liberties risks and their use could result in undesirable or unexpectedly biased outcomes.<sup>2</sup>

<sup>1</sup>According to Census Bureau estimates, by 2014 millennials outnumbered baby boomers as the largest living generation. The baby boomer generation consists of people currently ages 55–73 and the millennial generation of people currently ages 19–37.

<sup>2</sup>See GAO, *Data and Analytics Innovation: Emerging Opportunities and Challenges*, GAO-16-659SP (Washington, D.C.: Sept. 20, 2016); and *Artificial Intelligence: Emerging Opportunities, Challenges, and Implications*, GAO-18-142SP (Washington, D.C.: Mar. 28, 2018).

---

You asked us to provide an overview of insurtech activities in the property/casualty and life insurance sectors.<sup>3</sup> Specifically, this report (1) identifies uses of technologies and the benefits and challenges they might present for insurers and their customers, and (2) discusses what stakeholders identified as key challenges that could affect the adoption of new technologies, and actions that have been taken to address those challenges.

To address both objectives, we examined insurtech activities in the property/casualty and life sectors of the U.S. insurance market, including information on personal and commercial insurance where available. We did not include the health insurance sector because of significant differences between that sector and the property/casualty and life insurance sectors in terms of products offered and methods by which they are sold and regulated.<sup>4</sup> We conducted background research and a literature review to understand the most prominent, or key, technologies being used in the insurance industry and to identify any analyses of potential benefits and challenges that insurtech products and services may pose. Because insurtech is a fairly new field, we found few academic publications related to our objectives. We also conducted more than 35 semi-structured interviews with and reviewed documents provided by knowledgeable stakeholders to identify and obtain information about (1) current, in-development, and potential future uses of existing or new technology in the insurance industry; (2) stakeholder views on the potential benefits and challenges such technology presents to insurance companies and consumers; (3) which challenges may affect insurers' adoption of technology; and (4) actions the National Association of Insurance Commissioners (NAIC) and selected state insurance regulators

<sup>3</sup>Advances in technology and widespread internet and mobile device use also helped fuel the rise of fintech (the provision of traditional financial services by non-traditional technology-enabled providers). We issued a series of reports examining fintech and made recommendations to address areas including fintech regulation and use of alternative data sources in underwriting. See GAO, *Financial Technology: Information on Subsectors and Regulatory Oversight*, GAO-17-361 (Washington, D.C.: Apr. 19, 2017); *Financial Technology: Additional Steps by Regulators Could Better Protect Consumers and Aid Regulatory Oversight*, GAO-18-254 (Washington, D.C.: Mar. 22, 2018); and *Financial Technology: Agencies Should Provide Clarification on Lenders' Use of Alternative Data*, GAO-19-111 (Washington, D.C.: Dec. 19, 2018).

<sup>4</sup>Health insurance is the third-largest sector. It includes products from private health insurers, as well as government programs. Both the property/casualty and life sectors also write some health insurance.

---

have been taking or might consider to address these challenges.<sup>5</sup> The stakeholders included the Federal Insurance Office, NAIC, selected state insurance regulators, associations representing state agencies, academics, consumer groups, insurance providers and industry associations, actuarial professional associations, consulting groups, lawyers in the field, and technology providers.<sup>6</sup> We identified potential interviewees by conducting internet research, reviewing literature search results, and reviewing recommended interviewees from our initial interviews. Finally, we reviewed NAIC model laws and state laws to identify any relevant to the development and implementation of insurtech. See appendix I for more information on our scope and methodology.

We conducted this performance audit from April 2018 to June 2019 in accordance with generally accepted government auditing standards. Those standards require that we plan and perform the audit to obtain sufficient, appropriate evidence to provide a reasonable basis for our findings and conclusions based on our audit objectives. We believe that the evidence obtained provides a reasonable basis for our findings and conclusions based on our audit objectives.

---

## Background

Insurance allows individuals and businesses to manage risk by providing compensation for certain losses or expenses, such as those from car accidents, fires, medical services, or inability to work. According to NAIC, as of December 31, 2017, there were 2,509 property/casualty companies and 852 life insurance companies in the United States and its territories. In 2017, premiums written for the property/casualty sector totaled \$602.2 billion in 2017 and premiums written for the life and health sector totaled \$683.2 billion.<sup>7</sup>

---

<sup>5</sup>NAIC is the standard-setting and regulatory support organization created and governed by the chief insurance regulators from the 50 states, the District of Columbia, and five U.S. territories (Guam, American Samoa, Puerto Rico, U.S. Virgin Islands, and the Northern Mariana Islands).

<sup>6</sup>For our discussion of stakeholder views on benefits and challenges in the primary areas they identified as being affected by technology, we define "some" as stakeholders from three or four categories and "several" as stakeholders from five or more categories.

<sup>7</sup>The life and health sector consists mainly of life insurance and annuity products. Most private health insurance is written by insurers whose main business is health insurance, which is not discussed in this report. Premium data are from NAIC. See National Association of Insurance Commissioners, *2017 Insurance Department Resources Report*, vol. II (Washington, D.C.: 2018).

---

As we have noted in recent reports, advances in technology and widespread use of the internet have brought about significant changes in the financial industry.<sup>8</sup> For example, in recent years technology has changed consumer expectations and preferences, with younger consumers especially being well-versed in new technologies and looking to take a more hands-on approach to managing their finances. Similarly, over the last 5 years, established insurers and insurtech companies have used technology to offer simpler insurance products and streamlined customer experiences. Insurtech companies have been playing a variety of roles in the U.S. insurance market. Key players in insurtech include the following:

- **Insurtech companies (typically startups) that are licensed insurance companies.** Insurtech startups offer innovative products and services and are active in all major insurance products and all lines of business, with concentrations in the property/casualty business. For example, according to its website, Lemonade Insurance Company is a property/casualty insurer that sells products exclusively through mobile applications (apps) and its website. It offers renters, condominium, and homeowners insurance in several states. Another example is Root, which describes itself as an automobile insurance company that uses a smartphone app to understand individual driving behavior. Customers can download the Root app to their smartphones, obtain a personalized quote after a 2–3 week test drive, and purchase and manage their policy entirely within the mobile app.
- **Insurtech companies that do not provide insurance themselves, but offer technology solutions for insurers.** For example, according to the website for Groundspeed Analytics, they use AI and data science methods to provide information for the commercial property/casualty insurance industry to help identify potential areas of profit and enhance the customer experience. According to the website for Habit Analytics, they use real-time consumer data, sourced from smartphones and connected devices in homes, to create behavioral profiles that enable insurance companies to provide input for their risk models. Many established insurers have been acquiring such companies.
- **Established insurers that use technologies or partner with insurtech companies.** For example, the insurer Nationwide notes on its website that it created Nationwide Ventures to invest in startups,

---

<sup>8</sup>See GAO-17-361 and GAO-18-254.

pilot new technologies, and test new solutions and business models by exploring topics that range from analytics and automation technology to new insurance and financial services platforms.

According to analysis by the Deloitte Center for Financial Services and data collected by research firm Venture Scanner, as of mid-2018 there were more than 1,000 insurtech firms established in more than 60 countries, with more than half of those launched in the United States since 2008.<sup>9</sup>

#### State Licensing Regulation for Admitted and Nonadmitted Insurance Markets

Insurance companies are regulated principally by the states and are licensed under the laws of a single state, known as the state of domicile. Companies may conduct business in multiple states, but the state of domicile serves as an important regulator. State regulators license insurance agents, generally review and approve insurance products and premium rates, and examine insurers' financial solvency and market conduct. As we have previously reported, state regulators typically conduct financial solvency examinations every 3–5 years, while market conduct examinations are generally done in response to specific consumer complaints or regulatory concerns.<sup>10</sup> To help ensure that policyholders continue to receive coverage if their insurer becomes insolvent or unable to meet its liabilities, states also have guaranty funds (separate for life and property/casualty insurance), which are funded by assessments on insurers doing business in those states.<sup>11</sup>

Individuals who wish to sell, solicit, or negotiate insurance in the United States must generally be licensed as producers, a term including insurance agents and insurance brokers. Insurance agents typically represent only one insurance company. Insurance brokers represent multiple insurance companies and are free to offer a wider range of products to their clients. Brokers can search the market and obtain

<sup>9</sup>Deloitte Center for Financial Services, *InsurTech Entering Its Second Wave: Investment Focus Shifting from New Startups to More Established Innovators* (2018), and Venture Scanner web site <https://www.venturescanner.com/insurance-technology>, accessed on March 14, 2019.

<sup>10</sup>See GAO, *Insurance Reciprocity and Uniformity: NAIC and State Regulators Have Made Progress in Producer Licensing, Product Approval, and Market Conduct Regulation, but Challenges Remain*, GAO-09-372 (Washington, D.C.: Apr. 6, 2009).

<sup>11</sup>According to NAIC, all 50 states, Puerto Rico, the U.S. Virgin Islands, and the District of Columbia have a guaranty mechanism for the payment of covered claims arising from the insolvency of insurers licensed in their state.

---

multiple price quotes to fit their clients' needs. Producers must comply with state laws and regulations governing their activities. NAIC notes that as of September 2018, more than 2 million individuals and more than 200,000 business entities were licensed to provide insurance services across all lines of insurance in the United States.

Traditional insurers, sometimes referred to as admitted insurers, can be licensed to sell several lines or types of coverage to individuals or families, including personal lines—such as homeowners, renters, and automobile insurance—and commercial lines—such as general liability, commercial property, and product liability insurance. Admitted insurers can sell insurance in one or more states but, according to NAIC, must be licensed to operate in every state in which they sell coverage. To help ensure adequacy and fairness in pricing and coverage, state regulators oversee the insurance rates and forms of admitted insurers. State regulators also may require admitted insurance companies to maintain specific levels of capital to continue to conduct business.

