House of Representatives Committee on Financial Services
Task Force on Artificial Intelligence hearing

"Beyond I, Robot: Ethics, Artificial Intelligence, and the Digital Age"
October 13, 2021

Prepared Testimony of
Meg King
Director, Science and Technology Innovation Program
The Wilson Center

The application of AI is already having a profound effect on how the world works. As with any technological evolution, the benefits of AI come with associated costs and risks. Focusing only on the benefits in a particular industry misses the nuances of the potentials and pitfalls of this advance. As the title of this hearing makes clear, the risks are more subtle than a dystopic future populated by robot overlords.

To help the Committee understand the risks to any industry, and in particular the financial services industry, I will focus my remarks on the nature of AI generally to understand the environment in which creation is occurring.

Assessing current ethical AI frameworks

Today, there are not significant incentives for the private sector to include ethics directly in the development process. At the current pace of advancement, companies cannot afford to develop slowly – or a competitor might be able to bring a similar product to market faster.

Largely due to consumer trust concerns, international intergovernmental organizations, regions and private companies have all begun to issue ethical frameworks for AI. Most are very vague principles, with little guidance as to application. Two that this Committee should pay close attention to are those of the Organization for Economic Cooperation and Development (OECD) and the European Commission.

Adopted in 2019, the OECD's AI Principles aim to "promote use of AI that is innovative and trustworthy and that respects human rights and democratic values." Its five principles encourage inclusive growth, sustainable development and well-being; human-centered values and fairness; transparency and explainability; robustness, security and safety; and accountability. Perhaps most relevant to this Committee are the process and technical guidelines – ranging from pinpointing new research to making available software advances – that OECD is in the process of identifying and which will become part of a publicly available interactive tool for developers and policymakers alike.

Similarly, the European Commission issued "Ethics Guidelines for Trustworthy AI," which include 7 requirements: human agency and oversight; technical robustness and safety; privacy and data governance; transparency: diversity, non-discrimination and fairness; societal and environmental well-being; and accountability. This spring, European regulators announced a risk-based plan to prevent the sale of AI systems to the region with use-cases deemed too dangerous to safety or fundamental citizen rights (e.g. social credit scoring systems) and transparency requirements for others, including biometric identification and chatbots. Chatbots in particular are expected to have a significant impact on the financial services industry as many companies see value in customer service process improvement and the prospect of gaining more insight into customer needs in order to sell more financial products.

Determining that AI systems do not all pose equal risk of harm and should be evaluated based on level of risk to consumers, a new European AI Board will be created to manage compliance (e.g. record checks) and enforcement (e.g. financial penalties). As regulators ask developers more questions about the ethics of their AI systems, they have the potential to slow the process, which could cost businesses money. However, if ethical concerns are identified too late in the development process, companies could face considerable financial loss if problems cannot be addressed.

How to make AI ethics practical

No ethical AI framework should be static as AI systems will continue to evolve as will our interaction with them. Key components, however, should be consistent, and that list, specifically for the financial sector, should include: *explainability*, *data inputs*, *testing*, and *system lifecycle*.

As the Committee considers ethical AI frameworks, one of the near-term questions to ask about systems you will encounter in your oversight is how will COVID-19 pandemic experiences factor into these systems?

Explainability

Also known as XAI, this is a method to ask questions about the outcomes of AI systems and how they achieved them. XAI helps developers and policymakers identify problems and failures in AI systems, identify possible sources of bias, and help users access explanations. There are a number of techniques available as well as open source tools like InterpretML and AI Explainability 360, which make these techniques more accessible.

Questions can include:

- Why was the AI system developed?
- What are the outcomes intended?
- How can it fail?
- How can we report and correct errors?

There are various techniques to accomplish this process today, and going forward, the goal will be to design AI systems that explain their logic, identify strengths and weaknesses and provide prediction for how they will behave in the future. At least for now, the limits of human intelligence limit the evolution of more ethical AI systems – even those that learn without human intervention.

In the financial sector, explainability will become critical as predictive models increasingly perform calculations during live transactions—for example, to evaluate the risk or opportunity of offering a financial product or specific transaction to a customer. Establishing a clear process for explainability in the first place will be critical to address flaws identified in these real-time systems, and should be an area of focus for the Committee. Additionally, producing policies for how these systems will be used and in what context will be helpful.

