Statement by

Meredith Broussard

Associate Professor, New York University

Affiliate Faculty, NYU Center for Data Science

before the

Task Force on Artificial Intelligence

of the

Committee on Financial Services

U.S. House of Representatives

October 18, 2019

Congressman Foster, Ranking Member Hill, thank you. It is an honor to be asked to testify today regarding "AI and the Evolution of Cloud Computing: Evaluating How Financial Data is Stored, Protected, and Maintained by Cloud Providers."

I'd like to speak about cybersecurity and the cloud in general, and then offer an opinion on strategies to improve cybersecurity in financial technology. I am an associate professor at the Arthur L. Carter Journalism Institute of New York University and an affiliate of the NYU Center for Data Science. I started my career as a computer scientist, working at AT&T Bell Labs. I later worked on Wall Street at Prudential Securities, and at a tech startup that did secure document management for complex financial transactions. I then switched to journalism, where now I teach data journalism. Data journalism can be described as the practice of finding stories in numbers, and using numbers to tell stories. I do a specific kind of data journalism called algorithmic accountability reporting. Increasingly, algorithms are being used to make decisions on behalf of citizens; algorithmic accountability reporters hold algorithms accountable, as well as the people who make algorithms. My academic research focuses on artificial intelligence for investigative reporting. In other words, I build AI tools in order to do accountability reporting in the digital world. I am also the author of a book called *Artificial Unintelligence: How Computers Misunderstand the World,* which explores the inner workings and outer limits of technology.[1] Understanding the outer limits of tech is essential for making decisions about cloud computing.

A great deal of cybersecurity discourse focuses on defense from attacks. In understanding regulation like the Strengthening Cybersecurity for the Financial Sector Act, I would argue that it is important to think through the interplay between humans and technical systems in addition to

[1] Broussard, *Artificial Unintelligence*.

considering the usual attack vectors. Effective regulation depends on effective communication as well as technical competence.

I was asked to address several points:

(1) The types of cloud deployments and services and how they are used within financial services;

(2) How AI could help automate the various components within a cloud infrastructure;

(3) Best practices for regulatory examiners when engaging cloud service providers and other related third-parties utilized by their regulated entities;

(4) Ways to combat systemic risks, strengthen consumer privacy, and decrease the risks associated with data breaches;

(5) Regulatory and legislative proposals to strengthen federal oversight of cloud infrastructures utilized by financial institutions.

I will address each of these points in turn.

1. **Types of cloud deployments**

There is a lot of confusion about what the cloud is. The most common expression among computer scientists is, "The cloud is someone else's computer." A program that *runs in the cloud* means that the program is running on someone else's computer. Data *stored in the cloud* means data stored on someone else's computer. The cloud is a wonderful metaphor, but practically speaking, the cloud just means "a different computer, probably located with thousands of other computers in a large warehouse in the tristate area." We can even pinpoint exactly where those computers are. Amazon Web Services, which controls 48% of the cloud computing market, has

four major data centers in the United States.[2] These are located in Northern Virginia, Ohio, Oregon, and Northern California. These data centers are also called server farms.

Amazon, Google, Microsoft, and Alibaba together control 76% of the worldwide market for cloud computing. These companies own or lease server farms, and inside the server farm buildings they maintain thousands of physical computers. Some of those computers are dedicated to a single purpose or client; some of those computers are shared by multiple clients. The hearing memo outlines four different types of clouds: public clouds, private clouds, community clouds, and hybrid clouds.

One example of a private cloud is the AWS GovCloud, the secure set of servers that hosts data and programs for DHS, Treasury, DoD, Cloud.gov, and other agencies. The computers that power the AWS GovCloud are physically located in buildings on the East Coast (in Virginia) and the West Coast. Those locations are powered by electricity. If the power goes out or those locations flood or are affected by extreme weather, the network will be compromised. All of the GovCloud servers are connected by underground wires to the global network we refer to as the Internet. These wires are also subject to physical constraints. They may be dug up, they may wear out, they may flood, they may be affected by earthquakes, or they may be vulnerable to any other physical threat. It is useful to understand the physical reality of the cloud in order to think about security for cloud computers. Cybersecurity threats arise from the natural world as well as from the humans who seek to penetrate secure systems.