The surplus lines insurance market, also known as the nonadmitted market, can provide insurance coverage for risks that traditional insurers are unwilling or unable to cover. The risks covered can include potentially catastrophic property damage and liability associated with high-hazard products, special events, environmental impairment, and employment practices.<sup>12</sup> In the absence of the surplus lines market, NAIC notes that some insureds in those markets would be unable to secure coverage.<sup>13</sup>

In most states, surplus lines insurers cannot write insurance coverage that is available from admitted insurers and only may write coverage rejected by a number of admitted insurers, according to NAIC. Furthermore, in those states, the surplus lines insurance broker must conduct a "diligent search" of the admitted insurance market to determine if comparable coverage is available. The broker can write coverage only if a specified number of admitted insurers have declined to offer such coverage.

---

<sup>12</sup>For more information on surplus lines insurers, see GAO, *Property and Casualty Insurance: Effects of the Nonadmitted and Reinsurance Reform Act of 2010*, GAO-14-136 (Washington, D.C.: Jan. 16, 2014).

<sup>13</sup>According to data from NAIC, as of year-end 2016 (the most recent available), surplus lines premium volume across all lines of insurance was \$44.5 billion, which was 7.4 percent of the \$602.3 billion in premium volume in the admitted market.

---

According to NAIC, new and innovative insurance products for which there is no loss history may be difficult to appropriately price. According to stakeholders we interviewed, the nonadmitted market is therefore a common entry point into the insurance market for insurtech firms that want to sell insurance products. NAIC notes that, after a new coverage has generated sufficient data, the coverage often eventually moves to, and is sold by, insurers in the admitted market. For example, private flood insurance was developed and first offered in the nonadmitted market but now also is offered in the admitted market.

The nonadmitted market is generally regulated somewhat differently than the admitted market. According to NAIC, surplus lines insurers are subject to regulatory requirements and are overseen for solvency by their domiciliary state or country, but surplus lines transactions are regulated through the licensing of surplus lines brokers. NAIC states these brokers are responsible for ensuring that the surplus lines insurer meets eligibility criteria to write policies in the state and is financially sound. Furthermore, NAIC notes surplus lines brokers and producers must be licensed to sell surplus lines insurance in each state in which they operate. State insurance departments may have authority to suspend, revoke, or not renew the license of a surplus lines broker or producer. Unlike admitted insurers, surplus lines insurers may not have access to state guaranty funds that are available to help pay claims in the event of an insurer insolvency. In addition, according to NAIC, surplus lines insurers generally have more freedom to change policy coverages and premium rates than admitted insurers. NAIC stated that state regulators require both nonadmitted and admitted insurance companies to maintain specific levels of capital to continue to conduct business. According to NAIC, most state insurance regulators also can use their authorities under state statutes such as an unfair trade practices act to ensure consumers are protected (for example, to ensure that claims are paid and insurers or brokers do not misrepresent policy terms) and to remedy other bad conduct.<sup>14</sup>

---

<sup>14</sup>NAIC's model Unfair Trade Practices Act prohibits a number of specifically defined unfair trade practices if they are committed flagrantly and in conscious disregard of the Act or any rules implementing the Act, or have been committed with such frequency to indicate a general business practice. Under the model law, if, after a hearing, the state insurance commissioner finds a violation of the Act, the commissioner is to file a cease and desist order and may at their discretion impose a limited monetary penalty or suspend or revoke the insurer's license. See NAT'L ASS'N OF INS. COMM'RS, UNFAIR TRADE PRACTICES ACT, MDL-880-1, §§ 3, 8, 11 (2004).

---

---

**Other Participants in the  
Regulatory Framework for  
Insurance**

NAIC assists state regulators with various oversight functions. While NAIC does not regulate insurers, it provides services designed to make certain interactions between insurers and regulators more efficient. These services include providing detailed insurance data to help regulators understand insurance sales and practices; maintaining a range of databases useful to regulators; and coordinating regulatory efforts by providing guidance, model laws and regulations, and information-sharing tools.

The Federal Insurance Office was established in the Department of the Treasury (Treasury) by the Dodd-Frank Wall Street Reform and Consumer Protection Act.<sup>15</sup> The office is headed by a director appointed by the Secretary of the Treasury. The Federal Insurance Office monitors all aspects of the insurance industry (including by identifying issues or gaps in insurance regulation that could contribute to systemic risk in the insurance industry), and helps develop federal policy on international insurance matters, but is not a regulatory agency itself. The office also serves as an information resource for the federal government and coordinates with federal regulators, state insurance regulators, and NAIC. The Federal Insurance Office also represents the United States in the International Association of Insurance Supervisors and coordinates federal efforts in international insurance matters.

---

<sup>15</sup>Pub. L. No. 111-203, § 502, 124 Stat. 1376, 1580 (2010), codified at 31 U.S.C. § 313. The Dodd-Frank Wall Street Reform and Consumer Protection Act also requires the Secretary of the Treasury to advise the President on major domestic and international prudential policy issues in connection with all lines of insurance except health insurance. *Id.* at § 502(b)(3), codified at 31 U.S.C. § 321(a)(9).



---

### Emerging Use of Technologies Can Reduce Insurance Costs and Expand Product Choices but Creates Privacy and Other Challenges

In recent years, the insurance industry has begun to adopt several types of technology that are designed to provide a range of benefits to insurers and consumers (policyholders), including improved risk monitoring, reduced costs, and improved underwriting.<sup>16</sup> However, the use of these technologies also can create challenges for insurers and potential risks for consumers, including changed business models, pricing fairness, and privacy issues.

---

### Insurance Industry Increasingly Using Mobile Apps, Big Data, and Other Technologies

Based on our literature review and interviews with stakeholders, we identified six key technologies that have seen increased use in the insurance industry in recent years and one technology (blockchain) that has seen limited adoption and which the industry has been exploring for wider use.<sup>17</sup>

- **Mobile apps.** A mobile app is software designed to run on a mobile device, such as a smartphone or tablet computer. Insurance industry stakeholders told us that several insurers have adopted mobile apps to make their products and services available on mobile devices. For example, insurers have adopted mobile apps that allow consumers to purchase products online. An increased number of insurers in recent years also have adopted mobile apps that allow customers to complete tasks online such as submitting insurance claims and turning on-demand insurance coverage on or off. Insurers also have been using mobile apps to capture consumer data and usage patterns (behaviors).
- **AI, algorithms, and machine learning.** AI is the development of computer systems to perform tasks and make decisions that historically have required human intelligence to perform. Machine learning is a subset of AI and focuses on the ability of machines to

<sup>16</sup>Insurance underwriting is the process of evaluating a consumer's risk using specific data and information provided by the consumer, as well as other relevant information, determining the rate associated with the risk, and deciding whether to accept the risk and insure the consumer.

<sup>17</sup>Insurance industry participants may be applying technology to other uses beyond those we identify in this report, but we focused on those that our interviewees and the literature we reviewed most often identified. For example, peer-to-peer lending utilizing block-chain technology was mentioned in some articles, but not by our interviewees.

---

receive a set of data and learn for themselves, changing algorithms as they learn more about the information they process. (Algorithms are sets of rules that a computer or computer program follows to compute an outcome.) In the insurance industry, AI includes applications that provide specific expertise or allow for task completion. For example, AI provides on-line "chatbots" (sometimes called robo-advisory services) that answer questions specific to an insurance product or service.<sup>18</sup> When a consumer communicates with a chatbot, the chatbot takes the information the consumer provided and enters it into an algorithm. Based on protocols outlined in the algorithm, the chatbot provides a response to the consumer's question. As the conversation moves forward, the chatbot will adapt to answer more questions using machine learning in real-time. According to insurance industry stakeholders, insurers have been using algorithms to analyze information obtained from other technology sources to determine what a consumer's risk profile is and then determine the consumer's premium rate based on their risk profile.

- **Big data.** Big data are large volumes of data (often aggregated from multiple sources to develop data sets). As we have noted in other work, big data are frequently analyzed using predictive analytics, machine learning, and data mining to identify trends, patterns and characteristics.<sup>19</sup> The insurance industry uses big data in several ways, including analyzing consumer information, identifying risk patterns and pricing risk, and analyzing information related to risk pooling. Insurers also use big data to streamline and more accurately underwrite products. For instance, an insurer may use big data to determine whether consumers are high- or low-risk based on factors identified from extensive datasets such as what they purchase online or how they shop for insurance online. This is similar to lenders' usage of big data. In a previous report, we noted that lenders were

---

<sup>18</sup>In a prior report (presenting results of a forum GAO held to discuss AI), we described a chatbot as a program that interacts directly with users through a natural language process in a free-form conversation. We also noted AI had no single universally accepted definition. For instance, researchers have distinguished between narrow AI (applications that provide domain-specific expertise or task completion) and general AI (applications that exhibit intelligence comparable to a human). See GAO-18-142SP. This report focuses on narrow AI applications.

<sup>19</sup>American Academy of Actuaries, *Big Data and the Role of the Actuary* (Washington, D.C.: 2018). We examined issues surrounding big data and the use of algorithms in decision making. See GAO-16-659SP.