Data inputs
Without context, data pulled from a mix of public and private records, including credit score, banking activity, social media, web browsing, and mobile application use, can produce inaccurate results and discriminate access to financial products.

We have all heard horror stories of individuals who lost jobs because of the pandemic. In a hypothetical scenario, that person could be denied unemployment benefits because of incorrect data, causing delay or inability to pay rent. If a landlord sues, even if that lawsuit does not succeed because of a federal moratorium, it becomes part of public record, which could be used to decline future rental applications. Meanwhile, due to the data provided around these circumstances, this person is served ads for lower paying jobs and the same data about late rent payments could make it harder to secure financing for a car, necessary to transport the individual to a new lower paying job.

The cycle could continue without intervention or a redress process. In the longer term, investment advice, insurance pricing and customer support may be challenged if inputs are not equal. One promising possibility to address the data input problem might be to synthesize artificial financial data to correct for inaccurate or biased historical data (Efimov, Xu, Kong, Nefedov, Anandakrishnan, 2020).

Testing
While quality assurance is part of most development processes, there are currently no enforceable standards for testing AI systems. Therefore, testing is uneven at best.

Where the Committee can provide guidance and support to the private sector will be in the testing process. Developers will need more time and resources to involve those most likely to be affected by the AI systems being created for them.

Lifecycle of systems
Increasingly, users are far removed from AI system developers. Additionally, the software procurement process in the private sector is rarely transparent. Carefully assessing the growing field of MLOps (machine learning operations) and identifying ways to participate will be useful. Assessing the lifecycle of AI systems will be particularly important in gaining early warning about the possibility and risk of “black swans” in the financial system that could occur due to failure modes in AI systems.

Failure modes

AI breaks, often in unpredictable ways at unpredictable times.

Participants in the Wilson Center's Artificial Intelligence Lab trainings for Congressional and Executive branch staff have seen AI function spectacularly – some using a deep learning language model to produce the first ever AI-drafted legislation – as well as fail, when a particular image loaded into a publicly available Generative Adversarial Network produced a distorted picture of a monster rather than a human. Lab learners also study why accuracy levels matter as they use a toy supply chain optimization model to predict whether (and why) a package will arrive on time, and how to improve the prediction by changing the variables used, such as product weight and month of purchase.

While very successful at classifying images, language, and consumer preferences, deep learning – a subset of machine learning that uses neural networks – is challenged by inputs and any alterations to them. For example, if an image of a stop sign is provided to an AI system upside down or at an unusual angle, or if the stop sign itself is altered with pieces of tape, the system may not recognize the image as a stop sign. Failure modes become even more likely as the number of machine learning models in AI systems increases (e.g. image to text combined with language detection in the stop sign example), which can interact in different ways depending on the purpose of the system.

Beyond mistakes, some AI systems carry out tasks in ways humans never would. Many examples exist of scenarios producing results developers did not intend, such as a vacuum cleaner that ejects collected dust so it can collect even more (Russell and Norvig, 2010) and a racing boat in a digital game looping in place to collect points instead of winning the race (Amodei and Clark, 2016). In a recent paper from one of the Wilson Center's machine learning experts, this problem of reward hacking is made clear:

> *"Autonomous agents optimize the reward function we give them. What they don't know is how hard it is for us to design a reward function that actually captures what we want. When designing the reward, we might think of some specific training scenarios, and make sure that the reward will lead to the right behavior in those scenarios. Inevitably, agents encounter new scenarios (e.g. new types of terrain) where optimizing that same reward may lead to undesired behavior." (Hadfield-Menell, Milli, Abbeel, Russell and Dragan, 2017)*

Anyone who has played the game twenty questions understands this problem: unless you ask exactly the right question, you will not get the right answer. As more and more AI systems are built and then distributed widely with varying levels of user expertise (some are even designed to be easy for engineers of all abilities to use), this problem – especially in the financial services industry – will only continue. Establishing a framework of ethics for the development, distribution and deployment of AI systems will help spot potential problems and provide more trust in them.

Conclusion

It is not possible to understate the impressive capability of AI systems today, but also how narrow they remain. These systems are in many applications far better than humans at specific tasks but fail when posed with strategic or context-relevant ones. And these problems are not purely American: there are memes in China about unintelligent AI, including a popular one mocking a facial recognition system that accused a woman – on the ad of a bus driving through an intersection – of jaywalking.

AI breaks everywhere, and in places we are not looking.

Thank you. I look forward to your questions.