Despite analysts' predictions, it is unlikely that *all* bank operations will eventually move to the cloud. It is more likely that the current trend will continue. Some operations will be most

[2] https://aws.amazon.com/about-aws/global-infrastructure/regional-product-services/

efficient if done locally on one person's computer; other operations will be most efficient if done remotely on a more powerful cloud computer. Effective tech policy requires using the right tool for the task.

This gets us to the people part of the system. To get a good picture of how effective companies are at using cloud computing, it helps to hear from the IT professionals who manage local and cloud computers. A 2014 Ponemon Institute survey asked IT professionals to rate their organizations' effectiveness in securing data and applications used in the cloud.[3] Most (51%) rated their organizations as low in effectiveness. Based on their lack of confidence, these IT professionals also said the likelihood of a data breach in the cloud is increased.

Most banks currently do an effective mix of using the cloud for less-confidential services like email and HR, and keep their most secure information (consumer accounts, commercial accounts, customer data) in data centers that they manage themselves. Moving sensitive data to the cloud would require thoroughly vetting the cloud provider beforehand. In the same Ponemon survey, 62% of respondents did not agree or were unsure that cloud services were thoroughly vetted before deployment. Sixty-nine percent believed that their organizations failed to be proactive in assessing information that was too sensitive to be stored in the cloud.

If IT professionals have so little faith in their own organizations, and we know there is a high demand but low supply of IT professionals who are experts in cybersecurity, it seems that more regulation and oversight will help protect bank operations in the cloud.

2. How AI could help automate components of the cloud

[3] http://go.netskope.com/rs/netskope/images/Ponemon-DataBreach-CloudMultiplierEffect-June2014.pdf

Artificial intelligence, like cloud computing, is widely misunderstood. Hollywood images of AI, like the Terminator or Commander Data from Star Trek, are what most people think of when they think of AI. These Hollywood images are delightful, but they are not real. AI is best understood as a branch of computer science, the same way that algebra is a branch of mathematics. Inside AI, there are other branches: machine learning, expert systems, and natural language processing are just a few of them. However, machine learning is the most popular kind of AI in business right now. It is so popular that there has been linguistic confusion. When people say "I am using AI for my business," what they usually mean is, "I am using machine learning for my business." Machine learning is another misleading name; it sounds like the computer has sentience or agency. It does not. Machine learning is math. It's computational statistics on steroids.

Banks are using machine learning to help make decisions about things like who qualifies for a mortgage. When we use machine learning, the first thing we do is take some data and construct a machine learning model to predict a certain value in the data. I describe the process in my book *Artificial Unintelligence:*

> There are three general types of machine learning: supervised learning, unsupervised learning, and reinforcement learning. Here are definitions of each from a widely used textbook called *Artificial Intelligence: A Modern Approach* by UC Berkeley professor Stuart Russell and Google's director of research, Peter Norvig:[4]
>
> > **Supervised learning:** The computer is presented with example inputs and their desired outputs, given by a "teacher," and the goal is to learn a general rule that maps inputs to outputs.
> > **Unsupervised learning:** No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means toward an end (feature learning).
> > **Reinforcement learning:** A computer program interacts with a dynamic environment in which it must perform a certain goal (such as driving a vehicle or

[4] Russell and Norvig, *Artificial Intelligence*.