---

using big data to evaluate risk and make lending decisions using real-time nontraditional information gathered from social media sites.<sup>20</sup>

- **Internet of things.** The internet of things refers to semi-autonomous and internet-capable devices (such as machinery, home appliances, thermostats, and smartphones) that have sensors that interact with the physical environment and typically contain elements for processing and communicating information.<sup>21</sup> Some insurers stated that the internet of things could be used in the insurance industry to track and reduce risk, detect problems, and mitigate potential claims. For example, a homeowner could have a smart home thermostat that sends alerts when the power goes off and indoor temperature decreases. With the homeowner able to address the issue in real time, the homeowner could mitigate the risk of frozen pipes bursting and potentially prevent a loss and an insurance claim. According to CBInsights, insurers have partnered with insurtech firms that provide this technology to offer real-time monitoring.<sup>22</sup>
- **Drones.** Drones are remotely piloted aircraft systems. Insurers have been using drones for a variety of purposes in the insurance industry. For example, insurers use drones to obtain aerial footage over a disaster area to determine the amount of damage to a house or crop field. Insurance companies also use drones to verify information submitted by a policyholder in a claim or help determine the risk presented by difficult-to-reach areas of a property, such as a roof.
- **Telematics.** Telematics combines telecommunications and information processing to send, receive, and store information related to specific items such as automobiles or water heaters. Telematics often uses sensors to relay information such as global positioning system location, speed, and water levels. For example, sensors in an automobile can provide data on a driver's behavior (such as speed, hard braking, and turning radius). The insurer may use that information to determine the driver's risk profile and help determine the premium rate for that driver.

These technologies can be used together. For example, a telematics device can be used to provide data to a mobile app, which can then send

---

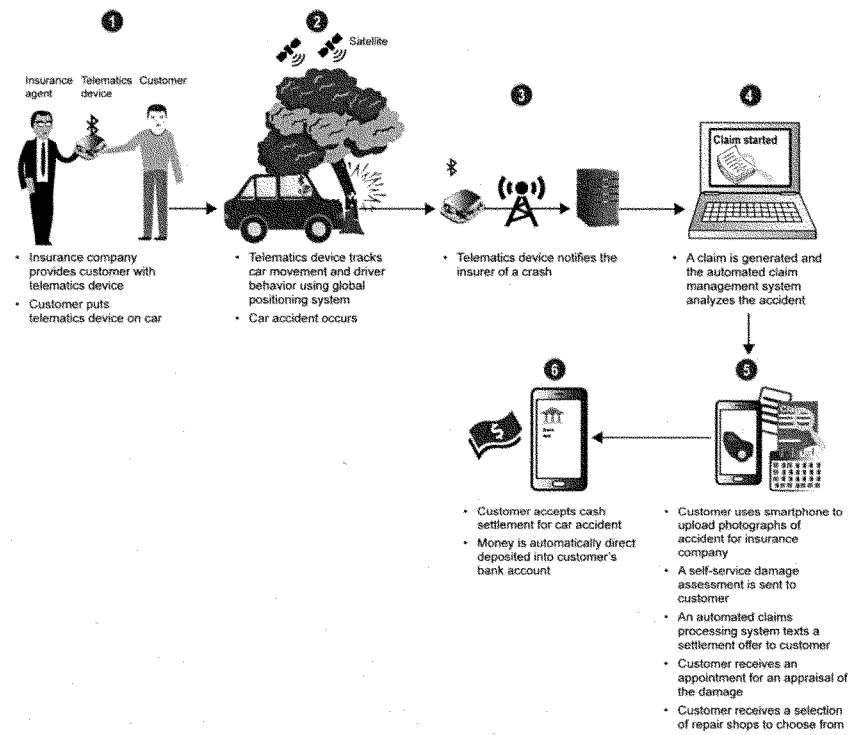
<sup>20</sup>See GAO-16-659SP.

<sup>21</sup>GAO, *Internet of Things: Enhanced Assessment and Guidance Are Needed to Address Security Risks in DOD*, GAO-17-668 (Washington, D.C.: July 27, 2017).

<sup>22</sup>CBInsights, *How Major Insurers Are Teaming Up with Internet of Things Companies in One Infographic* (Dec. 7, 2015).

the information to an AI algorithm to determine whether a claim should be paid. See figure 1 for examples of the types of technologies that insurers may use to automate the claims process.

Figure 1: Examples of How Technology Can Automate the Automobile Claim Process



Source: GAO. | GAO-19-423

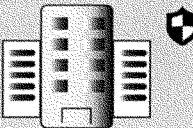
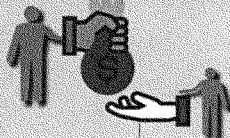
- 
- **Blockchain/ distributed ledger technology and smart contracts.** The insurance industry has been studying whether blockchain technology could be used to improve insurance processes. Blockchain refers to a type of distributed ledger technology—in which multiple entities and locations share and synchronize datasets—that facilitates and permanently records virtual transactions. Information is uploaded and recorded in a series of secured blocks; the information uploaded cannot be modified or erased once uploaded into the blockchain (thus providing an accurate history of specific transactions and information). According to insurers, blockchain could be used by the industry to track insurance coverage history, expedite the claims process, provide an audit trail of insurance transactions, and address cybersecurity issues. For instance, a blockchain could expedite the claims process by allowing agents, policyholders, and repair companies immediate, secure access to certain data that are part of the claim only as the data are needed. “Smart contracts” include provisions for contract performance that can be executed by a computer algorithm (for instance, on a blockchain). For example, an insurer stated that a smart contract for homeowners insurance might stipulate that if an earthquake of a specific size occurred in a policyholder’s residential area, a claim payment for damage in a specified dollar amount automatically would be made from the insurer to the policyholder. According to NAIC, adoption of blockchain technology in insurance is limited at this time.

---

**Technologies Can Create Benefits but Also Present Risks to Insurers and Consumers**

According to stakeholders with whom we spoke and literature we reviewed, the use of technology in the insurance industry creates potential benefits but also can create risks for both insurers and consumers. We present stakeholder views on the benefits and challenges technology presents in the primary areas they identified as being affected by technology, which include (1) pricing and risk evaluation, (2) consumer protection, (3) business operations and risk monitoring, and (4) product offerings. See figure 2 for a summary of the potential benefits and challenges we discuss.

Figure 2: Potential Benefits and Challenges Technologies Present for Insurers and Consumers

	Insurers		Consumers	
				
Insurance area affected by technologies	✓ Potential benefits	✗ Potential risks	✓ Potential benefits	✗ Potential risks
Pricing and risk evaluation	<ul style="list-style-type: none"> <li>✓ Increased underwriting accuracy</li> <li>✓ More individualized pricing</li> </ul>	<ul style="list-style-type: none"> <li>✗ Validating consumer data and models</li> </ul>	<ul style="list-style-type: none"> <li>✓ More individualized, risk-based pricing</li> </ul>	<ul style="list-style-type: none"> <li>✗ Quality of data used in pricing</li> <li>✗ Decreased pooling of risk</li> </ul>
Consumer protection		<ul style="list-style-type: none"> <li>✗ Protecting consumer data</li> </ul>		<ul style="list-style-type: none"> <li>✗ Consumer privacy concerns</li> <li>✗ Consumer protection concerns</li> </ul>
Business operations and risk monitoring	<ul style="list-style-type: none"> <li>✓ Reduced costs</li> <li>✓ Improved risk monitoring</li> </ul>	<ul style="list-style-type: none"> <li>✗ Connecting to legacy computer systems</li> <li>✗ Changing roles for insurers and agents</li> </ul>		
Product offerings	<ul style="list-style-type: none"> <li>✓ Ability to offer on-demand products</li> </ul>		<ul style="list-style-type: none"> <li>✓ Increased convenience</li> <li>✓ Increased consumer choice</li> </ul>	

Source: GAO analysis. | GAO-19-423

Pricing and Risk Evaluation

According to stakeholders we interviewed and literature we reviewed, the use of technology for determining insurance pricing and coverages creates several benefits and risks for insurers and consumers:

- **Increased underwriting accuracy.** Insurers and others told us that insurers have been using technologies that provide enhanced analytic capabilities or data from previously unavailable sources to increase the accuracy of underwriting. These technologies allow insurers to make new connections between policyholder characteristics and risk. That is, insurers are using big data, AI, and algorithms to obtain and

---

analyze more information about consumers than they previously had been able to obtain. For instance, a property/casualty insurer could collect data on when consumers set their home alarms and use this and other risk information to refine risk determinations for those individuals. Another example is when insurers use data collected from telematics devices in automobiles to inform the insurer about the policyholder's risk of being involved in an accident. A better understanding of the risk presented by policyholders can help insurers more accurately and effectively price and manage risks.

- **More individualized pricing.** Insurers also have been using technologies to underwrite policies in a way that results in more individualized pricing, which benefits insurers and could benefit some consumers. That is, big data can allow an insurer to use factors for which traditional underwriting typically has not accounted.<sup>23</sup> According to stakeholders we interviewed, doing so allows an insurer to place an individual in a smaller risk pool than if traditional underwriting factors were used and to price coverage for that individual more in line with the risk that individual presents. This can help an insurer better manage its level of risk by offering lower prices to lower-risk customers, charging more for higher-risk customers, or even declining to offer coverage to consumers it considers high-risk.

Some stakeholders told us that technologies allow consumers to receive more individualized premium rates, based on their risk characteristics, than had been possible. For example, some insurers have been using telematics devices to obtain information on policyholder driving habits and the risk level they present and adjust premium rates based on this information. As a result, consumers who engage in safer driving practices receive the benefit of lower premiums. Policyholders also could use such information to take actions that will lower their risk level and therefore their premiums. For instance, consumers could seek to reduce specific driving behaviors, such as fast stops or starts, which negatively affect their premium rate. However, consumers with higher-than-average risks could end up paying more or perhaps be declined coverage.

---

<sup>23</sup>According to the World Bank, after identifying risks, an underwriter will classify the insured into the appropriate risk class. Classifying risk into classes allows the insurance company to determine the appropriate premium rate that should be charged. Not differentiating risk classes would result in some insureds being charged too much premium while others would be "cross-subsidized" (they would be charged less than the actual cost for insurance). In a competitive market, this cross subsidy creates a serious competitive disadvantage for the insurance company.