> playing a game against an opponent). The program is provided feedback in terms of rewards and punishments as it navigates its problem space.
>
> Supervised learning is the most straightforward. The machine is provided with the training data and labeled outputs. We essentially tell the machine what we want to find, then fine-tune the model until we get the machine to predict what we know to be true.
>
> All three kinds of machine learning depend on *training data*, known datasets for practicing and tuning the machine-learning model. Let's say that my training data is a dataset of one hundred thousand credit card company customers. The dataset contains the data you would expect a credit card company to have for a person: name, age, address, credit score, interest rate, account balance, name(s) of any joint signers on the account, a list of charges, and a record of payment amounts and dates. Let's say that we want the ML model to predict who is likely to pay their bill late. We want to find these people because every time someone pays a bill late, the interest rate on the account increases, which means the credit card company makes more money on interest charges. The training data has a column that indicates who in this group of one hundred thousand has paid their bills late.
>
> We split the training data into two groups of fifty thousand names each: the training set and the test data. Then, we run a machine-learning algorithm against the training set to construct a model, a black box, that predicts what we already know. We can then apply the model to the test data and see the model's prediction for which customers are likely to pay late. Finally, we compare the model's prediction to what we know is true—the customers in the test data who actually paid late. This gives us a score that measures the model's precision and recall. If we as model makers decide that the model's precision/recall score is high enough, we can deploy the model on real customers.

It seems very attractive to create a model based on data, and then use the model to make decisions. It might also seem like it would be cheaper to use a machine learning model than to use a human staffer for making decisions. The problem is, almost every dataset is biased. Machine learning models discriminate by default.[5] If I have a dataset of people who have gotten mortgages, the data will be tainted by the history of redlining and residential segregation in the United States. Facial recognition systems are trained on datasets of faces; the bias is baked in

[5] Benjamin, *Race after Technology*.

based on who is in the dataset. In their groundbreaking "Gender Shades" project, Joy Buolamwini and Timnit Gebru showed that the most commonly used facial recognition systems are good at recognizing people with light skin, but fail to recognize darker skinned people.[6] In part, this is because the people who made the facial recognition technology (who are probably men with light skin, based on the dominant demographics of the tech industry) either failed to notice or failed to care that the tech failed for people with darker skin.

Good suggestions have been made by Cathy O'Neil, Jack Balkin, and others[7] about how to audit algorithms and machine learning models. If banks run machine learning models on sensitive data on-premises, it makes it easier to audit them, as will certainly be required in the near future. If the models run in the cloud, it makes it slightly more difficult to audit them because of the different levels of visibility available in cloud environments.

The issue here is not whether banks should use AI in the cloud or on-premises. When a bank uses AI, we should ask what the AI is used for, plus how it is used, what kind of AI is used, what specific data is used to train the model, and what specific data is used to make decisions after the model is trained. These questions need to be answered in addition to the basic questions of where the AI program and its associated data will run and be stored. One option: the questions above could be answered in plain language, and this information could be communicated as part of the regulatory examination.

3. Best practices for regulatory examiners

[6] Buolamwini and Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification."
[7] O'Neil, *Weapons of Math Destruction*; Balkin, "Information Fiduciaries and the First Amendment."

Every modern organization uses a combination of in-house (on-prem) and cloud (remote) resources. When a bank contracts with a Cloud Service Provider (CSP) like Amazon Web Services, they are required to do their due diligence just as they would with any outside vendor.

Regulators have been overseeing banks for hundreds of years, but cloud computing oversight is relatively new. Additional guidance is appropriate. There should be abundant oversight of CSPs, including regular on-site visits and more documentation of physical and virtual security practices. More regulators are likely needed to staff the regulatory positions. Cybersecurity is so complex that it is impossible for a single person to be an expert in both physical and virtual security. A team approach is necessary.

Another important thing to be aware of is the cultural conflict between tech and finance. In the tech world, which is the world that trains cybersecurity people and cloud computing people and IT people and AI people, nobody talks about regulatory compliance. The ACM, the membership organization that provides guidelines for computer science education globally, only developed ethical guidelines in the past two years. Regulatory compliance is not part of core computer science education. "Compliance" is a word most software developers don't hear unless they work in finance. The "move fast and break things" ethos is diametrically opposed to the mindset of compliance. It thus doesn't surprise me that in April 2019, when federal examiners visited the AWS site in Virginia, they did not notice the Capital One data breach in which 100 million customers' data was stolen.[8] "The examiners were greeted warily at the Amazon offices," the *Wall Street Journal* wrote of the visit. "Chaperoned by an Amazon employee, they were allowed to review certain documents on Amazon laptops, but not allowed to take anything

[8] https://www.wsj.com/articles/fed-examined-amazons-cloud-in-new-scrutiny-for-tech-11564693812

with them, some of the people said." The Amazon IT workers were behaving in a way that is culturally appropriate for their workplace: trying to keep data and procedures secret to protect what they perceive as Amazon's assets. Amazon's IT staff is not necessarily trained in financial industry compliance the way a bank's IT staff would be. One option is to require cloud providers' staff to be trained in financial regulatory requirements, just as staff who administer clinical trials must be trained in HIPAA security and procedures. The legislation could require training, then the specific training could be implemented at the level of, say, an industry association so the training could keep up with the pace of change in the technology world.

4. **Ways to combat systemic risks, including data breaches**

Liability in cyberspace should mirror liability in the physical world. A server farm is a bit like a hotel in that each bank is renting secure space (a hotel room in this analogy) from the server farm. If you suffer injury at a hotel, because of the hotel's negligence or an accident, the hotel is liable. In its FAQ about the 2019 data breach, Capital One writes: "Like many companies, we have a Responsible Disclosure Program which provides an avenue for ethical security researchers to report vulnerabilities directly to us."[9] It is ethical for a bank to constantly monitor its systems for vulnerabilities, just as it is for CSPs.

5. **Regulatory and legislative proposals to strengthen federal oversight of cloud infrastructures utilized by financial institutions**

[9] https://www.capitalone.com/facts2019/2/

This testimony has been prepared for a hearing entitled "AI and the Evolution of Cloud Computing: Evaluating How Financial Data is Stored, Protected, and Maintained by Cloud Providers." I am under the impression that the Committee is considering the proposed Strengthening Cybersecurity Act. The act proposes to extend existing regulation. Already, bank regulators have the power to oversee and examine third-party vendors for banks. This act proposes to extend this regulatory authority to the NCUA and FHFA so that credit unions, Fannie Mae, Freddie Mac, and FHLBs are similarly protected.

I would argue in favor of this or a similar act because additional protections and constant oversight are needed to protect Americans' financial information in the digital sphere, just as protections and oversight are needed in physical banks. We should approach the safety of our financial data with at least the same level of care that we devote to food safety in restaurants. Thinking about the physical reality of AI and cloud computing is important so that we don't make the mistake of thinking that tech is something different or special that demands exceptional treatment.

Citizens' rights and human rights must be protected online as they are offline. Effective regulation of financial technology in the cloud will allow us to foster innovation and competition while protecting consumers. CSP staff should be trained in regulatory compliance in order to serve bank customers, and more plain language explanations of complex AI technology should be made available by banks and CSPs so that regulators can adequately monitor the health of financial technology systems.

Thank you for the opportunity to contribute to this hearing. I look forward to answering your questions.

References

Balkin, Jack M. “Information Fiduciaries and the First Amendment.” *UC Davis Law Review, Vol. 49, No. 4, 2016* 49, no. 4 (February 3, 2016). https://ssrn.com/abstract=2675270.

Benjamin, Ruha. *Race after Technology: Abolitionist Tools for the New Jim Code*. Medford, MA: Polity, 2019.

Broussard, Meredith. *Artificial Unintelligence: How Computers Misunderstand the World*. MIT Press, 2018.

Buolamwini, Joy, and Timnit Gebru. “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.” In *Proceedings of Machine Learning Research*, 81:1–15, 2018.

O’Neil, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. First edition. New York: Crown, 2016.

Russell, Stuart J, and Peter Norvig. *Artificial Intelligence: A Modern Approach*, 2015.