---

Stakeholders including an industry representative and a law firm in the field indicated that insurers also might use data to exclude high-risk consumers from marketing. For example, an insurer might not choose to market to high-risk consumers to discourage them from buying their insurance. This approach, in theory, helps insurers decrease the number of high-risk policyholders they insure but could create difficulties for some seeking coverage.

Two industry representatives and an academic in the field indicated that the potential for decreased risk pooling creates a difficult question about the minimum extent of pooling that is socially desirable. For example, these stakeholders stated that when insurance underwriting becomes too individualized, it might no longer serve an insurance function; that is, there is very little pooling of risk. They stated it may be a desirable social benefit to have a certain level of risk pooling to allow more people to effectively manage their risk. In a November 2018 issue paper, the International Association of Insurance Supervisors noted the potential effect of more individualized underwriting on the fairness of consumer outcomes. Among other findings, the paper noted the collection of more data on policyholders may enable a more specific risk categorization that could affect risk pooling principles and lead to issues around affordability of certain insurance products or even availability (the potential for exclusion).<sup>24</sup> The association noted that insurance supervisors should monitor whether such negative consumer impacts become a trend and, if so, raise awareness at the appropriate policy and political level(s).<sup>25</sup>

- **Validating consumer data and models.** Insurers and insurtech firms increasingly have been using AI and data collection algorithms to gather data through mobile, wearable, and other internet-connected devices and from online sites. According to two academics in the field, collecting consumer data in large quantities and from multiple disparate sources, including social media, poses challenges for insurers in relation to validating those data. Insurers and insurtech firms also face challenges associated with validating models that use the data. Although AI and machine learning can help insurers and agents underwrite risk more accurately, these stakeholders said that

---

<sup>24</sup>International Association of Insurance Supervisors, *Issues Paper on Increasing Digitalisation in Insurance and its Potential Impact on Consumer Outcomes* (Basel, Switzerland: November 2018).

<sup>25</sup>International Association of Insurance Supervisors, *FinTech Developments in the Insurance Industry* (Basel, Switzerland: February 2017).



---

these tools and processes can increase risk because the collected information may be inaccurate or inappropriately used in determining premium rates. For example, while models may indicate that certain factors developed by AI from social media and other sources are associated with increased policyholder risk, it may be difficult or impossible for insurers to validate the accuracy of such data.

In addition, it can be a challenge for insurers to ensure that the use of such data and models does not result in the use of prohibited factors in determining premium rates, such as race or sex.<sup>26</sup> For example, several stakeholders told us that certain factors, while not specifically disallowed by insurance regulations, could end up serving as a proxy for a disallowed factor. One example cited by a stakeholder was the use of information on consumer magazine subscriptions, which are not prohibited on their own, but could serve as proxies for factors that are prohibited.

Finally, it can be a challenge for insurers to document and explain to regulators how rating models that use AI and machine learning work and provide assurance that the rates produced by the models are not unfairly discriminatory toward policyholders.<sup>27</sup> For example, some industry stakeholders we interviewed said that these models are often developed by data scientists and not actuaries, as had been the case in the past. Unlike actuaries, they said data scientists who develop rating models may not fully understand insurance-specific requirements, such as setting premium rates that are not unfairly discriminatory, and may struggle to measure the impact of new variables used in the models. Furthermore, data scientists may be unfamiliar with insurance rules and regulations and may not understand how to communicate their work to state insurance regulators. One regulator described to us how one insurance company was unable to explain how one of the factors that it entered into its advanced risk model—proximity of a home to a day care center—related

---

<sup>26</sup>For example, NAIC Model Law 880 defines unfair discrimination as, in part, refusing to insure, refusing to continue to insure, or limiting the amount of coverage available to an individual because of the sex, marital status, race, religion or national origin of the individual. See NAT'L ASS'N OF INS. COMM'RS, UNFAIR TRADE PRACTICES ACT, MDL-880-1, § 4 (2004).

<sup>27</sup>For an example of how insurance underwriting can lead to unfair discrimination against policyholders, see New York Department of Financial Services, *RE: Use of External Consumer Data and Information Sources in Underwriting for Life Insurance*, Insurance Circular Letter No. 1 (Jan. 18, 2019); accessed at [https://www.dfs.ny.gov/industry\\_guidance/circular\\_letters/cl2019\\_01](https://www.dfs.ny.gov/industry_guidance/circular_letters/cl2019_01).

---

to the risk that a consumer posed. An actuarial group suggested a greater collaboration between actuaries and data scientists could provide greater assurance that such rating models meet regulatory requirements.

- **Quality of data used in pricing.** According to some stakeholders, insurers' use of nontraditional data and AI to develop insurance pricing models creates two potential risks for consumers that parallel some of the risks for insurers. First, as previously mentioned, insurer's use of nontraditional data and AI can create a risk that factors unrelated to the risk presented by a consumer could be used to set his or her premium rate. Stakeholders including a regulator said that algorithms or big data may allow insurers to correlate certain factors with higher claim rates, although the factors do not actually relate to risk and may even act as a proxy for a prohibited factor such as race or sex. As a result, some stakeholders noted that using such information to determine a premium rate could be unfairly discriminatory. Some stakeholders also said that such factors unintentionally could become proxies for prohibited rating factors—such as race. For example, using information on a consumer's purchase history could serve as a proxy for race.

Second, some stakeholders indicated that when insurers use AI to generate information on consumers, it is difficult to ensure these data are accurate. Because the data were not explicitly provided by the consumer, the consumer does not have a chance to correct or dispute the data. For example, if an insurer uses AI to pull data from a consumer's social media accounts, those data could be incorrect or outdated, but the consumer would not know the data were being used as a factor in determining his or her premium rate. This would prevent the consumer from correcting the information if it was wrong. Some stakeholders indicated that if an insurer has difficulty understanding the factors and algorithms being used to price the insurance product, the consumer most likely will not be able to understand them.

#### Consumer Protection

According to stakeholders with whom we spoke and literature we reviewed, some uses of technology can pose risks in terms of the protection of consumer data. In addition, the use of the nonadmitted market by insurtech companies and insurers may result in more limited financial protections for consumers.

- **Cost of protecting consumer data.** As noted earlier, insurers collect and use consumer data in large quantities and from multiple disparate sources, including social media, posing challenges for protecting those data. For example, according to representatives of one

---

property/casualty industry association we interviewed, it can be expensive to maintain the appropriate level of cybersecurity (including technical and organizational measures) to prevent any unauthorized access or use of the additional volumes and types of customer information used in recent years.

- **Consumer privacy concerns.** Stakeholders noted that insurers' expanded use of consumer data raises concerns about the privacy of such data. For example, an automobile insurer may collect data on a consumer using a telematics device installed in the consumer's vehicle. While an insurer may use data on the consumer's driving habits for the purpose of adjusting premium rates, the device also may collect information on where and when a consumer drives. This is information consumers may not wish others to possess.<sup>28</sup>

One academic also said there is concern about the ownership of the data collected through telematics and other technologies, such as AI, for the purposes of insurance. For instance, if an insurer obtained data from a policyholder's automobile with a telematics device, a question exists about whether policyholders would have the right to take those data to another insurer if they switched insurers or whether the data belong to the first insurer. As we have described in other work, this presents a larger privacy issue as it may not be possible for a consumer to know exactly what is collected, or when and how the data are used.<sup>29</sup> This lack of knowledge reduces the consumer's control over their personal information and limits their ability to track what data belong to them.

Some stakeholders mentioned concerns about insurers collecting information from social media and other sources that consumers did not explicitly consent to provide to insurers. The European Union (EU) General Data Protection Regulation, which includes regulations governing consumer consent, had an entry into force and application date of May 25, 2018.<sup>30</sup> According to an industry analyst, the General Data Protection Regulation applies to insurance companies around

<sup>28</sup>Consumers may not fully understand, anticipate, or consent to the end-user agreements or privacy statements related to data privacy or understand how the information collected from a telematics device could be used, shared, or sold. See GAO, *Vehicle Data Privacy: Industry and Federal Efforts Under Way, but NHTSA Needs to Define Its Role*, GAO-17-656 (Washington, D.C.: July 28, 2017): 24.

<sup>29</sup>GAO-17-656.

<sup>30</sup>See "Regulation (EU) 2016/679, Article 99, of the European Parliament and of the Council of April 27, 2016," *Official Journal of the European Union*, L 119 (May 4, 2016).

---

the world, including those in the United States, that process the personal data of EU residents, regardless of the nationality of the person in question or the location of the company.<sup>31</sup> Furthermore, the analyst notes that the regulation strictly defines legal uses of individuals' data and requires companies to ensure individuals can explicitly and individually consent to other uses of their data.<sup>32</sup> In prior reports, we also noted data privacy concerns in relation to lender use of financial technology.<sup>33</sup>

- **Consumer protection concerns due to use of the nonadmitted market.** The nonadmitted market is a common entry point for insurtech firms because of that market's usefulness for innovative insurance products with little loss history. However, the sale of consumer insurance through nonadmitted insurers raised concerns among several stakeholders. As we noted in a prior report, nonadmitted insurers may face fewer regulatory constraints than traditional insurers in the prices they can charge and their ability to create and offer new products.<sup>34</sup> While data do not exist on the number of insurtechs using the nonadmitted market, industry representatives told us that because of this greater regulatory freedom, a number of insurtechs choose to operate as nonadmitted insurers or as brokers selling policies through nonadmitted insurers. As described in the Background, when consumers purchase insurance from nonadmitted insurers, they do not have some of the same consumer protections they would have if they purchased

---

<sup>31</sup>See Mitchell Wein, "GDPR for North American Insurers," *CIPR Newsletter* (NAIC Center for Insurance Policy and Research: January 2019): 7.

<sup>32</sup>See Regulation 2016/679. As stated in Article 6(1), "[p]rocessing shall be lawful only if and to the extent that at least one of the following applies: (a) the data subject has given consent to the processing of his or her personal data for one or more specific purposes; (b) processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract, (c) processing is necessary for compliance with a legal obligation to which the controller is subject, (d) processing is necessary in order to protect the vital interests of the data subject or of another natural person, (e) processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller, (f) processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data, in particular where the data subject is a child."

<sup>33</sup>GAO-19-111.

<sup>34</sup>To help insurers operate in this environment, regulators generally provide nonadmitted insurers with greater pricing flexibility. See GAO-14-136.

---

coverage from an admitted insurer. For example, regulators conduct limited reviews of the prices charged and the products sold by nonadmitted insurers. And as noted earlier, if nonadmitted insurers became insolvent, state guaranty funds may not be available to help pay policyholder claims.

As we previously reported, some regulations serve to push potential policyholders toward the admitted market because of the better financial protections it provides (such as rate approvals and access to state guaranty funds).<sup>35</sup> For example, as noted earlier, a broker placing coverage with a nonadmitted insurer generally must conduct a diligent search for available coverage in the admitted market every time a potential policyholder requests coverage in the nonadmitted market. This helps ensure coverage is purchased from an admitted insurer as often as possible.

Stakeholders offered differing assessments on the extent of any related risks to consumers resulting from insurtech use of the nonadmitted market. For example, an industry representative said the nonadmitted market is not appropriate for most consumer products because of the lower consumer protections as compared with the admitted market. Two insurtech firms also have raised questions about the ability of insurtech companies and other market participants to properly comply with diligent search requirements. For example, an industry representative told us it does not seem possible to satisfy the diligent search requirement when products are sold on-demand through a mobile app. Furthermore, the representative raised the question of how a broker could legitimately search the admitted market for coverage in cases in which an insurer offers immediate coverage as soon as consumers complete applications on their smartphones.

Conversely, some insurers, regulators, and NAIC said that nonadmitted insurers are appropriately regulated and consumers are not necessarily at any greater risk than when purchasing coverage from admitted insurers. Also, several states have eliminated the diligent search requirements. However, a consumer advocate noted that such deregulation raises further consumer protection issues in a market where less regulation is already a concern for consumers.

Business Operations and Risk  
Monitoring

---

<sup>35</sup>GAO-14-136.

---

According to the literature we reviewed and stakeholders we interviewed, insurers have been using various technologies to reduce their operating costs but may face risks that affect their operations and business models.

- **Reduced costs.** Stakeholders described how adopting various technologies has led to reduced costs in four operational areas for insurers:

**Communicating with customers.** Insurers have been using mobile apps and chatbots to reduce the cost of providing information to potential customers. For example, a consumer might be shopping online for an insurance policy late in the evening. The insurer can use a chatbot to interact with that consumer and answer questions about insurance products. In the past, this might not have been possible if an agent was not available to work nonstandard business hours or insurers might have needed to hire and retain more agents to work evenings and weekends.

**Underwriting.** Insurers have been using technology to reduce the cost of underwriting insurance. For example, according to two insurtech firms and one industry representative we interviewed, some insurers review multiple sources of data with AI to automatically review the information in a consumer's insurance application, rather than incurring the costs of hiring staff to do so. Through the industry article review and stakeholder interviews, we found that insurers also use the internet of things to obtain data from smart home alarms to monitor consumer usage of alarm systems and thereby assess consumer risk levels. This reduces the costs associated with determining and analyzing risk factors.

**Claims processing.** According to some stakeholders we interviewed, insurers now have the capability to digitally collect and automatically analyze claim evidence, thereby reducing staffing needs and realizing cost savings. For example, consumers can use their smartphones to take photographs of their vehicles after an accident and send the photographs and other information to their insurers through mobile apps. On receipt of the photographs, insurers can use AI algorithms to verify the damage shown—decisions that historically required human intelligence to perform—and automatically start the claims process for the consumer.

**Fraud.** Insurers are able to detect fraud, or decide which claims need to be investigated further by employees, with information verified

---

using big data, the internet of things, and telematics.<sup>36</sup> For instance, an insurer may verify information provided in a claim against information obtained from a smart device to determine if the information provided by the policyholder was accurate. An insurer also might identify a false burglary claim by verifying whether an alarm was set during the time frame identified in the claim and reviewing video from home security cameras.

- **Connecting to legacy computer systems.** Some industry stakeholders and association representatives we interviewed stated that established insurers face significant challenges using new technologies because they first have to replace legacy computer systems or customize their systems to interface with new technologies properly. According to industry stakeholders, legacy computer systems were, in some ways, built around satisfying regulatory requirements rather than enhancing the consumer experience or providing more desirable products. They noted it can be costly and difficult to replace such systems or to modify them to interface with more consumer-centered systems, such as those being developed by insurtech companies.
- **Changing roles for insurers and agents.** According to some insurance industry stakeholders, emerging uses of key technologies and innovative business models could lead to changes in insurers' roles and products. For example, with the advent of self-driving vehicles, the liability for accidents could shift from the driver to the vehicle maker or the company that produced the self-driving system. In such cases, they said insurance coverage primarily would be sold to those entities rather than the consumer, and the demand for and amount of consumer automobile coverage sold could decrease substantially. This could cause a shift in demand for products from consumers to commercial lines, resulting in the potential loss of business for some agents and insurers. Some industry stakeholders we interviewed also told us that as more technologies (such as telematics or other smart devices) were adopted to help consumers mitigate risk, insurers likely would have to shift their business model. That is, they would have to move from a model focused on sales of policies, in which agents play a central role, to a model focused on

---

<sup>36</sup>Fraud can occur when policyholders provide false information to their insurer when making a claim. According to the Federal Bureau of Investigation, the estimated total cost of insurance fraud (for nonhealth insurance) is more than \$40 billion a year. Federal Bureau of Investigation, *Insurance Fraud*, accessed December 31, 2018, <https://www.fbi.gov/stats-services/publications/insurance-fraud>.

---

providing consulting services to consumers to help them prevent and mitigate risk and loss.

- **Risk monitoring.** Insurers have been using big data with data aggregation and mining to improve monitoring of insured risks. More specifically, several stakeholders told us that these tools and analytical methods can help insurers quickly analyze volumes of data from many sources in or near real time. For example, several stakeholders gave the example of an insurance company using sensors or other devices to continuously collect verified data on movements of insured ships and their cargo. Such data can be useful to insurers for understanding the risks associated with providing insurance coverage and even can be used to provide the ship carrying the cargo the appropriate insurance documentation required for the port of entry. Several stakeholders also told us that some insurtech companies have been using telematics to collect real-time data on driver behavior, which they combine with other information such as credit scores, to develop a fuller and more accurate picture of the risk presented by a given policyholder. Insurers then can use these risk profiles to determine whether to change a policyholder's rates or continue to insure them. Several stakeholders indicated that such real-time information is likely more accurate than previous risk-assessment methods.

#### Product Offerings

According to stakeholders we interviewed and literature we reviewed, the use of various technologies to create new product offerings has created several benefits for insurers and consumers.

- **Ability to offer on-demand products.** Technologies have been helping insurers tailor products to specific consumer needs and expand offerings to niche markets. Some insurtech companies have started offering on-demand insurance (insurance that policyholders can turn on and off as needed). For example, one regulator and an academic said that market research data demonstrated that consumers want to be able to turn on insurance for their drones when the drones are in use and turn it off when the drones are idle. Insurers also have been developing similar on-demand products for drivers working for rideshare companies such as Lyft and Uber and for Airbnb and VRBO rentals (to cover the gaps that traditional homeowners insurance, which generally provides coverage on a long-term basis, might have in relation to short-term rentals of homeowners' properties). On-demand products allow insurers to diversify their product lines and attract more consumers, which is discussed later in this report.



- 
- **Increased convenience.** With some insurers providing mobile apps and chatbots, consumers are able to access insurance products and information 24 hours a day. For example, consumers can use mobile apps to get immediate quotes and underwriting decisions from some insurers. In the past, consumers likely would have had to visit an insurance agent or fill out a lengthy application and wait much longer for an underwriting decision. And as previously discussed, some insurers allow their policyholders to submit claim information and photographs of damage through a mobile app without speaking with an agent.
  - **Increased consumer choice.** According to NAIC and an insurtech firm, consumers can benefit from the increased choice that comes from insurers using technology to offer additional products and services. For example, consumers obtain the ability to purchase insurance for certain time periods for certain items such as drones and action cameras, home sharing, or mile-based automobile insurance. NAIC and the insurtech firm said that some insurers that offer insurance to rideshare operators allow the policyholders to turn the coverage on when they are working and off when they are not. This can reduce premium rates for policyholders who only occasionally work as rideshare drivers.

According to the industry articles we reviewed and the stakeholders with whom we spoke, insurers' use of technology also has benefitted consumers by leading to the development of aggregator websites that bring together quotes from multiple insurers and allow consumers to comparison shop for insurance products. Some insurers said technology may soon give consumers the added ability to further customize their insurance policies by allowing them to select among various available coverages and terms and essentially create a policy that best suits their needs.

---

**NAIC and State Regulators Initiated Actions to Address Challenges That Stakeholders Said Could Affect Adoption of Technologies**

NAIC, state regulators, and others have initiated a number of actions intended to monitor and address industry and regulator concerns associated with insurtech, including any insurance rules and regulations that could affect insurers' adoption of technologies. These actions address challenges in areas including (1) evaluation of underwriting methodologies, (2) approvals for new insurance products, (3) customer notification methods and time frames, (4) anti-rebating laws, (5) cybersecurity, and (6) regulator skillsets and resources.

---

**NAIC and State Regulators Have Taken Actions Designed to Monitor Insurtech Concerns and Maintain Insurer Oversight and Consumer Protection**

NAIC and state regulators have initiated a number of actions intended to monitor concerns that regulations could affect insurers' adoption of innovative technologies while maintaining oversight of consumer protection issues. First, to monitor technology developments that may affect the state insurance regulatory framework and to develop regulatory guidance, as appropriate, NAIC created an Innovation and Technology Task Force. According to NAIC, this task force provides a forum for regulator education and discussion of innovation and technology in the insurance sector. For example, the task force has held discussions on the collection and use of data by insurers and state insurance regulators—as well as new products, services, and distribution platforms—to educate the regulators on how these developments affect consumer protection, privacy, insurer and producer oversight, marketplace dynamics, and the state-based insurance regulatory framework. In addition, the task force has held forums on emerging issues related to companies or licensees leveraging new technologies. Areas discussed included developing products for on-demand insurance purposes, reviewing new products and technologies affecting the insurance space, and potential implications for the state-based insurance regulatory structure.

In addition, in 2012 the EU-U.S. Insurance Dialogue Project was formed, in which EU and U.S. insurance regulators discuss emerging technology issues in the international insurance industry.<sup>37</sup> During the project's sixth

---

<sup>37</sup>The EU-U.S. Insurance Dialogue Project began in early 2012 as an initiative by the European Commission, the European Insurance and Occupational Pensions Authority, the Federal Insurance Office, and NAIC to enhance mutual understanding and cooperation between the European Union and the United States for the benefit of insurance consumers, business opportunity, and effective supervision. The Board of Governors of the Federal Reserve System has since joined the project.

forum in November 2018, the regulators and representatives from industry and consumer organizations discussed challenges and opportunities relating to issues including cyber risks, the use of big data, and AI. According to a project publication, the dialogue project enhanced mutual understanding of respective regulatory frameworks and initiatives between the United States and European Union, which will help ensure effective coordinated supervision of cross-border insurance groups for the benefit of policyholders.<sup>38</sup> In 2018, the project published an issues paper on big data.<sup>39</sup> The paper discusses data collection, portability, quality, and availability and how insurers and third parties use data in marketing, rating, underwriting, and claims handling. Future work by the project may include discussion of insurers' use of third-party vendors, disclosures to applicants, and insurers' use of AI models.

**NAIC and State Regulators Initiated Actions to Address Specific Insurtech Challenges**

NAIC and state regulators have initiated a number of actions intended to address industry and regulator concerns about certain insurance rules and regulations that a number of them said could affect insurers' adoption of technologies.

**Evaluating Underwriting Methodologies That Use Technology**

Stakeholders, including regulators, told us that regulators can face challenges in assessing new underwriting methodologies, such as those that use predictive analytics or AI. Reviewing predictive analytics can be a challenge for regulators because of the amount of data used to develop a model, the complexity of techniques, and limited staff resources (discussed in more detail later in this section). In addition, insurers employ different technological approaches, and their documentation and explanation of the methods and approaches differ. Finally, the data and models insurers use dynamically change and may have to be re-submitted for review even before regulators have an opportunity to review the original submission.

One state regulator and an industry stakeholder also told us that while an insurer may know the universe of factors from which an AI system pulls, the insurer may not know, or be able to describe for regulators, how the

<sup>38</sup>EU-U.S. Insurance Dialogue Project, *New Initiatives for 2017–2019: Focus Areas for 2018* (2018).

<sup>39</sup>EU-U.S. Insurance Dialogue Project, *Big Data Issue Paper* (Oct. 31, 2018).

---

system uses those factors to determine a premium rate. In turn, this may prevent regulators from understanding the system or validating the insurer's assertions about the system. For example, one state regulator told us that after presenting a rate scheme based on nontraditional factors, an insurer was unable to provide assurances or explanation to the regulator that the resulting premium rates were not unfairly discriminatory.

In 2018, NAIC's Casualty Actuarial and Statistical Task Force began developing a white paper on best practices state regulators can use when reviewing predictive models and analytics filed by insurers to justify rates and guidance they can use for their review of rate filings based on predictive models. NAIC officials told us the Casualty Actuarial and Statistical Task Force will receive comments on the white paper and then evaluate how to incorporate best practices into the Product Filing Review Handbook and recommend such changes to other NAIC working groups.

#### Approvals for New Products

Insurtech firms and other stakeholders told us that working through other regulatory processes, such as the insurance product filing and approval process, often can be inefficient and time consuming because insurers must file in every state in which they wish to sell a product and state requirements can vary. We have noted such difficulties in the insurance market in general.<sup>40</sup> These challenges can be exacerbated by rapid technological evolution in insurer products and risk models. In addition, some stakeholders noted that a lengthy product approval process can be challenging for technology-oriented products. For instance, an insurtech firm may develop a new product quickly to meet consumer demand but might not be able to get the product to market quickly. Some also said that products might become obsolete before the filing approval process was completed. Some stakeholders told us that such challenges can motivate insurtechs to sell insurance through nonadmitted insurers because such insurers have more freedom in altering and selling new products. As we have noted, doing so can bring risks for consumers.

In December 2017, the American Insurance Association proposed the Insurance Innovation Regulatory Variance or Waiver Act (Proposed

---

<sup>40</sup>See GAO, *Financial Regulation: Complex and Fragmented Structure Could Be Streamlined to Improve Effectiveness*, GAO-16-175 (Washington, D.C.: Feb. 25, 2016).

---

Model Law) to NAIC.<sup>41</sup> The proposed model law would urge allow regulators to create regulatory “sandboxes,” wherein certain regulatory requirements would be waived for insurers seeking to pilot innovative products.<sup>42</sup> Specifically, the proposed model law would authorize insurance regulators to grant variances, waivers, or no-action letters with respect to statutory or regulatory requirements that make it more difficult to introduce new insurance technologies, products, or services. Under the proposed model law, regulators also would be authorized to attach terms and conditions meant to protect consumers to such variances or waivers.<sup>43</sup> Some stakeholders with whom we spoke believed that regulatory sandboxes would not work in the U.S. state-based regulatory framework. For example, some stakeholders told us it would be inappropriate for a state to change legal or regulatory requirements for some but not all insurers or grant exceptions to laws passed by a state legislature to some insurers and not others, as it would no longer be a level playing field.

State regulators generally told us they believe that the current regulatory framework provides state regulators with enough flexibility to allow for technology-based innovation. Accordingly, some states have been promoting the use of innovation in the insurance industry by hosting technology sandboxes, where technology companies meet regularly with state regulators to improve companies’ knowledge of insurance regulations and also educate regulators about how the technologies work. According to stakeholders, these technology sandboxes are not the same as regulatory sandboxes that have been established in other nations, as they do not allow waivers of laws and regulations for insurtech companies to test their products.

#### Paper Notification Requirements

Insurtech firms we interviewed told us that regulations that require paper notifications and U.S. mail delivery for certain processes can make it difficult or more costly for them to offer products with features such as immediate underwriting or on-demand policies. For example, according to

<sup>41</sup>As of January 1, 2019, the American Insurance Association and the Property Casualty Insurers Association of America merged to form the American Property Casualty Insurance Association.

<sup>42</sup>Any model laws adopted by NAIC must be passed by individual state legislatures to be effective in a given state. State legislatures also may pass modified versions of model laws.

<sup>43</sup>As of May 2019, NAIC had not adopted the proposed model law.

---

insurers and other industry stakeholders, some state laws require that insurance policy cancellation notices be sent by U.S. mail rather than by email. One insurtech firm told us that it would be very costly to meet requirements for mail delivery of insurance policies and cancellation notices because they would have to set up another delivery mechanism (in addition to their electronic notification system).

Industry stakeholders also told us that certain laws and regulations that require a minimum period of time before a consumer-initiated policy cancellation takes effect can present challenges for products designed to allow consumers to immediately turn certain coverage on or off. For instance, if consumers used a mobile app to indicate they wanted to turn their automobile insurance coverage off temporarily, it could be unclear if this constituted an actual policy cancellation. Some stakeholders are concerned that states may require an insurance company to give the policyholder a written notice of cancellation at least 30 days before the end of the policy term. Similarly, industry stakeholders told us that some current state regulations could impede on-demand coverage because policies usually must indicate that coverage begins at 12:01 a.m. on the day after a policy is signed and approved. For instance, for on-demand policies that allow on/off subscription at the consumer's request, it can be unclear whether they are covered the minute that they initiate the coverage, or if they must wait until the following day for coverage to be effective.

According to NAIC, many states have taken steps to work within or modify existing laws and regulations to adapt to the increased use of technology in the insurance industry. For example, to address concerns that insurers are required to provide customers with a written, 30-day notice of a policy cancellation, NAIC conducted an analysis in 2018 that found that many states instead require "adequate" notice and that approximately 44 states allow notices to be provided electronically. However, some stakeholders in the insurance industry told us that state cancellation notice requirements are still a barrier to innovation.

#### Anti-Rebating Laws

According to industry stakeholders, many states have anti-rebating laws that generally prohibit insurers from providing consumers with anything of value as an inducement to purchase insurance. NAIC Model Law 880 states that unless expressly provided by law, no insurer may knowingly pay any rebate or incentive to an insured to induce them to purchase a

specific product.<sup>44</sup> Insurers, industry stakeholders, and regulators (including NAIC's Innovation and Technology Task Force) told us that anti-rebating laws can be a barrier to innovation because they could preclude insurers from offering devices that could be used to help insurers and consumers monitor risk. For example, if an insurer offered a policyholder free use of a telematic device (to help insurers collect real-time data and potentially help the policyholder make driving habits safer), it could be considered an inducement and violate anti-rebating laws. The same possibility exists if an insurer were to provide a policyholder with a device to monitor the operating conditions of a boiler to prevent potential water damage should a problem arise. As a result, anti-rebating laws may make it difficult for insurers to make use of certain technologies that could benefit both insurers and policyholders.<sup>45</sup>

In contrast to the consensus on the legitimacy of electronic communications, there is little consensus among states on addressing insurers' concern that anti-rebating laws are a barrier to innovation. According to NAIC, states vary widely on the types of items insurers are allowed to provide for free to customers, with some states having dollar limits on allowable items or allowing items that are specifically linked in a policy. In other cases, it is unclear what is allowable. At NAIC's fall 2018 meeting, participants noted that some of the NAIC bulletins related to the anti-rebating model law have not addressed whether technologies such as telematics that provide benefits to consumers are considered rebates. According to NAIC, others noted that states typically have taken the position that if a rebate or incentive reduces risk that is the most important issue for all parties involved. NAIC officials noted during the fall 2018 meeting that they will continue to monitor the issues involved.

#### Cybersecurity

NAIC adopted a model law and states have passed new laws governing cybersecurity and data protection to safeguard the increasing amount of personal data used by insurers. In 2017, NAIC approved the Insurance Data Security Model Law, which creates a legal framework for requiring insurance companies to operate cybersecurity programs.<sup>46</sup> The law

<sup>44</sup>NAT'L ASS'N OF INS. COMM'RS, UNFAIR TRADE PRACTICES ACT, MDL-880-1, § 4 (2004).

<sup>45</sup>Telematics and other devices typically involve investment on the part of insurers, consumers, or both. Cases of consumers having to make the investment to benefit from these technologies could raise the issue of insurance affordability, especially for low-income consumers, or potentially create a dual or segmented insurance market that could advantage high-income consumers.

<sup>46</sup>NAT'L ASS'N OF INS. COMM'RS, INSURANCE DATA SECURITY MODEL LAW, MDL-668-1 (2017).

---

outlines planned cybersecurity testing, creation of an information security program, and incident response plans for breach notification procedures. The NAIC model law is only a guideline until adopted by individual states, but NAIC noted that in 2018 and 2019, Michigan, Ohio, Mississippi, and Alabama adopted laws based on the NAIC model and additional states have pending legislation. In an October 2017 report, Treasury endorsed the model law and recommended that Congress consider preempting the states if the law were not adopted over the next 5 years.

At the state level, New York's Department of Financial Services noted it was the first state agency to establish cybersecurity regulations, which became effective March 1, 2017. In May 2018, South Carolina enacted the South Carolina Department of Insurance Data Security Act, which NAIC has characterized as an adoption of the model law. In December 2018, Michigan adopted a similar law. Separately, in June 2018 California passed a law giving consumers more control over their personal information.<sup>47</sup> California's law generally requires companies to report to customers, upon their request, the categories of personal information they collected about the customer, the business or commercial purpose for collecting and selling such personal information, and what categories of third parties received it.

#### Hiring and Retaining Staff with Technical Expertise

According to industry and regulatory stakeholders, the complexity and evolving nature of the models and approaches used by insurers may outpace the rate at which regulators can educate themselves on those models and approaches. For example, regulators trained in the current rating models may need to acquire new skills to understand and validate advanced and evolving models.

In addition, stakeholders told us that new technologies used by insurers can pose significant challenges to regulators partly because of the resource requirements. For instance, regulators and other stakeholders told us that regulators often do not have enough staff with technical expertise, such as data analytics skills, and find it challenging to hire and retain such staff due to limited resources.

NAIC has initiated actions to address concerns that state insurance regulators may not have staff with the knowledge or skill sets to address

<sup>47</sup>The California Consumer Privacy Act of 2018 is scheduled to take effect on January 1, 2020. See CAL. CIV. CODE § 1789.198(a) (2018).



---

more complex predictive models. For example, in 2018 NAIC management conducted a survey of states regarding the appropriate skills and potential resources NAIC membership may need to deal with big data. Subsequently, in April 2019, NAIC management made recommendations to its Big Data Working Group to hire a technical staff resource to provide technical support for state insurance regulators in the review of actuarial models; develop a tool for state insurance departments to share information on model reviews; and develop a training and education program. NAIC officials told us they also plan to develop a white paper to provide state regulators with guidance on the use of chatbots and AI in the distribution of insurance and the regulatory supervision of these technologies.

As many of the regulatory initiatives that NAIC and states have undertaken to address challenges associated with the implementation of new technologies are under development (or recently developed), the impact of these actions on innovation and consumer protection is unknown. It will be important for NAIC and state insurance regulators, as well as the Federal Insurance Office, to continue monitoring developments in these areas.

---

### Agency and Third Party Comments

We provided a draft of this report to Treasury and NAIC for review and comment. Treasury and NAIC provided technical comments that we incorporated as appropriate.

We are sending copies of this report to the appropriate congressional committees, the Secretary of the Treasury, the Chief Executive Officer of the National Association of Insurance Commissioners, and other interested parties. In addition, the report is available at no charge on the GAO website at <http://www.gao.gov>.

---

If you or your staff have any questions concerning this report, please contact me at (202) 512-8678 or [ortiza@gao.gov](mailto:ortiza@gao.gov). Contact points for our Offices of Congressional Relations and Public Affairs may be found on the last page of this report. GAO staff who made major contributions to this report are listed in appendix II.



Anna Maria Ortiz  
Acting Director, Financial Markets and  
Community Investment

---

*List of Requesters*

The Honorable Patrick McHenry  
Ranking Member  
Financial Services Committee  
House of Representatives

The Honorable Rick Allen  
House of Representatives

The Honorable Earl L. "Buddy" Carter  
House of Representatives

The Honorable Michael McCaul  
House of Representatives

The Honorable Scott Tipton  
House of Representatives

The Honorable Ann Wagner  
House of Representatives

The Honorable Rob Woodall  
House of Representatives

---

## Appendix I: Objectives, Scope, and Methodology

---

This report (1) identifies uses of technologies and the benefits and challenges they might present for insurers and their customers, and (2) discusses what stakeholders identified as key challenges that could affect the adoption of new technologies, and actions that have been taken to address those challenges.

While insurance technology (insurtech) does not have a standard definition, for the purposes of this report we defined it as the use of emerging technologies by insurance companies. We focused on insurtech activities in the property/casualty and life sectors of the U.S. insurance market, including information on personal and commercial insurance where available. We did not include the health insurance sector in our scope because of significant differences between that sector and the property/casualty and life insurance sectors in terms of the types of products offered and the methods by which they are sold and regulated.

To identify technologies being used in the insurance industry and gain insights about their (potential) benefits and challenges for insurers and customers, we conducted a literature review of scholarly and peer-reviewed material, trade and industry articles, government reports, conference papers, general news, association, nonprofit, and think tank publications, hearings and transcripts, and working papers that described these technologies and their uses. We conducted searches of the ProQuest and HeinOnline databases to identify studies published from January 2015 through June 2018 that were relevant to our research objectives. Because insurtech is a fairly new field, we found few academic publications related to our objectives. We also conducted background research for examples of technologies being used in the insurance industry and their associated benefits and challenges.

We also conducted semi-structured interviews with cognizant stakeholders and reviewed documents provided by them to obtain information on and descriptions of current, in-development, and potential future uses of existing or new technology in the insurance industry. We also obtained their views on the benefits and challenges experienced or expected by insurance companies as well as the (potential) benefits and challenges for consumers. We conducted more than 35 interviews with representatives of regulatory organizations, including the Federal Insurance Office; National Association of Insurance Commissioners (NAIC); state insurance regulators in Arizona, California, Connecticut, and Michigan; and the National Council of Insurance Legislators. We also interviewed three academics, representatives of one consumer group, 13 traditional insurance and reinsurance providers and industry associations,

---

two actuarial professional associations, four consulting groups, two law firms in the field, and seven insurtech firms. We identified potential interviewees by conducting internet research, reviewing literature search results, and reviewing recommended interviewees from our initial interviews. We selected interviewees based on their relevance to the scope of our review. Based on our literature review and interviews with stakeholders, we identified seven recently used and emerging technologies in the insurance industry: (1) mobile applications; (2) artificial intelligence (AI), algorithms, and machine learning; (3) big data; (4) internet of things; (5) blockchain/ distributed ledger technology and smart contracts; (6) drones; and (7) telematics.

To obtain information about challenges that could affect the adoption of innovative technologies, we identified relevant laws and regulations pertaining to insurance technology innovation by reviewing prior GAO reports on financial regulation, interviewing regulators and industry participants, and analyzing relevant documents, including relevant NAIC model laws and state laws and regulations. We also conducted semi-structured interviews with and reviewed documents provided by the key stakeholders identified in the first objective to identify (1) any actions NAIC and selected state insurance regulators were taking on new insurance technologies, and what challenges, if any, insurers' use of new technologies creates for regulators; (2) what is known about the impact of any actions taken by NAIC and state insurance regulators on innovation among insurance companies and on consumer protection; and (3) stakeholders' views on the applicability of foreign regulatory actions for U.S. insurtech markets.

We conducted this performance audit from April 2018 to June 2019 in accordance with generally accepted government auditing standards. Those standards require that we plan and perform the audit to obtain sufficient, appropriate evidence to provide a reasonable basis for our findings and conclusions based on our audit objectives. We believe that the evidence obtained provides a reasonable basis for our findings and conclusions based on our audit objectives.

---

## Appendix II: GAO Contact and Staff Acknowledgments

---

---

### GAO Contact

Anna Maria Ortiz, (202) 512-8678 or ortiza@gao.gov

---

### Staff Acknowledgments

In addition to the contact named above, Patrick Ward (Assistant Director), Deena Richart (Analyst in Charge), Gina Hoover, Hadley Nobles, Akiko Ohnuma, and Tyler Spunaugle made key contributions to this report. Also contributing were Ernei W. Li, Barbara Roesmann, Jena Y. Sinkfield, Frank Todisco, and Helen Tulloch.

<b>GAO's Mission</b>	The Government Accountability Office, the audit, evaluation, and investigative arm of Congress, exists to support Congress in meeting its constitutional responsibilities and to help improve the performance and accountability of the federal government for the American people. GAO examines the use of public funds; evaluates federal programs and policies; and provides analyses, recommendations, and other assistance to help Congress make informed oversight, policy, and funding decisions. GAO's commitment to good government is reflected in its core values of accountability, integrity, and reliability.
<b>Obtaining Copies of GAO Reports and Testimony</b>	The fastest and easiest way to obtain copies of GAO documents at no cost is through GAO's website ( <a href="https://www.gao.gov">https://www.gao.gov</a> ). Each weekday afternoon, GAO posts on its website newly released reports, testimony, and correspondence. To have GAO e-mail you a list of newly posted products, go to <a href="https://www.gao.gov">https://www.gao.gov</a> and select "E-mail Updates."
<b>Order by Phone</b>	The price of each GAO publication reflects GAO's actual cost of production and distribution and depends on the number of pages in the publication and whether the publication is printed in color or black and white. Pricing and ordering information is posted on GAO's website, <a href="https://www.gao.gov/ordering.htm">https://www.gao.gov/ordering.htm</a> .  Place orders by calling (202) 512-6000, toll free (866) 801-7077, or TDD (202) 512-2537.  Orders may be paid for using American Express, Discover Card, MasterCard, Visa, check, or money order. Call for additional information.
<b>Connect with GAO</b>	Connect with GAO on Facebook, Flickr, Twitter, and YouTube. Subscribe to our RSS Feeds or E-mail Updates. Listen to our Podcasts. Visit GAO on the web at <a href="https://www.gao.gov">https://www.gao.gov</a> .
<b>To Report Fraud, Waste, and Abuse in Federal Programs</b>	Contact FraudNet: Website: <a href="https://www.gao.gov/fraudnet/fraudnet.htm">https://www.gao.gov/fraudnet/fraudnet.htm</a> Automated answering system: (800) 424-5454 or (202) 512-7700
<b>Congressional Relations</b>	Orice Williams Brown, Managing Director, <a href="mailto:WilliamsO@gao.gov">WilliamsO@gao.gov</a> , (202) 512-4400, U.S. Government Accountability Office, 441 G Street NW, Room 7125, Washington, DC 20548
<b>Public Affairs</b>	Chuck Young, Managing Director, <a href="mailto:youngc1@gao.gov">youngc1@gao.gov</a> , (202) 512-4800, U.S. Government Accountability Office, 441 G Street NW, Room 7149, Washington, DC 20548
<b>Strategic Planning and External Liaison</b>	James-Christian Blockwood, Managing Director, <a href="mailto:spel@gao.gov">spel@gao.gov</a> , (202) 512-4707, U.S. Government Accountability Office, 441 G Street NW, Room 7814, Washington, DC 20548



Please Print on Recycled Paper.



In our original post, "Why Lenders Shouldn't 'Just Use SHAP' To Explain Machine Learning Credit Models," we raised several issues lenders face when trying to explain their machine learning (ML) credit models with "just" the popular open-source package SHAP. Computer scientist Scott Lundberg, one of the primary authors of SHAP, has been gracious with his time and responded to many of our questions and comments on his github page. We agree with most of what Scott says, but we want to clarify our position on a few issues.

Mainly we want to clarify that one must take care to ensure the margin space/score space transition is handled correctly so that the explanations generated are an accurate assessment of how a model-based decision is made. Although this may seem like a technical nit, this is keenly important both for generating accurate adverse action reasons and for doing adequate fair lending analysis. It's also why we spent so much time and care to develop the core explainability math inside ZAML: So that lenders can accurately assess the reasons for a model-based decision in score space, without requiring unrealistic assumptions such as variable independence, or that missing values are missing completely at random.

Let's first consider the process of generating adverse action reasons. Under the Fair Credit Reporting Act, when an applicant is denied a loan, the lender must respond with an adverse action letter notifying the applicant of the denial of credit and listing the top five reasons they were denied. These top five reasons are provided for two purposes: first, so that the applicant can see if there are any errors in the information provided to the lender, and, second, so that the applicant can figure out what to do to raise their likelihood of being approved in the future.



Scott calls out one of our statements in his response and raises an important issue. We say:

If you compute the set of weighted key factors in margin space, you'll get a very different set of factors and weights than if you compute them in score space, which is where banks derive their top five explanations for rejecting a borrower.

And then Scott says:

[The additivity of margin space] is the same reason I often encourage people to think about explanations in margin space and not just use probability space... I talked with Zest about this and it seems like it could be better to do the explanations in score space for finance, but that conclusion is not 100% clear cut.

We respectfully disagree.

It is important to understand why the distinction between margin space and score space matters as you consider a method for generating adverse action reasons. The lender needs to tell the applicant the five most important reasons they were denied -- but what does "most important" actually mean? Lenders approve a fixed fraction, say  $k$  percent, of all applications -- that is, they look at the highest ranked  $k$  percent of the applications they receive. That's a ranking problem, so the natural score one would wish to discuss is the 'score space' in which the distribution of outputs would be uniform. The applicant does not benefit from knowing what five values would improve the difference between their marginal score and an abstract threshold which corresponds to the desired rank and so, lenders are required to disclose which five factors would most reduce the difference between the rank of their current application and the threshold acceptance rank.

Thus, an explainer powering adverse action reasons needs to provide accurate "score space" reasons. While reasoning in margin space is convenient, ultimately financial services applications of ML explainability require something different. This issue isn't limited to finance applications, it applies to any modeling problem which requires a probability assignment.

As Scott rightly points out, score space explanations are available from the various explainers within SHAP, but the score space explanations generated by SHAP are an approximation.

Per the SHAP documentation, TreeExplainer probability outputs (which are required for applications like generating adverse action reasons in credit, as explained above) are only available when the variables upon which the model depends are statistically independent. This just isn't realistic for a credit risk scenario, in which many of the variables are dependent on each other and any adequate risk model must capture that fact.

To see how this is problematic in a domain like credit risk, consider two common input variables: total debt to income (DTI) and revolving credit utilization. These variables are not independent: revolving credit utilization is a subset of total debt and therefore a component of total debt to income. Imposing an independence assumption means you can't explain one of machine learning's greatest strengths -- the ability for ML models to capture interactions among variables.

GradientExplainer, the part of SHAP that can be used for explaining continuous models like neural networks, assumes the input features are independent as well. From the SHAP README page, "If we approximate the model with a linear function between each background data sample and the current input to be explained, and we assume the input features are independent then expected gradients will compute approximate SHAP values." Unfortunately, feature independence is not a safe assumption in credit risk, and, as far as we can tell, not a necessary assumption to make here.

KernelExplainer, a more sophisticated implementation of LIME that computes Shapley values, suffers from a different flaw. It makes the assumption that missing values can be filled with an average, as though they are missing at random. This is not valid in most real-world datasets, in general, and in datasets arising from financial services applications, in particular. To see why, consider the meaning of a common credit risk variable, an applicant's credit score. That someone has a missing credit score provides information about the distribution of the other signals corresponding to the application. A missing credit score usually indicates a lack of credit history, which, in turn, suggests that many of the variables associated with a credit history will be differently distributed for that population than they would be for the population in general. This, in turn, means that creating a population of artificial completions for items from which the credit score is omitted becomes more complicated than it may at first seem. KernelExplainer will erroneously impute a spectrum of values drawn from the population as a whole. This will inevitably lead to providing the wrong adverse action reasons to the consumer.

The same issues that come up when considering how to generate adverse action reasons also come up when considering fairness. The Equal Credit Opportunity Act, requires lenders to make decisions without regard to race and ethnicity, gender, and other protected statuses. The act further requires the identification of disparity in approval rate and pricing terms, and that if disparate impact exists, e.g., that the approval rate or pricing for applicants within a protected class is unfavorable when compared to the unprotected baseline, that the lender quantify and understand the drivers of such disparity, and mitigate them or document them accurately. The law provides for stiff enforcement penalties. You can easily see how quantifying the drivers of a difference in approval rate also requires reasoning in “score space”. The factors that drive an applicant to be approved or not are based on the rank ordering of the applicant’s credit risk.

The point of all this is not to discredit the work of our esteemed colleague Dr. Lundberg. (BTW, congratulations, Scott, on successfully defending your dissertation and receiving your Ph.D.) Clearly, a team of data scientists with enough time and care can make the improvements and accommodations required to use open source packages like SHAP safely in financial services.

But we think it’s important to understand limitations, assumptions, and safe operating parameters before applying algorithms and techniques, especially in a domain like credit, where significant life-changing events are at stake, such as the ability to own a home or to get financing for a car you need to drive to work. We spent significant resources to develop the core explainability math inside ZAML so that lenders can accurately assess the reasons for a model-based decision, and thereby safely make use of the significantly better predictive power offered by modern machine learning techniques.



ZESTFINANCE.COM

